Financial Sentiment Beyond Text: A Multimodal Approach to understanding Financial Market Dynamics and Investor Behaviours

Andrew Todd

A Thesis Submitted to the University of Strathclyde in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

February 2025

Declaration

This thesis is the result of the author’s original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:

Date:

Acknowledgements

Throughout this PhD there have been many ups and downs but looking back over the course of this degree I can confidently say I have developed both in a professional and a personal capacity. I would like to thank my supervisors, Dr James Bowden and Professor Mark Cummins, for their continued guidance, constructive feedback and patience over these last few years. I would like to extend my gratitude towards all the staff in the Business School who have assisted me over the years. Furthermore, to my family and friends who have been in constant support of my research over the last few years, this thesis would not have been possible without your help.

Finally, I would like to dedicate this thesis to my late Gran and Nana who unfortunately passed away in the last year of my degree. Both were my biggest supporters and were my motivation to keep going when times got tough.

Table of Contents

Table of Figures

List of Tables

Abstract

Over the past two decades, financial literature has extensively examined textual data to identify key drivers of market activity, with sentiment analysis emerging as a pivotal analytical tool. This thesis investigates the application of a multimodal sentiment analysis model for earnings conference calls, combining textual and audio data, to enhance the accuracy of sentiment classification and allow for deeper insights into financial behaviours to be examined. The research identifies a critical gap in existing literature, in that basic sentiment analysis methods dominate despite their underperformance when compared to state-of-the-art natural language processing (NLP) techniques (El-Haj et al. 2018).

Advancing sentiment analysis in the financial domain, Chapter 4 shows that multimodal sentiment analysis significantly outperforms commonly used classifiers in extant literature, in terms of classification accuracy and forecasting capabilities. Chapter 5 reveals that the multimodal model is highly adept at forecasting Cumulative Abnormal Returns (CARs), uncovering a return reversal dynamic between sentiment and CARs that lends support to behavioural theory. This chapter also identifies that framing bias—discrepancies in how information is presented by managers and analysts—intensifies market reactions, emphasizing the role of framing in financial decision-making and providing evidence towards a potential driver of the returns reversal dynamic. Chapter 6 explores the relationship between multimodal sentiment and Cumulative Abnormal Volume (CAV). The findings demonstrate a significant association between multimodal sentiment and long-term trading volume, underscoring the impact of non-verbal communication on investor behaviour. Furthermore, it establishes that sentiment divergence, indicative of disagreement, leads to heightened volume, underscoring the market's sensitivity to conflicting signals.

This research demonstrates the power of multimodal approaches in capturing nuanced financial sentiment and its implications for market behaviour. The findings advance both the technical and theoretical understanding of financial sentiment analysis, particularly highlighting the importance of incorporating audio characteristics alongside textual data in the financial domain.

1. Introduction

1.1 Overview and Background

The study of natural language using machines, or at least the concept, has existed since the 20th century and has been a principal topic for discussion since the inception of electronic computers (Hoard, 2002). Natural Language Processing (NLP) can be traced back to the mid-20th century, where pioneers such as Alan Turing and Noam Chomsky laid the foundations for the understanding of computational and theoretical complexities surrounding the interpretation of natural language by machines. The main aim of NLP research is to explore the capacity that computers have for understanding, processing and now generating natural language (text or speech) for various utilitarian tasks (Chowdhury, 2020). Extrapolating from this definition, it can then be said that NLP researchers' aim is to gain a comprehensive understanding into how humans interpret and use natural language to provide machines with the appropriate basis for completing NLP tasks efficiently. Within NLP there are multiple different tasks that machines have been trained for, primarily entity recognition, text classification, language translation, chatbots, and sentiment analysis. However, the domain extends to any task which requires a machine to understand linguistic complexities and use its knowledge to perform some function.

The specific NLP task that this thesis focuses on is sentiment analysis, which is defined as the computational process of understanding the opinions, attitudes and emotions disseminated in natural language towards an entity (Medhat, Hassan and Korashy, 2014). Mantyla, Graziotin and Kuutila (2018) highlight that sentiment analysis is one of the fastest growing fields of research within the computer science domain, with the increase in popularity due to its ability to accurately capture the publics' opinion in response to a given event. The capacity to extract sentiment from qualitative disclosures, and public discourse, is particularly useful within financial markets as market participants continuously monitor and use financial information in their decision-making process. For example, Mishev et al. (2020) note that the ability to extract accurate sentiment from publicly available and relevant financial information is important for investment decision making by traders, portfolio managers and market participants in general.

A considerable amount of financial literature has attempted to measure the underlying sentiment within qualitative information in an attempt to gain a greater understanding of the dynamic between sentiment and the market. Chapter 2 shows that most studies adopt traditional and comparatively basic approaches to define financial sentiment, such as dictionary methods (Tetlock, 2007; Tetlock, Saar-Tsechansky and Macskassy, 2008; Bollen, Mao and Zeng, 2011; Davis and Tama-Sweet, 2012; Twedt and Rees, 2012; Siganos, Vagenas-Nanos and Verwijmeren, 2014:2017) and supervised machine learning methods (Antweiler and Frank, 2004; Das and Chen, 2007; Li, 2010; Sprenger et al. 2013; Gu, Shi and Tu, 2016; Renault, 2020). Such models typically focus on the frequency with which specific words that indicate sentiment are used within a text, but do not take into account the context within which the word is used. As such, they are easily explainable but do not consider potentially important information, hindering classifier accuracy.

In contrast, state-of-the-art approaches, such as those utilising neural network architecture, are now commonly used within the NLP domain to understand textual information (Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Gillioz et al. 2020; Alamoudi and Alghamdi, 2021).[1] These include Vaswani et al.'s (2017) introduction of transformer architecture, which has rapidly revolutionised the capabilities that machines have with respect to NLP and understanding (Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Gillioz et al. 2020; Alamoudi and Alghamdi, 2021). Various models that have utilised and built upon this architecture have set new benchmarks across multiple NLP tasks, such as Generative Pretrained Transformers (GPT), Text-to-Text Transformer (T5), and Bidirectional Encoder Representations from Transformers (BERT).[2]

A recent autoregressive language model containing 175 billion parameters, GPT-3, has achieved exceptional results across various language tasks, notably scoring the highest accuracy of 86.4% on the LAMBADA language modelling task (Brown et al. 2020).[3] Newer versions of GPT have allowed models to understand both image and text inputs (Open AI, 2023), and have been shown to produce human-level performance across several professional and academic benchmarks.[4] T5, established by Raffel et al. (2020), has also achieved state-of-the-art benchmarks on multiple tasks, including the SQuAD question-answering task.[5] BERT was created by Devlin et al. (2019) to consider the context of textual input from both directions, significantly improving its understanding of language nuances. BERT has shown exceptional performance in NLP tasks such as GLUE,[6] as well as returning superior performance in sentiment classification (Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Alamoudi and Alghamdi, 2021).

Summarising these advances, the methods used in the NLP domain have changed rapidly in recent years, in terms of new benchmarks and model development. As a result, the performance of sentiment analysis classifiers, as a subset of NLP, have also increased. Sentiment analysis has traditionally focused on textual content, as shown within the relevant literature established in Chapter

[1] A comprehensive overview of each of these methods can be found within Chapter 2 of this thesis. Section 2.2.1 provides an overview of dictionary methods and the application of this method in finance research. Section 2.2.2 introduces machine learning approaches in the context of finance research, and Section 2.2.3 highlights state-of-the art models, including transformer architecture.

[2] Chapter 3.8 also provides an in-depth discussion of transformer architecture with a focus on the BERT model.

[3] LAMBADA tests a model's computational ability of understanding text using a word prediction task. This task involves predicting the final word of a sentence using an input paragraph for context.

[4] Interestingly, the model was shown to score in the top 10% of test takers when faced with a simulated bar exam.

[5] The SQuAD Q&A task presents a model a paragraph with a question about the paragraph. The goal of the model is to effectively answer the question posed. The answers to the questions give insight to how well a model can understand text.

[6] The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training and evaluating NLP models.

2. Though these approaches are typically explainable and straightforward to interpret, they lack accuracy relative to more recent, computationally demanding approaches. For example, new models have recently been created to take into consideration multiple modalities, in order to further understand the content being disseminated (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021). Mehrabian (1968) suggests that 7% of human emotion is communicated through the semantic contents of a message, compared to 38% through vocal attributes and 55% via facial expressions. This rule accentuates the lack of information conveyed through the textual modality alone if the message is not transmitted in writing, highlighting that a sentiment classifier which only takes into consideration textual information when there are other available modalities will struggle to accurately measure the underlying sentiment.

Within finance there is one common financial disclosure that offers the ability to utilise more than one modality: the earnings conference call. This disclosure contains textual information via transcripts and paralinguistic information through audio recordings, therefore presenting a suitable medium for the application of multimodal analysis. The combination of text and audio information has been shown within previous psychology literature on several occasions to return improved sentiment classification accuracy over singular modality models (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yang, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021), as the inclusion of additional modalities provides models with deeper behavioural insights into the aspects of human natural language, subsequently improving their abilities to accurately classify sentiment.

The application of these cutting-edge techniques to financial information is scarce and academic scrutiny in this area is in its infancy. Informed by previous literature from the finance, NLP and psychology domains, this thesis contributes to the research area of financial sentiment analysis through the development and application of a multimodal sentiment analysis model that assigns sentiment based on audio and textual information, improving sentiment classification accuracy and therefore offering fresh insights concerning the interplay between sentiment and trading behaviour. It builds upon existing finance research which indicates that the better the understanding a model has of natural language, particularly within a financial context, the greater the ability it has of capturing an accurate sentiment variable and subsequently the greater ability it has in understanding market behaviours. As a result of the interdisciplinary nature of this thesis, the research is also built on the foundations of computational linguistic literature which has identified, across various domains and natural language tasks, that transformer architecture is the most adept model in terms of understanding natural language and returning state-of-the-art results for sentiment analysis projects. Finally, it links to literature from psychology which indicates that, when leveraging the additional behavioural cues provided in non-textual modalities, sentiment classifiers gain a deeper understanding of the communication process.

Through the creation of a finance-specific multimodal classifier we are better able to assess whether this new model possesses a greater understanding of natural language in a financial context, potentially leading to new understandings about how market participants respond to information disseminated on earnings conference calls. Though this classifier represents a technical contribution to the area of finance research, the application of a multimodal classifier also has the potential to further our theoretical understanding by evaluating the magnitude of the market's response to earnings conference call information, both textual and audio. Specifically, this thesis aims to contribute to the understanding of behavioural theory within the asset pricing literature by leveraging state-of-the-art multimodal NLP techniques to define sentiment, allowing for deeper insights into the psychological responses of investors to audio-textual cues in the context of financial markets. This enhanced understanding of financial context, grounded in behavioural perspectives, allows for additional insight into the role of sentiment in asset pricing.

Traditional finance suggests that, in contrast to behavioural theory, market prices fully represent all available relevant information as market participants identify, process, and use the information to make financial decisions (i.e., to buy, sell or hold a given share). This rests on an underlying assumption that market participants are rational and make financial decisions based on new information using probability theory to solely maximise one's wealth. This assumption of rationality leaves little consideration for the emotional aspect of financial market participants, despite evidence that emotional states may impact upon decision making (Conley, Lind and O'Barr, 1978; Erickson et al. 1978; Apple et al. 1979; Wallbott, 1982; Brooke and Ng, 1986; Bradac, Mulac and House, 1988; Chattopadhyay et al. 2003; Feinberg et al. 2005; Klofstad, Anderson and Peters, 2012; Martín-Santana et al. 2015; Wang et al. 2018; Chua et al. 2020; Song et al. 2020). At its core, behavioural theory challenges this assumption of rationality and seeks to evaluate the behaviour of financial market participants under different conditions, and the implications for market activity in each case.

Strikingly, behavioural theory argues that market participants do not always act rationally and are prone to sub-rational decision-making. Specifically, market participants make decisions using heuristics which are often influenced by cognitive biases. For example, the framing bias is a cognitive bias whereby people react differently to the same information depending on how it is portrayed (McMahon, 2005). Earnings conference calls, particularly when evaluated using both modalities on these calls, provide an optimal environment to assess whether market participants are liable to framing bias in the context of financial markets and subsequently evaluate the assumption of rationality.

This thesis consists of three empirical chapters which comprehensively evaluate the capabilities of a multimodal sentiment classifier within the financial domain. The first (Chapter Four) outlines the design of a multimodal sentiment classifier for financial decision making, informed by advances in the finance, computer science and psychology literature. A comparative analysis is then presented, in which the performance of the multimodal sentiment classifier is compared to sentiment classifiers commonly used within the academic finance literature. To do so, an extensive earnings conference call dataset (textual and audio) is used to evaluate the relative accuracy of each classifier. Specific components of an earnings conference call (for example, during the 'management discussion' or 'question and answer' components) for which the multimodal classifier outperforms (and underperforms) are also explored, thus establishing the relative strengths (and weaknesses) of the classifier and identifying the aspects of a call for which multimodal sentiment classification might be most effective. Finally, we explore the specific paralinguistic features (e.g. pitch or tone) that are most informative in the sentiment classification process. Ultimately, the multimodal sentiment classifier is found to return the highest accuracy in comparison to all other models considered. Of particular interest, the results show that the addition of paralinguistic features, through the audio modality, to a state-of-the-art transformer model returns increased classification ability. The performance gain is marginal relative to other state-of-the-art models but marks a considerable outperformance of as much as 30% when compared with popular classifiers such as Loughran and McDonald's (2011) dictionary approach.

Applying this new multimodal classifier to earnings conference calls for constituent firms of the Standard & Poors 100 (S&P100) index, Chapters Five and Six seek to provide fresh insights into previously explored associations between earnings conference call sentiment and both abnormal returns (Chapter Five) and abnormal trading volume (Chapter Six). With respect to abnormal pricing, the findings of Chapter Five are broadly consistent with previous literature, in that multimodal sentiment has: (i) a significant, positive association with short-term abnormal returns; and (ii) a significant, negative association with longer-period abnormal returns. This implies that as the sentiment on earnings calls increases the reaction in market activity, as measured by abnormal returns, increases over an initial three-day period before decreasing over the longer horizon of 60 days. Thus, a returns reversal is identified. These results also indicate that the multimodal model may be more adept at forecasting abnormal returns over short and long-term time horizons than those models used in the prior literature, identified through a stronger coefficient of determination (R^2). This finding is magnified when focusing on short-term abnormal returns, as the multimodal model returns a substantial increase of 0.6 in R^2 over prior related studies of market sentiment.

Additionally, Chapter Five explores the magnitude of market reaction when the dataset of earnings conference calls is isolated to calls exhibiting the highest divergence of paralinguistic features between the managerial and analyst participants, which we consider as an indicator of variation in how information is conveyed through vocal expression. Focusing on paralinguistic features that have been shown to create psychological implications of speaker perception allows for an investigation into whether these psychological biases have any significant impact on financial market behaviours. From this, it is found that abnormal returns are larger in magnitude for those earnings calls characterised by a high divergence of paralinguistic features, particularly short term returns, which may suggest that the manner in which information is disseminated on conference calls has a significant impact on how call content is processed and used in the decision making process. Combined, the findings of Chapter Five are broadly consistent with behavioural theory highlighting investors' psychological biases in relation to the way in which information is framed on earnings calls (framing bias). The analysis indicates that investors may overreact to the information divulged during earnings conference calls, due to the way in which this information is conveyed, leading to a reversal over longer time horizons.

Similarly, Chapter Six analyses market reactions, with a focus on abnormal trading volume as the dependent variable. Previous studies have analysed the relationship between financial sentiment and trading volume (Antweiler and Frank, 2004; Tetlock, 2007; McKay Price et al. 2012; Garcia, 2013; Sprenger et al. 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014:2017) however the association between sentiment and trading volume has been understudied in comparison to that of sentiment and returns. The analysis conducted within Chapter Six indicates that multimodal sentiment has a negative and significant relationship with longer period trading volume. However, no statistical significance is identified in relation to the short-term measure. Chapter Six then evaluates the market reaction to calls which produce significant differences between managerial and analyst sentiment. Building upon prior research, we consider the difference between participant sentiment as a proxy for call participant disagreement (Siganos, Vagenas-Nanos and Verwijmeren, 2014:2017) and evaluate the impact that disagreement among call participants has on the market behaviour of trading volume (Karpoff 1986; Harris and Raviv, 1993; Hong and Stein, 2007; Banerjee and Kremer, 2010; Atmaz and Basak, 2016). Broadly, we find that calls displaying significant disagreement return increased market reactions in relation to abnormal trading volume over the short- and longer-term.

The remainder of this chapter is organised as follows. Section 1.2 outlines the aims and objectives for this thesis. Section 1.3 provides a theoretical justification for this research. Section 1.4 highlights the key contributions of this research. Finally, Section 1.5 outlines the structure of this thesis.

1.2 Research Aims and Objectives

There have been numerous attempts by researchers to define the most robust financial sentiment classifier, with a substantial amount of comparative analysis conducted on financial information sources including annual reports, earnings conference calls, and social media. However, given the performance gains realised in recent NLP models, such as transformer architecture, there has been a lack of advanced NLP models being included within the finance domain. Previous studies have focused on testing the differences between general dictionary-based approaches, specific dictionary approaches and machine learning approaches to sentiment classification (Guo, Shi and Tu, 2016; Renault, 2017, 2020; McGurk, Nowak and Hall, 2020). Generally, these studies have returned a consensus that specifically tailored financial sentiment dictionaries are more adept in classifying financial sentiment than general dictionaries that are not domain specific. However, machine learning methods are commonly recognised as outperforming dictionary-based methods in general. Focusing purely on the comparison of machine learning models, previous research continues to identify that more advanced ML approaches can generate a better understanding of sentiment than generic baseline models (Das and Chen, 2007; Tabari et al. 2019). Contemporary comparisons of sentiment analysis models in finance have compared the performance of machine and deep learning classifiers, with results indicating that deep learning models are more robust at defining sentiment than machine learning models in a financial context (Hiew et al. 2019; Qaiser et al. 2021).

This thesis introduces a multimodal sentiment classifier that takes into consideration both textual and paralinguistic data collected from earnings conference calls. It leverages a financially pretrained transformer architecture (FinBERT) to generate numerical representations of the textual modality produced on these calls, before combining paralinguistic information together in a deep neural network to classify earnings call sentiment. Several studies have demonstrated the abilities of transformer architecture for various NLP tasks outside of the financial domain (Vaswani et al. 2017; Devlin et al. 2019; Brown et al. 2020; Raffel et al. 2020) with results suggesting particularly impressive performance of BERT on sentiment analysis tasks (Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Alamoudi and Alghamdi, 2021). Furthermore, FinBERT has been shown to obtain greater capabilities of classifying sentiment in the financial domain compared to the generic transformer (Hiew, 2019; Yang et al. 2020).

The desire to create a multimodal sentiment classifier arose from the substantial research on emotion classification which has identified that the inclusion of more than one modality allows models to have a richer understanding of the communication process and consequently perform better on downstream tasks (Morency, Mihalcea, and Doshi, 2011; Houjeij et al. 2012; Wollmer et al. 2013; Bhaskar, Sruthi and Nedungadi, 2014; Poria, Cambria and Gelbukh, 2015; Yan, Xu, and Gao, 2020; Dair, Donovan and O'Reilly, 2021). To the authors’ knowledge, Mayew and Venkatalchalm (2012) represents the first study of non-verbal content using financial disclosures to classify the underlying emotional state. They provide initial results indicating that the use of paralinguistic information in financial context is beneficial for sentiment detection. However, the researchers do not focus on textual content, using vocal cues as the single modality in their analysis.

Interactions between sentiment and financial market activity has been extensively studied in previous literature. Relationships between sentiment and various forms of financial information sources have been considered, including financial disclosures (Henry, 2006:2008; Li, 2010; Loughran and McDonald, 2011; Davis and Tama-Sweet, 2012; Jegadeesh and Wu, 2012; Twedt and Rees, 2012; Jiang et al. 2019), news (Tetlock, Saar-Tsechansky and Macskassy, 2008; Grob-Klubmann and Hautsch, 2010; Garcia, 2013; Ferguson et al. 2015; Sun, Najand and Shen, 2016; Audrino and Tetereva, 2019) and social media (Antweiler and Frank, 2004; Bollen, Mao and Zeng, 2011; Mao and Bollen, 2011; Sprenger et al. 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Siganos, Vagenas-Nanos and Verwijmeren, 2017; Gu and Kurov, 2020). However, these papers do not use state-of-the-art classification methods, relying instead on dictionary or supervised machine learning methods, with the most popular technique used to define sentiment being Loughran and McDonald’s (2011) domain-specific dictionary approach.

Building on the previously published model comparisons in finance, as well as the literature identifying that deep learning, transformer architecture and multimodal analysis have been shown to improve sentiment classification in alternative contexts, a contemporary comparison of the available sentiment analysis methods is required. To this end, Chapter Four measures the abilities of the most used financial sentiment classifiers (General Dictionary, Specific Dictionary, Machine Learning, Generic Transformer) against a tailored multimodal model developed by the author. This comparative analysis provides an understanding of the extent to which the models shown to be at the forefront of sentiment classification in other domains can gain a deeper understanding of financial context than existing models, and consequently increase the capabilities of financial sentiment classification. Specifically, this research explores the extent to which the inclusion of the underutilised paralinguistic (audio) modality offered by earnings conference calls adds to our understanding of financial sentiment.

A consensus exists in the literature focused on earnings conference call sentiment and cumulative abnormal returns (CARs), namely that positive short-term CARs coincide with positive earnings call sentiment. However, a limitation of related studies in the area concerns the tendency to focus on short-term return data, while rarely considering the impact of sentiment over longer time periods. Among the limited studies that do examine longer periods, there is a clear lack of agreement as to the nature of the relationship between earnings call sentiment and abnormal returns (Mayew and Venkatachalam, 2012; McKay Price et al. 2012; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Jiang et al. 2019). However, applications of sentiment analysis to financial information sources other than earnings conference calls find an inverse relationship between earnings call sentiment and longer period abnormal returns is expected, presenting a clear gap in the literature that this research seeks to address.

Addressing this gap, the second research question builds on the comparative analysis by seeking to ascertain the capabilities of a multimodal sentiment classifier for forecasting firm level abnormal returns. To do so, Chapter Five applies a tailored multimodal classifier to earnings calls from 95 of the largest US-listed firms between 2006 and 2021. Firstly, the analysis in Chapter Five aims to understand whether multimodal sentiment offers new insights on the relationship between sentiment and abnormal returns, and whether the relationship is consistent with prior text-based measures of sentiment. Secondly, due to the use of paralinguistic features in our multimodal model, the research aims to identify the extent to which trading activity is affected by well-documented behavioural responses to vocal (audio) cues, such as pitch and tone. Specifically, it evaluates the extent to which market behaviour is impacted by calls when there are significant differences in the average levels of participant vocal cues, particularly those which have been evidenced in the related psychology literature as impacting the decision-making process.

The third research question continues to assess the relationship between multimodal sentiment and financial market behaviour but is differentiated through its focus on abnormal trading volume as a measure of financial market activity. A considerable amount of prior research has investigated the extent to which financial sentiment impacts upon trading volume (Antweiler and Frank, 2004; Tetlock, 2007; Garcia, 2013; McKay Price et al. 2012; Sprenger et al. 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014, 2017), and the results indicate that high levels of positive or negative sentiment induce heightened market reaction in trading volume over the short-term. As is the case for the second research question, the relationship between sentiment and longer-period trading volume remains understudied. Indeed, to the author’s knowledge McKay Price et al. (2012) represents the only study addressing this topic, with results suggesting that an initial uptick in trading volume surrounding the call date is followed by a continued rise in trading volume over a longer time horizon. A common theme amongst studies that focus on abnormal trading volume is the tendency to also address the impact that disagreement (measured by variance in sentiment) has on market behaviours (Karpoff, 1986; Harris and Raviv, 1993; Antweiler and Frank, 2004; Banerjee and Kremer, 2010; Atmaz and Basak, 2016), although this analysis is underexplored in the context of earnings conference calls and thus worthy of further scrutiny in this thesis.

Similar to the prior literature exploring associations between sentiment and abnormal returns, the models used in prior studies to explore the relationship between sentiment and trading volume lack understanding of natural language that the techniques at the forefront of NLP – such as transformer models - possess. Therefore, the third research question explores the interaction between multimodal sentiment and abnormal trading volume in order to conduct a robust analysis using more recent methods. Firstly, Chapter Six seeks to ascertain whether a relationship exists between multimodal earnings call sentiment, and the extent to which that relationship is consistent with prior literature. Secondly, we evaluate the extent to which multimodal sentiment has stronger forecasting capabilities for abnormal trading volume compared to prior studies adopting single modalities. Finally, we relate our findings to prior models of disagreement and trading volume (Karpoff, 1986; Harris and Raviv, 1993; Antweiler and Frank, 2004; Banerjee and Kremer, 2010; Atmaz and Basak, 2016) to ascertain whether multimodal sentiment provides any further insights toward the dynamic between sentiment and trading activity, as measured by abnormal trading volume.

Briefly summarising key results, we find that a heightened market response to calls characterised by significant differences in participant sentiment, a result that is more consistent with behavioural theory, given the underlying assumption of traditional theory that market participants are rational and rely on fundamental information when making decisions. However, if reactions to earnings calls are disparate, based upon the way in which information is communicated, and not because of the information itself, then this suggests that psychological influences and biases may affect financial market participant behaviours. Alternatively, if the response to earnings calls remains consistent, regardless of divergence of opinion measures, then our conclusions would perhaps be more aligned to traditional theory, implying that market participants make decisions based on the content of earnings calls and disregard the way in which the information is communicated.

1.3 Key Contributions

This thesis provides both technical and theoretical contributions to the existing body of research which has investigated the impact of sentiment on trading activity. It creates a technical contribution to the financial sentiment analysis field through the introduction of a comparatively more robust sentiment classifier than other commonly used models, and it establishes a theoretical contribution by using a more accurate multimodal classifier to gain further insights concerning financial market participants behaviour in reaction to corporate earnings conference calls. This section first focusses on the technical contributions of this thesis, before addressing the theoretical contributions.

Firstly, the utilisation of the largest financial multimodal dataset to date (to the authors' knowledge) marks a significant technical contribution to the field. This tailored dataset contains a full repository of aligned paralinguistic features, at the sentence level, for 4,860 earnings calls translating into 637,220 sentences. Aligning calls and generating paralinguistic features is a challenging endeavour which has proven to be time consuming and requiring of significant compute power. In comparison, similar explorations of the audio modality leverage considerably smaller datasets. Mayew and Venkatachalam (2012) analyse 466 calls, whereas Chen, Han, and Zhou (2023) examine 848 calls, and Li et al. (2020) use 3,443 calls.[7] The newly-created dataset contains earnings conference call information for constituents of the S&P100, which is unprecedented in its scale and alignment accuracy. By leveraging this new multimodal dataset, the research in this thesis adopted state-of-the-art techniques from the NLP domain to introduce a novel framework for financial sentiment analysis.

Secondly, in the comparative study of Chapter Four, the multimodal sentiment classifier developed in the methods section is shown to excel in classifying sentiment of earnings calls, outperforming the commonly adopted methods used in previous studies. This framework surpasses dictionary-based models, machine learning techniques and generally pretrained transformer architecture-based models in terms of classification performance by combining state-of-the-art NLP models and the paralinguistic modality, which has been significantly understudied in previous literature (see Chapter 2). The multimodal classifier returns the highest testing accuracy of 74.88% in the comparative analysis, which slightly outperforms a similar single-modality textual sentiment model (74.64%) and returns a substantially more robust classification accuracy in comparison to Loughran and McDonald's (2011) specific dictionary approach (47.39%).

The inclusion of paralinguistic features further enhances classification performance – albeit marginally – demonstrating the added value of considering multiple modalities in sentiment analysis through an ability to capture more nuanced aspects of communication. These results support prior assertions made across various domains that the inclusion of additional modalities above and beyond singular modality classifiers increases accuracy (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021). Furthermore, the sizeable increase (+32.49%) in classification accuracy over Loughran and McDonald's (2011) dictionary approach highlights a measurable difference in the performance capabilities of the multimodal classifier introduced in this thesis. In the context of the full dataset of 637,220 sentences used in this dataset, the

[7] Li et al. (2020) focus solely on the initial management discussion section of earnings calls and exclude any paralinguistic information from the Q&A portion. As a result, despite having a call count similar to this study, the number of sentences that include a complete set of paralinguistic features is considerably smaller—394,277 sentences compared to the 637,220 used here.

sizeable increase in classification accuracy over the domain-specific dictionary method translates into 207,033 more sentences being correctly classified.

Thirdly, the multimodal classifier shows a highly significant relationship with both short-term and longer-term abnormal returns. The results indicate superior prediction capabilities, evidenced by a notable increase in the coefficient of determination (R^2 value of 0.6) for short-term Cumulative Abnormal Returns (CARs) compared to the comparative results in relevant studies (Doran et al. 2012; Mayew and Venkatachalam, 2012; McKay Price et al. 2012; Brockman, Li and McKay Price, 2015). The classifier also marginally improves the forecasting of longer-term CARs, showcasing its robustness across different time horizons (McKay Price et al. 2012). The analysis of earnings conference call sentiment has been subject to a substantial amount of interest in recent years, with many papers returning results which have improved our understanding of financial market behaviour (Mayew and Venkatachalam, 2012; Larcker and Zakolyukina, 2012; Doran et al. 2012; McKay Price et al. 2012; Davis and Tama-Sweet, 2012; Wang and Hua, 2014; Davis et al. 2015; Brockman, Li and McKay Price, 2015; Blau, DeLisle and McKay Price, 2015; Milian and Smith, 2017; Borochin et al. 2017; Chen, Nagar and Schoenfeld, 2018; Fu et al. 2019; Bochkay et al. 2020; Amoozegar et al. 2020). As this thesis puts forward a more robust multimodal sentiment classifier, there arises the potential to further understand the relationship between sentiment and financial market activity, subsequently providing the potential to develop and advance the insights of market participant behaviour in relation to asset pricing theories.

Finally, Chapter Six reveals that multimodal sentiment has a strongly significant negative association with longer-period abnormal trading volumes. In comparison to the only other study that evaluates the relationship between earnings conference call sentiment and trade volumes over long time horizons (McKay Price et al. 2012), the findings indicate that multimodal sentiment is more effective at explaining longer period trading fluctuations. Again, this result is identified through a stronger R^2, underscoring the importance of considering multimodal sentiment when investigating market behaviour over extended periods.

In terms of theoretical contributions, the findings of this research lend further support to behavioural theory, specifically in explaining phenomena within the subdomain of asset pricing. By highlighting the role of investor psychology and sentiment, this research strengthens the behavioural perspective as a key explanatory framework in understanding market dynamics. Chapter Five provides significant theoretical contributions by identifying a pattern of mean reversion which is potentially induced by a framing bias arising in market participants decision making. The multimodal sentiment classifier reveals a highly significant, positive relationship with short-term CARs, and shows a highly significant, negative relationship with longer period CARs. These results are consistent with existing literature (Antweiler and Frank, 2004; Lemmon and Portniaguina, 2006; Tetlock, 2007; Henry, 2008; Tetlock, Saar-Tsechanksy and MacKassy, 2008; Schmeling, 2009; Loughran and McDonald, 2011; Doran et al. 2012; Garcia, 2013; Ho and Hung, 2012; Mayew and Venkatachalam, 2012; McKay Price et al. 2012; Twedt and Rees, 2012; Bathia and Bredin, 2013; Corredor, Ferrer and Santamaria, 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Brockman, Li and McKay Price, 2015; Ferguson et al. 2015; Azar and Lo, 2016; Bannier et al. 2017; Gao and Yang, 2017; Jiang et al. 2019) affirming the relationship between financial sentiment and market returns. The initial positive reaction in CARs, subsequently followed by a reversion in returns over longer time horizons, lends support to behavioural theory. Chapter Five identifies a potential initial overreaction to the information provided on earnings calls, moving prices away from their fundamental values. Over the longer horizon, the negative relationship between multimodal sentiment and market returns indicates a reversal of this overreaction, with prices eventually correcting.

This thesis posits that the initial overreaction arises due to the presence of a framing bias influencing the decision making of financial market participants. Given that this research uses a multimodal sentiment classifier that includes paralinguistic information known to impact persuasion and decision making, the classifier potentially identifies nuances in participant communication and identifies that the way in which information is disseminated on these calls significantly impacts market participants' decisions. This framing bias potentially creates the initial overvaluation, which is corrected over time as the market reassesses the information provided on earnings calls. Where framing exists, identified through differences in communication style, discrepancies arise in the decision-making process of wider market participants, leading to further overvaluation and correction. This insight thus contributes to the understanding of how information framing impacts upon market behaviour.

Chapter Six further contributes to the theoretical understanding of financial decision-making by integrating psychological insights which have been shown to impact individual decision-making into the analysis of trading volume behaviour around earnings calls. By using the multimodal sentiment classifier, this research provides a more nuanced approach to examining the impact of sentiment on trading volumes. The findings reveal that multimodal sentiment is significantly associated with long-term trading volume and highlights the influence of emotional and non-verbal communication on investor reactions. The initial positive reaction in trading volume aligns with both sentiment and informational theories, though our findings lack statistical significance. The negative and statistically significant relationship between CAVs and multimodal sentiment over longer-term time horizons suggests a reversion of trading volume to baseline levels, which is consistent with informational theory. Overall, the results of this analysis indicate that neither traditional nor behavioural theories fully account for the observed relationship regarding absolute values of multimodal earnings conference call sentiment and trading volume.

Moreover, Chapter Six advances the literature on Divergence of Sentiment (DoS) by demonstrating its effectiveness as a predictor of both short- and long-term trading volume. Building upon the foundational work by Siganos et al. (2017), which employed a DoS measure as a proxy for investor disagreement, Chapter Six evaluates the relationship between calls characterised by high levels of participant disagreement and abnormal trading volume. To the author's knowledge, this is the first study to identify disagreement using both linguistic and paralinguistic content in earnings conference calls. The DoS analysis reveals that calls marked by high (low) levels of disagreement provoke a larger (smaller) short-term market reaction compared to the main results. This indicates that DoS, which captures significant differences in sentiment between managers and analysts, serves as a robust predictor of short-term cumulative abnormal volume. While both managerial and analyst optimism predict short-term trading volumes, calls characterized by greater managerial optimism trigger a more pronounced market reaction in. This pattern is consistent with subsets of calls showing DoS across the overall call, as well as the DoS associated particularly with the Q&A section.

In summary, where disagreement exists between managers and analysts (proxied by divergence of sentiment), increased abnormal trading volumes are observed in the short term, particularly when managers are more optimistic than analysts. The emergence of statistical significance for multimodal sentiment in the DoS analysis aligns with theoretical models of disagreement which have suggested that divergence among market participants drives heightened trading activity (Karpoff 1986; Harris and Raviv, 1993; Antweiler and Frank, 2004; Banerjee and Kremer, 2010; Atmaz and Basak, 2016). The analysis also explores the relationship between sentiment disagreement during earnings calls and its impact on longer-term trading volumes, representing one of the first studies of its kind to do so. The findings indicate that calls exhibiting sentiment disagreement between managers and analysts lead to a reduction in trading volume over the longer term. This may imply that sentiment disagreement during earnings calls reveals new fundamental information, facilitating better initial decision-making and stabilising prices at their fundamental values, thereby reducing long-term trading activity.

The earnings conference call is shown to be a useful platform in evaluating the presence of heuristics and behavioural biases, as it offers the opportunity to apply the understudied paralinguistic modality on these disclosures. Although there has been extensive research into earnings conference call sentiment and market characteristic reactions, there is still a great deal of research that can be conducted via the application of multimodal analysis to these disclosures. Overall, this thesis contributes to our collective understanding of how multimodal sentiment analysis can more accurately capture the nuances of earnings conference calls, ultimately influencing market behaviour. By contributing both technically and theoretically to the field, this research offers valuable insights into the dynamics of financial sentiment and its impact on market characteristics, with its conclusions providing further insights into the asset pricing domain.

1.4 Theoretical Justification

1.4.1 Finance Theory

The Efficient Markets Hypothesis (EMH), introduced by Fama (1970), became widely accepted since it was first proposed, and stands as an extremely influential study (Malkiel, 2003) that has provided a valuable framework for understanding how financial markets operate. Fama (1970) states that an efficient market is one in which share prices fully represent all fundamental information at any given time. In other words, if share prices "fully reflect" all available fundamental information they can be considered efficient. Market efficiency is built upon multiple assumptions but the one that is commonly criticised, particularly by behavioural theory, is its rational agent belief. In market efficiency, all agents of the market are assumed to be fully rational. Blatuseen (2009) notes that rational agents are fully rational individuals who make decisions based upon probability theory and update their preferences and probabilities when new information comes to light, optimise over all investment alternatives, only consider wealth or value gain in their decisions and are risk averse/neutral in all situations. These characteristics create the foundations for behavioural theory, which attempts to show deviations in general human decision-making behaviour that subsequently moves share prices away from their fundamental values.

Shiller (2003) remarks that market efficiency peaked in interest and was widely considered to be proven beyond doubt amongst the academic community during the 1970s. As with any theory that becomes widely adopted it must withstand critique to maintain its status as the dominant force in explaining some phenomenon. Behavioural theory is the framework which challenges market efficiency. Baltussen (2009) outlines behavioural finance as a theory built upon groundings in psychology and sociology, that attempts to improve the traditional understanding of financial markets. It theorises that investor behaviour affects decisions, rendering them sub-rational and these decisions in turn spill over to market values which subsequently move prices away from fundamental values and render markets inefficient. Hence, behavioural theory posits that some features of asset pricing can plausibly be identified as deviations from fundamental values and these deviations stem from the presence of market participants who are sub-rational. The EMH counters this by asserting that the market comprises both rational and sub-rational agents. As sub-rational agents drive prices away from fundamental values, rational agents quickly negate any mispricing through arbitrage, thereby ensuring market efficiency. However, the theory of limits to arbitrage contradicts the EMH, arguing that rational market participants often lack the ability to correct mispricing. Behavioural theory has studied the structure of these deviations from fundamental values through the lens of human behaviour and decision making.

Four main aspects of human behaviour have been extensively studied and have been shown to deviate human decision-making away from the purely rational assumption defined by the EMH; cognitive ability, heuristics, factors which influence decisions and risk preferences.[8] As defined by the EMH, a rational agent makes decisions solely based upon the laws of probability. In other words, rational agents use perfect, logical deductive reasoning in the decision-making process (Arthur, 1994). Deductive reasoning refers to the process of making decisions based on a top-down approach. It begins with a hypothesis and evaluates all possibilities to reach a specific logical conclusion. Arthur (1994)

[8] Barberis and Thaler (2003) note however that behavioural theory argues that phenomena seen within financial markets can still plausibly be explained when market participants are not fully rational.

notes that humans do not possess the cognitive abilities to be perfectly deductive - especially in a financial setting:

> "If we were to imagine the vast collection of decision problems economic agents might conceivably deal with as a bottomless sea or ocean, with the easier problems on top and more complicated ones at increasing depth, then deductive rationality would describe human behaviour accurately only within a foot or two of the surface." (Arthur, 1994, p.1)

This breakdown in deductive reasoning is largely due to one main factor; past a certain level of complexity, humans do not possess the cognitive ability to successfully apply this logical process (Tversky and Kahnmen, 1974; Arthur, 1994:1995). Thus, the use of probability theory to optimize over all relevant investments to make financial decisions is not common practice for the average market participant. There exists a great deal of literature evaluating the cognitive abilities of humans and the extent to which we can solve complex problems. Humans still possess the ability to solve complex problems, however not to the extent of complexity often found within financial markets. Tversky and Kahnmen (1974), show that humans reduce the complex tasks of assessing probabilities and predicting values to simpler judgemental tasks using heuristics. In their paper they discuss three heuristics that are commonly used in financial settings: representativeness heuristic, availability heuristic, and adjustment and anchoring heuristic.

The representative heuristic uses the similarity of an event or individual when making a judgement surrounding a similar event or individuals' likelihood. Baker and Nofsinger (2010) note errors stem from this heuristic due to the lack of relevance given to base rate information prior to a judgement and the reliance on small samples for future forecasts. Kahnmen and Tversky (1982) show this bias through an early experiment surrounding the judgement of occupations based upon a given profile. The experiment showed that participants disregarded base rate data and based decisions upon stereotypical characteristics of occupations. Boussaidi (2013) demonstrates the effect of this bias in a financial setting and finds evidence suggesting that investors tend to react with excessive optimism (pessimism) to a series of good (bad) earnings news. The effect is extrapolating good (bad) information too far into the future and inciting overvaluation (undervaluation) of share prices, the assumption being that past news publications (a small statistically invalid sample) are representative of long term-future firm performance.

Tversky and Kahnmen (1974) define the availability heuristic as the assessment of the frequency of an event by how easily information regarding said or similar events comes to mind. In other words, the weighting one gives to the probability of an event occurring depends on how easy examples come to mind.[9] Typically, the use of this heuristic is accompanied by the overweighting of

[9] In other words, if an event occurring can be recalled with ease, individuals are likely to overestimate the probability of such an event occurring.

current information instead of processing all relevant information (Kilger and Kudryavtsev, 2010). Shiller (1998) asserts that investors' attention to certain categories of investments are influenced by alternating waves of public attention or inattention. Furthermore, Barber and Odean (2008) show that investors choosing to buy stocks tend to only take into consideration shares that are in the news, shares experiencing high levels of trading volume and shares with extreme one day returns. These types of shares are usually well documented and hence at the forefront of market participants' attention. Thus, the ease of recall gives rise to the availability heuristic and indeed the accompanying biases. Kilger and Kudryavtsev (2010) documented that for both analyst recommendation upgrades and downgrades, abnormal event-day stock returns were significantly higher if the contemporaneous return on the composite market index was positive, supporting prior evidence of the availability heuristic being used within financial markets.

The final heuristic discussed by Tversky and Kahneman (1974) which is commonly discussed in financial decision-making is anchoring and adjustment. An investor may make a comparative assessment of whether a share price will increase or decrease. After this initial assessment they will then go on to judge the magnitude of change based upon the comparative assessment - making an absolute estimate. Through this process, the investor then makes predictions of future share prices, otherwise known as the anchoring and adjustment heuristic (Pena and Gomez-Mejia, 2019). However, this heuristic along with the others is liable to bias. Epley and Gilovich (2002) note that bias occurs from an insufficient adjustment away from the initial anchored value. Baltussen (2009) gives an example of this bias in a financial setting. When there has recently been a movement in the price of a share that corrected a mispricing, investors may still anchor on this past price and expect it to continue. Thus, a disproportionate weight is assigned to the past price in future decisions in the decision-making process, invoking insufficient adjustments and creating price volatility. These heuristics can evidently speed up our decision-making process and are very useful in certain situations where human cognitive abilities cease to be of use. However, the use of heuristics often leads to biases leading to inaccurate decisions. These erroneous decisions stray from the decision outcomes of the rational agent and often, as shown in examples above, move prices away from fundamental values and hence markets away from efficiency.

Based upon the profile of a rational agent, it is considered that financial market participants solely consider utility or wealth when making financial decisions. In other words, external variables are considered unimportant and not used in reaching conclusions. Fehr and Schmidt (2006) note that traditional theory still routinely assumes that material self-interest is the sole motivation of all people. Extensive experimental testing has been done on the self-interest hypothesis, with many studies using experimental games such as the ultimatum game. Guth et al. (1982) conduct an experiment surrounding bargaining behaviour using the ultimatum game. Results of the study show that people often base their strategy on fair outcomes. The authors note that participants make decisions in the game based upon what they consider to be a fair or justified result, showing that people are altruistic and take into consideration others' final positions, not just their own. Thus, this study suggests deviations from the self-interest hypothesis. The altruistic nature of humans is evident in a well-known market anomaly - sin stocks. Blitz and Fabozzi (2017) note that historical evaluation of sin stocks has shown they deliver significantly positive abnormal returns. Yet many investors shy away from investing in such opportunities as they deviate from their moral compass. Fabozzi, Ma and Oliphant (2008) find evidence that sin stocks are initially under-priced. Hence, to a rational agent a sin stock would be extremely attractive. However, multiple external factors drive investors away from these stocks, mainly societal and moral.

The last departure from a rational agent's decision-making process surrounds risk preferences. Kahneman and Tversky (1979) give a critique of Expected Utility Theory (EUT)[10] and develop their own model of decision making under risk - Prospect Theory. A main component of Prospect Theory which continuously arises and is in contradiction to EUT is risk preferences. In EUT, which is based upon a rational agent's decision-making process, agents are considered risk averse or risk neutral. However, Kahneman and Tversky (1979) show that risk preferences differ from rational expectations and are in fact affected by behaviour. Specifically, humans are risk averse over gains but risk seeking over losses, a finding which is supported by earlier work from Markowitz (1952). The findings identify the same risk pattern for outcomes with large magnitudes - risk averse over gains and risk seeking in losses. Markowitz (1952) also shows that for the opposite, outcomes with small magnitudes, people exhibit risk seeking tendencies in gains and risk aversion in losses. Coval and Shumway (2005) demonstrate the presence of this effect in a financial setting, noting traders who are exposed to losses in the morning session have an increasing propensity to be risk seeking in the afternoon. In line with prospect theory these traders who end up experiencing losses become risk seeking.

Furthermore, Kahneman and Tversky (1979) show that people care disproportionately more about losses than they do about gains. The aggravation that one experiences in losing a sum of money appears to be greater than the pleasure associated with gaining the same amount. Building upon this point the authors confirm that people perceive outcomes as gains and losses rather than final wealth positions. These gains and losses are defined by some reference point. Multiple human behaviours can impact the value of this reference point, past decisions, aspirations, expectations, social comparison, other available alternatives and outcomes. These deviations from EUTs risk preference structure provide contradictory conclusions and create evidence toward market participants being sub-rational decision makers.

Out with the three heuristics discussed by Tversky and Kahnmen (1974), one judgemental heuristic which directly aligns with the work conducted in this thesis is the framing heuristic. The framing effect highlights that individuals make decisions based upon how information is presented, or 'framed', rather than only considering the pure facts. In other words, this heuristic indicates that

[10] EUT is a decision theory that models a rational market participants decision-making under uncertainty.

differing decisions can be made about the same information based upon how information is framed. McMahon (2005) highlights that experimental studies have found that the way in which information is framed has a significant impact on the ultimate choice that is made surrounding the given information. Particularly, pertinent to financial information, there can be a significant difference in outcomes if information is framed in terms of losses as opposed to gains.[11] The framing heuristic stands in contraction to traditional theories assumption of rational agents as a principle of rational decision-making is that choices should be independent of the way in which information is communicated and solely be based upon the content provided.

1.4.2 Psychology Theory

In the above text, the decision-making process for rational agents and normal market participants has been discussed. The main point surfacing is the effect that behavioural, cognitive and emotional aspects can have on decision-making, which subsequently impacts market efficiency. However, the points laid out above discuss general factors which induce bias in decisions. Earnings calls disseminate information through two modalities of information, textual and audio. The psychology literature, to the author's knowledge, contains multiple experimental studies that analyse the response to natural language that is framed in various ways. Particularly, prior studies have evaluated the extent to which varying levels of paralinguistic cues impact the decision-making process and incite greater levels of persuasion on listeners (McCroskey and Mehrley, 1969; Mehrabian and Williams, 1969; Miller et al. 1976; Conley, Lind and O'Barr, 1978; Erickson, Lind, Johnson, and O'Barr, 1978; Apple et al. 1979; Hollandsworth et al. 1979; Wallbott, 1982; Brooke and Ng, 1986; Bradac, Mulac and House, 1988; Burgoon, Birk and Pfau, 1990; Chattopadhyay et al. 2003; Feinberg et al. 2005; Klofstad, Anderson and Peters, 2012; Kennedy, Anderson and Moore, 2013; Martín-Santana et al. 2015; Gaertig and Simmons, 2018; Guyer, Fabrigar and Vaughan-Johnston, 2018; Wang et al. 2018; Buelow et al. 2020; Chua et al. 2020; Song et al. 2020; Van Zant and Berger, 2020).

Although across most financial disclosures the main modality in which information is disseminated is text, earnings conference calls communicate financial information using natural language utilising both the textual and audio modality. Hence, evaluating earnings conference calls offers an ability to understand not only the extensively studied impact textual information has on market participants and subsequently market characteristics, but also the impact paralinguistic cues may have on financial markets. Guyer, Fabrigar and Vaughan-Johnston (2018) remark that a great deal of research has revealed that the content of what we say matters, but also emphasise that the way in which we communicate matters. How we speak conveys substantial information beyond the content of communication. There is a considerable body of psychology literature that relates to vocal characteristics and their impact on persuasion and decision making more broadly.

[11] Kahnmen and Tversky (1979) discover that individuals are risk averse over gains but risk seeking over losses.

Burgoon, Birk and Pfau (1990) investigate the relationship between nonverbal behaviour and persuasion, evaluating which specific nonverbal behaviours return the strongest relationships with persuasion. The authors consider how nonverbal cues relate to perceptions of pleasantness, immediacy, potency, dominance and arousal (or its opposite, relaxation). Prior studies have shown that pleasantness related vocal attributes undermine persuasiveness and attitude change (McCroskey and Mehrley, 1969; Mehrabian and Williams, 1969; Erickson et al. 1978; Hollandsworth et al. 1979). Potency cues such as loudness, tempo, intensity, dynamic and confidence are documented in the literature as having increased persuasive abilities (Mehrabian and Williams, 1969; Miller et al. 1976; Apple et al. 1979). Furthermore, a faster speech rate and increased vocal intensity due to louder amplitudes, greater intonation, greater fluency and faster tempo forming a more confident speaking style in turn enhance persuasiveness (Mehrabian and Williams, 1969; London, Meldman and Lanckton, 1970; Erickson et al. 1978; Edinger and Patterson, 1983).

To assess the above vocal cues on persuasiveness the authors videotaped participants delivering a persuasive speech. Following each persuasive speech, audience members completed a persuasiveness scale. From the analysis Burgoon, Birk an Pfau (1990) confirmed that persuasiveness increases with greater fluency, and greater pitch variety but not greater voice quality. Furthermore, findings indicate that a faster tempo, more tempo variety and greater loudness have no significant effect on persuasiveness. Of all the variables stated, speech fluency showed the strongest persuasive power. This result pertains to both the introductory and question-and-answer section of an earnings call. In the introductory statements if managers speak in a fluent manner they will be perceived as competent and confident. This in turn will enhance the probability of analysts and investors on the call being persuaded toward the manager's particular train of thought. Furthermore, answering questions effortlessly without the use of filler words will increase the persuasive power of the managers due to the confident speaking style conveyed.

In a more recent study, Van Zant and Berger (2020) complete a comprehensive review and analysis of the effects of paralinguistic cues on persuasion. The authors note that most persuasion research has focused on what people say such as the above studies in language intensity, extremity and vividness but less is known about how they say it. They assert that communicators often modulate their vocal features, such as speaking at different levels of loudness or varying their pitch. The study sets out to understand how individuals use linguistic cues in persuasion attempts and if these adjustments aid persuasion. Van Zant and Berger (2020) suggest two possibilities as to why paralinguistic attempts may boost persuasion: detectability and confidence. They contend that vocal cues may be effective in persuasion attempts as they evade detection. Friestand and Wright (1994) conclude that the extent to which persuasion cues succeed depends on the level of detection of the persuasive intent: if it is extremely obvious one is attempting to persuade, its chances of realisation are slim. Tenney et al. (2019) demonstrate the difficulty in inferring a communicator's intent through vocal attributes, noting it becomes especially hard when the sender is motivated to conceal their intentions (Bond, Kahler and

Paolicelli, 1985; DePaulo et al. 2003; Ten Brinke et al. 2014). Hence, due to detectability reasons certain vocal cues may potentially be more persuasive than linguistic attributes. However, Van Zant and Berger (2020) contend that the successfulness of persuasion via paralinguistics may be diminished with the inclusion of linguistic persuasion cues (such as the ones noted above) due to their higher probability of detection.

Gaertig and Simmons (2018) show that confidence is a powerful tool for persuasion as it shapes message receivers' judgement. Similarly, Kennedy, Anderson and Moore (2013) suggest that people continue to be influenced by nonverbal confidence displays even when they know the source to be biased. Hence, even if the communicator's attempts to persuade are noticed (as long as they are not so excessive it is thought of as disingenuous) then paralinguistic confidence has the potential to enhance persuasion. Prior literature has examined persuasive language, but relatively little is known about how persuasive people are when they attempt to persuade through paralanguage, or acoustic properties of speech (e.g., pitch and volume).

Van Zant and Berger (2020) set up a total of five experiments to understand the overall influence of vocal attributes on persuasion. Discussion of all studies would be too extensive for this thesis, therefore only the relevant analysis has been highlighted. Initially, the authors asked speakers to read a transcript about a product review twice - first normally and second using persuasive paralinguistic features to make the product more attractive. These recordings were then played to listeners, and they were asked if they would purchase the product and if so what their anticipated satisfaction with the product would be. The statements read with enhanced vocal cues significantly increased persuasion, with listeners who were played the paralinguistic attempt viewing the product significantly more positively. In a follow up investigation, Van Zant and Berger (2020) included a disclosure statement which brought to the listeners attention that it was a persuasion attempt. This allowed the authors to understand whether the detectability or confidence condition was at play. Results show that irrespective of the listener’s knowledge of a persuasion attempt, the paralinguistic persuasion succeeded. Hence, engaging in paralinguistic persuasion attempts made speakers appear more confident which falls in line with the confidence account. Overall, these results indicate that paralinguistic persuasion attempts increase persuasion regardless of whether the listener knowns speakers are trying to persuade them.

Continuing their extensive research into the impact of paralinguistic cues on persuasion, Van Zant and Berger (2020) evaluated whether the same vocal cues were being used to portray confidence and in turn persuade the receiver. Looking particularly at volume, pitch and speech rate measures, the result indicates that paralinguistic attempts to persuade showed speakers modifying their vocal cues in a similar fashion across the studies - with the exception of pauses. Specifically, they found that in attempts to persuade using paralinguistics, speakers spoke at a higher volume, spoke at a higher pitch, varied their volume to a greater extent and spoke at a faster rate. Additionally, Van Zant and Berger (2020) find that the two paralinguistic cues most adept at heightening persuasion are increased volume and a more varied level of volume. These effects also were shown to portray the speaker in a more confident manner which in turn, as discussed above, increases persuasion.

Guyer, Fabrigar and Vaughan-Johnston (2018) set out to investigate vocal attributes that they identify as being related to persuasion, but also being understudied to date, namely vocal speed, vocal intonation and vocal pitch. They further evaluate whether these cues remain persuasive and impact attitude change at different levels of elaboration surrounding the message.[12] The researchers identify more topic-relevant thoughts among the group that were assigned the high elaboration condition compared to the low group. In relation to the vocal attributes, the authors find that increased speech rate and falling vocal intonation was rated significantly as more confident by participants. Furthermore, there was no interaction between their elaboration variables and speaker confidence. This suggests that speaker confidence is not influenced by the amount of receiver information processing. Further tests show that the magnitude of effect for vocal speed and intonation on persuasion were similar.

Guyer, Fabrigar and Vaughan-Johnston (2018) use the high and low elaboration condition as an assessment tool to understand the differences in attitude change when individuals are focused on an argument or not. Initially, using a chi-squared statistic the authors show that the two conditions are significantly different from each other in the way that they affect individuals. This finding falls in line with Petty and Cacioppo's (1986) Elaboration Likelihood Model of Persuasion (ELM). Such a model implies that the influence of persuasion on an individual varies based upon the extent of careful thought surrounding an argument. With extensive thought (the high elaboration condition) the ELM implies that a variable should bias thought-favourability, which in turn guides the formation of attitude. For the opposite, the low elaboration condition, vocal cues should still influence perceptions of speaker confidence, however confidence should no longer bias thought-favourability. Instead, it directly influences attitudes as a peripheral cue. Their results confirm the ELM's statements surrounding the effect of confidence on attitude formation. They show in the high elaboration condition that perceptions of speaker confidence - created through the discussed vocal cues - can bias thought-favourability. However, for the low-elaboration condition speaker confidence directly affects attitudes as a peripheral cue - not through thought-favourability. The study from start to finish was replicated for vocal pitch with results returning the same as described above. Hence, it can be concluded that vocal speed, vocal intonation and vocal pitch all create perceptions of speaker confidence which in turn are persuasive variables for attitude change, all of which change attitudes differently based upon the level of information processing conducted by the receiver.

The above findings are interesting in the context of earnings calls as, from these, it would be reasonable to assume that the level of information processing conducted on an earnings call is quite high. Analysts and investors are seen as intelligent individuals who will look to process details discussed

[12] See Guyer, Fabrigar and Vaughan-Johnston (2018) p.165 for an in-depth explanation of their method to assess the persuasiveness.

on the call. Therefore, Guyer, Fabrigar and Vaughan-Johnston (2018) findings suggest that even if the content of a message is being thoroughly examined and processed by a receiver, vocal cues that create the perception of confidence still have a persuasion effect that can change attitudes.

Buelow et al. (2020) evaluate the effect of prosody on decision making, particularly assessing the impact of speech rate. Testing this effect, the authors primed college students and preschool kids with either fast or slow speech before giving them the Hungry Donkey Test (HDT).[13] The findings showed college students who were exposed to fast speech were unable to differentiate between the advantageous and disadvantageous outcomes. However, the slow prime condition showed preference towards advantageous decisions, implying a better learning of the task. Hence, riskier decisions are made when exposed to the fast prime. Thus, in the context of corporate verbal communication between management and interested parties an overall faster level of speech may lead interested parties into making riskier decisions based on the information provided, potentially leading to incorrect forecasts and deviation from fundamental valuation.

This section has so far discussed separate linguistic and paralinguistic manipulations that have been shown to have persuasive effects on message receivers. The main driver of persuasiveness stemming from the literature is perceptions of speaker confidence, which can be created through linguistic and paralinguistic cues. Van-Zant and Berger (2020) alluded to the possibility that the inclusion of more recognisable linguistic manipulations along with paralinguistic persuasion attempts may diminish persuasive power. However, the literature points towards a confidence theory which shows that even when receivers understand that the communicator is trying to persuade, if the linguistic and paralinguistic attempts create an image of confidence surrounding the communicator, it can still shape the receiver's judgement. In addition, whether the message content is being processed extensively by the receiver or not, Guyer, Fabrigar and Vaughan-Johsnton (2018) show that vocal cues still affect attitude change even if through different means.

From the literature discussed, perceptions of confidence created through linguistic and paralinguistic cues have a strong relationship with persuasion and subsequently attitude change. Although there has been extensive analysis into the textual content of financial information sets and subsequent market reactions to such financial information there has been little analysis conducted on the impact that paralinguistic information has on market participants and subsequent market reactions. The paralinguistic cues extensively discussed in the psychology literature, to the author's knowledge, have not been evaluated in a financial context. The findings from the psychology domain would suggest that language extremity, language intensity, language vividness, vocal volume, variation of vocal

[13] The HDT was created as a task for children which mimics Bechera's Iowas Gambling task. The task tests cognition and emotion and was originally developed to assist in detecting decision-making impairment in patients with prefrontal cortex damage. It involves the participant choosing from four doors, each with a cost or reward in apples. The objective of the participant is to give the donkey the most apples possible. HDT is often completed using a computer and carried out in real time to resemble real world contingencies.

volume, vocal pitch, vocal speech rate, vocal intonation and fluency all have persuasive effects on behaviour/attitude, and such could be incorporated into financial decision-making information set.

The inclusion of paralinguistic features to the sentiment classifier used within this thesis, uniquely allows for further evaluation as to which asset pricing theory is more adept at explaining market behaviour. Particularly, the analysis directly evaluates whether market participants are subject to judgemental framing when making financial decisions based upon information conveyed. It has been identified within the psychology literature that certain linguistic and paralinguistic modifications heighten the persuasiveness of communication in a general setting. Incorporating these paralinguistic cues which are thought to impact decision making, in a financial context will heighten the literatures' understanding of how market participants respond to such information. Do they only consider the fundamental information or is the way this information is framed a prevalent factor?

In line with semi-strong market efficiency, earnings conference call information will be used in rational market agents' financial decision-making processes and realised quickly and efficiently in share prices. A rational agent, according to the EMH, will only take into consideration fundamental information that directly speaks to the performance and future expectations of a given firm. Hence, the way in which information is communicated should not be a factor in the way in which such information is received, understood and used in this decision-making process. Therefore, it is expected that if agents of the market are rational and subsequently markets are efficient, share prices following the release of an earnings call will quickly incorporate the information disseminated on these calls, setting prices to fundamentals.

Behavioural theory, on the other hand, leans on literature in psychology and sociology relating to human decision making and suggests that investor behaviour impacts financial decision-making and consequently creates sub-rational decisions. As market agents are inherently sub-rational, these sub-rational decisions affect share prices and create deviations away from their fundamental values, creating market inefficiencies. Results stemming from this thesis that indicate that market participants do take into consideration the way in which textual and paralinguistic modalities frame information, and not just the information itself, would lean in favour of behavioural theory as the superior explainer of market activities. If this is indeed the case, an evaluation of the manner with which information is communicated to investors could provide insights into how investors make decisions influenced by behavioural cues in financial markets. This potentially allows for a greater understanding of financial decision making based upon the earnings call information set.

By calculating the sentiment of content produced on earnings calls, using both modalities, the results of this thesis will give insight into how market participants digest and use such information. Specifically, by evaluating the relationship between multimodal sentiment and market behaviours, the conclusions of the analysis conducted within this thesis will give insight into how market participants react and whether these reactions are in line with that of a rational or sub-rational agent. This may consequently provide a deeper understanding as to which of the two asset pricing theories discussed are more adept at explaining market behaviour.

In a similar light to Shiller's (1998) review of the literature, this section focussed primarily on behavioural theory given that it is more aligned to the findings of this thesis. The aim of this section is to give theoretical justification for the application of a sentiment analysis classifier which considers both the textual and paralinguistic information on earnings conference calls in classifying financial sentiment and understanding responses to such information.

1.5 Thesis Structure

The remainer of this thesis is as follows: Chapter Two critically evaluates the literature relating to financial sentiment, with a specific focus on studies that have used earnings conference calls as a medium for textual sentiment analysis. Furthermore, research gaps in the domain are highlighted within this chapter. The research design, dataset and methods introduced and used within this thesis are then discussed in Chapter Three. Chapter Four is the first of three empirical studies included within this thesis, presenting a comparative study which identifies the robustness of a comprehensive selection of sentiment analysis classifiers that are either commonly used in finance research, or are at the cutting-edge of NLP research. Chapter Five represents the first of two event studies evaluating the relationship between multimodal earnings conference call sentiment and abnormal returns. Similarly, Chapter Six employs an event study methodology to evaluate the relationship between multimodal sentiment and cumulative abnormal trading volumes. Finally, Chapter Seven summarises the key findings of the research, discusses the implications, and suggests areas for future research. It also reflects on the limitations of the study and the contributions it makes to the existing body of knowledge.

2. Literature Review

2.1 Introduction

This chapter examines the body of work that has applied natural language processing (NLP) techniques - particularly sentiment analysis - in the finance domain. Typically, such work seeks to understand the intricacies of qualitative information and its relationship with financial markets. Over the last two decades, relevant literature has evaluated various sources of textual data to determine statistically and economically significant drivers of financial market activity. To the authors' knowledge, Antweiler and Frank (2004) conducted the first study to apply sentiment analysis to the finance domain. Their paper's findings have given birth to a wide ranging field of analysis which builds upon various sources of qualitative information and sentiment analysis techniques.

Through the analysis of different mediums of information, authors have created and applied various methods to measure sentiment. In academic finance, the earliest and most practiced method used to determine sentiment is the dictionary approach (see page 3 below). In attempts to create more accurate models, authors have utilised more computationally demanding machine learning approaches and - more recently - deep learning approaches. Similarly, various stock market metrics have been evaluated in relation to sentiment. To review the relevant literature systematically and comprehensively, the remainder of this chapter will critically evaluate information mediums, methods to determine sentiment and metrics of interest to create a full representation of the field. Furthermore, maintaining the forward-looking approach taken in survey papers such as Li (2010), Loughran and McDonald (2016) and El Haj et al. (2018) the aim of this literature review is to focus not only on the literature published solely in the accounting and finance domain, but also incorporate relevant studies originating from computer science. This allows for established methods and taxonomies used within computational linguistics to be mapped into an accounting and finance context, therefore establishing clear directions in which the accounting and finance literature can be advanced. The remainder of the chapter is as follows: section 3.2 discusses the application of sentiment analysis techniques to multiple different financial information sources and covers the relationship sentiment measures have with market characteristics. Section 3.3 looks specifically at sentiment analysis literature surrounding earnings conference calls. It highlights the different specific sentiments that can be drawn from earnings calls and again identifies the associations that such measures have with market characteristics. Finally, section 3.4 summarises the numerous studies discussed throughout this Chapter, highlighting key themes and identifying gaps in the extant literature.[33]

2.2 Applications of Sentiment Analysis in Finance

Information sharing has drastically improved since the early 1800s, when information could take weeks to months before arriving at its destination (Dombkowski, 2021). In a finance context, this presented opportunities for market participants to take advantage of information asymmetries resulting from the slow speed in relaying information through official channels. For example, Sharf (2007) explains how investment banker Mayer Rothschild’s extensive network of carrier pigeons delivered the news that England had beaten France in the battle of Waterloo before any other person in London. With this timely information he went long on British Government bonds when the general market consensus at the time was that there would be imminent news of a British defeat, and hence a drop in such bonds value.

With the adoption of the internet on a commercial scale, there has been a drastic increase in internet users worldwide, from 413 million in 2000 to upwards of 3.4 billion in 2016 (Roser, Ritchie and Ortiz-Ospina, 2015). The population adoption of the internet has created a world where information dissemination is instantaneous. The rise of the internet has revolutionised twenty first century living and of particular importance streamlined human communication channels through the creation of email,

[33] A condensed version of this chapter has been published in Intelligent Systems in Accounting, Finance and Management titled “Text-based sentiment in finance: Synthesising the existing literature and exploring future directions”.

websites, instant messaging and social media. Financial communication channels,[34] which are of particular importance to this research, have also increased their information outreach and the speed in which information is delivered to the masses.[35] Peress (2014) highlights that technology has increased access to a wealth of diverse information. For example, the information stemming from the financial communication channels noted above. The successful creation and application of models that accurately determine the underlying meaning of dense information sets creates actionable signals for investors. Similar to the early example of Rothschild's carrier pigeons, such models can create an information advantage in financial markets.

This section examines the existing body of work that has applied NLP techniques - particularly sentiment analysis - to financial information sources.[36] Earnings Conference Calls are excluded from this section as they have a dedicated section following this commentary due to said source being the focus of the proceeding empirical research. The extant literature to date has attempted to understand the intricacies of qualitative information as characteristics stemming from the information, particularly sentiment of the mentioned modalities, is understood to have a relationship with financial market metrics (Das and Chen, 2007; Tetlock, 2007; Loughran and McDonald, 2011; Mao and Bollen, 2011).

Along with the various information sources that have been used within this area of research, there have also been multiple techniques employed to determine sentiment. This section is structured by first focussing on earlier approaches to textual analysis in finance, before progressing through the published literature towards more recent and computationally demanding approaches. Cambria and White (2014) highlight the continuous search for more accurate approaches is due to automatic analysis of text involving a deep understanding of natural language by machines which is a reality we are still yet to reach.

2.2.1 Dictionary Approach

The first and most prominent approach within the literature is the dictionary approach (also known as the word count approach). The concept behind this approach is comparatively intuitive compared to more recent methods, in that the sentiment of the overall piece of text is determined by the sentiment of the specific words within the text (Li, 2010). Dictionaries are created containing positive and negative words respectively. Using these dictionaries, the frequency of words contained within a body of text are counted. After this count is complete, a calculation of the difference infers how positive or negative the text is.

[34] At the market level financial news outlets (Financial Times, Bloomberg, Yahoo! Finance) have website, apps and social media accounts. These methods of information dissemination are also used by individual firms.

[35] Whereas, in times prior to the mass-adoption of the internet, an investor who wanted to learn about potential investment opportunities might pay a company directly for the latest financial reports, an internet search in the modern day may return every financial report, financial headline and internet posting in relation to a particular public company free of charge.

[36] Financial disclosures, newspaper articles, financial news headlines, social media and internet message boards.

Tetlock (2007) uses a word count approach to understand whether financial news media induces, amplifies or simply reflects investors' understanding of stock market performance. He uses the Wall Street Journal's (WSJ) frequently published "Abreast of the Market" opinion piece and interprets its influence on trading activity for constituent firms of the Dow Jones Industrial Average (DJIA). Specifically, the author tests whether high media pessimism[37] (referred to as the pessimism factor) is related to low investor sentiment, which in turn results in downward pressure on prices.

Tetlock (2007) runs Vector Auto Regressions (VAR) to evaluate the pessimism factor's ability to predict returns and volume. In doing so, the author identifies that the pessimism factor returns a statistically significant negative influence on the next day's returns. Specifically, a one standard deviation change in the pessimism factor drives an 0.081% change on DJIA returns. However, this fluctuation is shown to be almost fully reversed by the end of the trading week. Further analysis suggests that the pessimism factor is a significant negative predictor of volume. Tetlock (2007) notes that this result was expected as high absolute values of pessimism are a proxy for disagreement between noise traders and rational traders, hence leading to an increase in trading volume on the next day. This finding agrees with Hirshleifer (1977) and Harris and Raviv (1993), who suggest that disagreement drives higher levels of trading volume because trading occurs when market participants assign different values to an asset.

To assess the pessimism factor's economic significance, Tetlock (2007) creates a trading strategy. His approach borrows at the riskless rate when the prior days negative word count is in the bottom third of the prior year's negative word distribution and sells them back one day later. On the contrary, one day after the negative words are in the top third of the prior year's negative word distribution, he borrows all the stocks on the Dow and buys them back one day later. This produces an average daily return of 0.044% which translates into a 7.3% annual return. The results are strongly significant and imply economic importance however do not take into consideration trading costs.

Expanding on the prior work of Tetlock (2007), Tetlock, Saar-Tsechansky and Macskassy (2008) expand on the above paper using a similar technique, but applying it to all articles contained within the WSJ and Dow Jones News Stories (DJNS) for constituents in the Standard and Poor's (S&P) 500. The authors demonstrate that news stories surrounding a specific company are highly concentrated around the time of the earnings announcement.[38] The firm-specific news stories are centred around one day before, the day of and one day after earnings announcements, thus suggesting that media reporting plays an important role in communicating earnings announcement information to a wider audience.

[37] Calculated using a word count approach with the Harvard IV psychosocial dictionary. The Harvard dictionary was first created for content analysis in the behavioural sciences (Stone and Hunt, 1963). 77 predefined categories are created from all words within the Harvard dictionary. This pessimism factor is a linear combination of 4 categories from said dictionary namely: Negative, Weak, Fail and Fall word categories.

[38] See Tetlock, Saar-Tsechansky and Macskassy (2008, p.1444) for a histogram of the relationship between the number of firm-specific news stories around firms' earnings announcements.

Tetlock, Saar-Tsechansky and Mackassy (2008) deploy the negative word category of the Harvard IV dictionary unlike Tetlock (2007) who leverages multiple categories to create a pessimism factor as noted above. They find that negative words consistently predict lower earnings for Standardised Unexpected Earnings (SUE). SUE is 0.255 standard deviations lower when negative word counts rise from two standard deviations below to two standard deviations above its mean value. Thus, they state that even this 'crude' (p.1449) measure of qualitative fundamentals can predict earnings more accurately than professional analysts' forecasts. In terms of firm performance, the authors show that negative words in firm-specific news stories robustly predict slightly lower returns in shares of the mentioned firm on the following day. The coefficients of this regression imply that a one standard deviation increase in negative words translates into a 0.032% reduction in next day abnormal returns.

Evaluating the economic significance of the negative word count measure, Tetlock, Saar-Tsechansky and Macskassy (2008) create a trading strategy involving a long portfolio that consists of firms subject to positive news stories and a short portfolio consisting of firms subject to negative news stories, with portfolios balanced weekly. This strategy produces cumulative raw returns of 21.1% per year when transaction costs are ignored, and in 21 out of 25 years the strategy returns positive abnormal returns.[39]

Twedt and Rees (2012) are the last paper to be discussed that use the general Harvard IV dictionary to determine sentiment. Specifically, the authors examine the tone of analyst reports to assess whether sentiment is useful in further understanding analyst forecasts and recommendations. The authors run regressions to test two hypotheses. First, they examine whether the sentiment of financial analyst reports (that is incremental to the information contained in earnings forecasts and stock recommendations) induces market reactions. Second, they assess whether the sentiment of financial analyst reports affects the market's reaction to information contained in earnings forecasts and stock recommendations.

The authors construct a regression analysis to evaluate the above hypothesis' with results showing that a change from the lowest quartile of analyst report tone (most pessimistic) to the highest quartile of analyst report tone (most optimistic) results in an average increase in return of 0.7%, holding all else equal. This implies that investors do view the sentiment of analysts' reports as an import source of information. However, the authors conclude that the effect of report sentiment on the firm's share price reaction does not appear to be on the news contained in the analyst report sentiment as an indicator of how they should react to the report, but instead that sentiment is incrementally informative by itself. Hence, a report that is not captured in potentially biased summary outputs due to analysts' conflicts of

[39] When including reasonable transaction costs – 0.01% in this case - the strategy is no longer profitable. However, the authors suggest this could potentially be mitigated through more sophisticated trading-based rules.

interest sentiment may be used by investors to understand analysts' underlying opinion about a firm that is not captured in potentially biased summary outputs due to analysts' conflicts of interest.[40]

Davis and Tama-Sweet (2012) evaluate managers' use of language across annual 10K statements and Earnings Press Releases (EPRs) assessing whether managers are strategic in their disclosure of information. To test this question, the authors deploy the computer software DICTION 5.0[41] to assess the use of optimistic and pessimistic language in such disclosures. Evaluating the means of optimistic and pessimistic language across 10Ks and EPRs, the authors show that on average 1.08% of the words in 10Ks are optimistic and 1.27% of words in EPRs are optimistic. On average 1.01% and 0.46% of words are pessimistic for 10Ks and ERPs respectively. This implies that on average EPRs exert higher levels of optimism and lower levels of pessimism in respect to 10Ks. Davis and Tama-Sweet (2012) highlight that prior literature suggests that information contained in EPRs is processed more efficiently than that of information contained within 10Ks (Stice 1991; Louis, Robinson, and Sbaraglia 2008; Levi 2008). Hence, managers expect more substantial market reactions from EPRs and strategically tailor their choice of language towards optimistic interpretations.

Looking at the relationship between optimistic and pessimistic (correlation of 0.09) word counts in 10K filings returns a balanced dissemination of results which is noted by the authors as expected due to the 10K filling being a more comprehensive and regulated disclosure in comparison to EPRs.[42] However, a comparatively weaker correlation (-0.04) is found between the optimistic and pessimistic factors in the case of EPRs, thus showing that managers have more flexibility in EPRs and can strategically select which results they wish to include. Further evaluating whether managers manipulate word choice, Davis and Tama-Sweet (2012) find that the underlying current period and future firm performance does not drive managers' decisions to choose pessimistic language in EPRs. This suggests that managers respond to greater incentives[43] to reduce the amount of negative news reported at the time of the earnings announcement to avoid high penalties for negative surprises. Taken together, the results provide evidence consistent with managers omitting or shifting pessimistic language from their EPR when they face greater strategic reporting incentives.

Bollen, Mao and Zeng (2011) cite that emotions can profoundly affect individual behaviour and decision making. To test this hypothesis, they use two mood tracking tools - Google-Profile of Mood

[40] Twedt and Rees (2012) look to confirm economic significance through a hedge portfolio going long on analyst reports in the highest quartile of tone and short on reports in the lowest, however results proved to be insignificant.

[41] DICTION is a dictionary-based language analysis program that analyses the implied meaning of a text by searching it with the assistance of some 40 dictionaries or word lists (Given, 2008). The textual analysis software uses a series of five main dictionaries to search for sentiment features – Activity, Optimism, Certainty, Realism and Commonality – as well as thirty-five sub-features.

[42] SEC regulations require a balanced perspective of operating performance and future expectations in MD&A disclosure (Davis and Tama-Sweet, 2012)

[43] Managers generally have incentives to minimize (maximize) the stock price effects of negative (positive) news reported.

States (GPOMS) and OpinionFinder (OF)[44] - to assess whether public mood (drawn from Twitter) is a predictor of economic indicators. The time period utilised in the study (February to December 2008) includes major sociocultural events, such as the 2008 US presidential election, allowing the authors to understand public mood. Evaluating the textual content measures relationship with public mood, the authors show that both tools successfully identify the publics' initial response to the presidential election. A significant drop in the GPOMS calm measure (from 0.262 to 0.065) prior to the election implies heightened anxiety levels with significant increases in vital, happy and kind scores on election days, with a significant but short-lived uptick in public positive sentiment post-election day (from 0.085 to 0.620).

Bollen, Mao and Zeng (2011) then evaluate the relationship between the GPOMS/OF mood measures and closing values for the Dow Jones Industrial Average (DJIA). The GPOMS calm mood category is the only category to return a significant granger causality coefficient with the association being found to be strongly significant.[45] Building upon these findings, the authors create a forecasting experiment to test whether mood variables increase the accuracy of stock market forecasting. They build a Self-organising Fuzzy Neural Network (SOFNN) that initially makes predictions on the next day's up or down change in DJIA value based upon the three previous days DJIA closing values. They then add various permutations of the mood time series to this initial model, finding that when considering the calm sentiment indicator from GPOMS in addition to the previous prices the accuracy increased to 86.7%. Thus, showing that sentiment indicators can be robust in increasing market value forecasting.

The last two papers to use general dictionaries within this review are provided by Siganos, Vagenas-Nanos and Verwijmeren (2014; 2017), both of which adopt Facebook's Gross National Happiness Index.[47] The author's initial study (2014) examines the relationship between daily Facebook sentiment and trading behaviour across twenty international markets. Initially, the authors show that the sentiment measure is significantly positively related to returns. They demonstrate that an increase of 0.1 in the sentiment measure translates into a 0.031% increase in returns.

The authors then assess cross-sectional results where firms are disaggregated into small, large, growth and value categories. In line with expectations, they find that the results are strongest for small firms – a one standard deviation increase in sentiment translates into a 0.043% increase in index returns. The authors show that the relationship between sentiment and returns is stronger for value firms (regression coefficient of 0.041) in comparison to growth firms (regression coefficient of 0.028).

[44] GPOMS is a textual content measure that quantifies mood in terms of six dimensions - Calm, Alert, Sure, Vital, Kind and Happy. OF is a mood tracking tool that measures positive vs negative mood. Both textual content measures fall under the category of general dictionaries.

[47] Facebook's Gross National Happiness (FGNH) indexes the positive and negative words used in the millions of status updates submitted daily by Facebook users. FGNH has face validity: it shows a weekly cycle and increases on national holidays. (Wang, Kosinski and Stillwell, 2012).

Evaluating the causality between their sentiment measure on a given day and market returns on the following day, they again find a 0.1 increase in sentiment relates to an increase of 0.021% in returns on the following day.[48]

In their follow-up study, Siganos, Vagenas-Nanos and Verwijmeren (2017), introduce the concept of divergence of sentiment (DoS) to the finance literature. They define DoS as the difference between positive and negative sentiment each calculated through positive and negative word counts respectively. The authors introduce the DoS measure noting that a more diverging sentiment implies more diverging views on risk and prospects which in turn implies more diverging views on the value of a share. Further stating that the measure's introduction to the literature was an attempt to improve sentiment representation. In the initial analysis, the authors look at the relationship DoS has with trading volume and strongly significant relationships between DoS and both measures, suggesting that DoS is related to a contemporaneous daily increase in trading volume (2.829) and volatility (0.004) – a result which falls in line with Hirshleifer (1977), Harris and Raviv (1993) and Tetlock (2007) findings that disagreement (in this case portrayed through diverging sentiment) leads to increased trading due toing different values to an asset.

Siganos, Vagenas-Nanos and Verwijmeren (2017) were not the first researchers to attempt to more accurately capture sentiment in financial markets. In fact, a body of literature is continuously attempting to create more accurate measures or models to classify sentiment in a robust fashion. Henry (2006) introduced a specific word dictionary for finance to overcome the domain-specificity limitation[49] inherent in general dictionaries (Chan et al. 2020). Gonzalez-Bailon and Patloglou (2015) and Ribeiro et al. (2016) both show the limitations of general dictionaries' understanding of sentiment when applied to new datasets from different domains. A suggestion made by Diesner and Evans (2015) and Grimmer and Steward (2013) is to create domain-specific dictionaries, where adding words to an existing dictionary and deleting irrelevant words or words with different meanings within a specific context would be beneficial; a solution given further support by Riberio et al. (2016).

Henry (2006) creates specific positive and negative word lists through the inspection of past EPRs, rather than a general dictionary approach. Using the basic word count method, she identifies sentiment from these EPRs, among various other textual/quantitative characteristics, and evaluates these variables' relationship with the S&P 500 index. The author splits the independent variables into three categories: (i) firm characteristics, earnings information, other financials (financial information captured within earnings announcements), (ii) keywords, and (iii) writing style.[51] From these variables

[48] Siganos, Vagenas-Nanos and Verwijmeren (2014) find that sentiment also has a significant negative contemporaneous relationship with trading volume and volatility. These results show that sentiment defined from Facebook has a positive relation with stock market returns and a negative relation with trading volume and volatility – showing a causal relationship between sentiment and market characteristics, highlighting the importance of behavioural finance.

[49] Sentiment accuracy suffers when general purpose dictionaries are used for specific domains that are not well represented by general language.

[51] Category (i) constitutes financial information, whereas categories (ii) and (iii) constitute verbal content.

the author builds six different models to evaluate S&P500 forecasting accuracy of the combinations. The first model only includes firm characteristics and returns an accuracy of 56.46%. A model only using financial variables returns an accuracy of 54.12%. However, when using all variables (including the qualitative measures) the classification accuracy rises to 59.52%, showing that the inclusion of qualitative variables improves prediction accuracy by 5.4% in this case.

These results indicate that there is indeed some degree of market reaction to the verbal contents of EPRs. The relatively lower performance of financial variables could be put down to the financial information being already incorporated into share prices in comparison to the verbal components of an EPR being considered new information, giving a deeper insight into what the company feels it needs to address surrounding past and future performance.

Henry (2008) is a two-part study: first, the author applies qualitative analysis to further understand the dual purpose of EPRs: namely, their informational-promotional roles. Second, the study uses a qualitative approach to measure the reaction of the stock market to EPRs, particularly inspecting sentiment and stylistic features.

The author notes the importance and high regard EPRs are held in the investor population due to lesser regulatory requirements. This freedom for companies to discuss past results and identify future opportunities more freely in comparison to stricter disclosures is a potential reason for EPRs' popularity among the financial community. However, this freedom also gives rise to the potential dual purpose that EPRs may have: informational and promotional. The informational role, as with any other financial disclosure, is to inform readers about the company's performance, whereas the promotional role is to favourably influence readers views about firm performance.[52] In her initial analysis Henry (2006) shows that firm profitability and the length of press release are negatively related, implying that more profitable firms have shorter EPRs. These results align with the promotional aspect of EPRs with firms only wanting to highlight positive results and having less to discuss.

Henry (2008) further shows a positive association between sentiment and abnormal returns implying that companies with greater levels of positive tone within their press release experience higher abnormal returns, even after controlling for financial results.[54] Henry (2008) shows that market reaction becomes stronger the greater the positive tone produced, up until a certain point[55].

[52] Henry (2008) provides examples of the promotional role of EPRs by explaining the same quantitative results with four different explanations, showing that the same facts can portrayed by poor performing metrics or more realistically positive elements. The author then goes on to show promotional techniques used within published EPRs to draw the reader's attention to positive aspects. Promotional techniques can be considered as; bullet points at top of page using non-GAAP measures, comparisons of results to a benchmark instead of previous results, repetition of positive figures (headline, first paragraph and so on), positive evaluative comments, negative evaluative comments.

[54] Unexpected Earnings, log of the market value of the firm's common equity, an indicator variable equal to one if earnings exceed analysts' forecast, an indicator variable equal to one if earnings are greater than zero.

[55] Henry (2008) implies that past a certain point of positive tone, market reactions stop increasing. However, this specific level of tone is note defined.

Loughran and McDonald (2011) show that general dictionaries misclassify terms used within a financial context, noting that 73.8% of negative words within the Harvard dictionary are not considered negative in financial text. They build finance-specific dictionaries that include word categories relating to negative, positive, uncertainty, litigious, strong modal and weak modal words highlighting that their primary focus is the negative dictionary. To create these word lists the authors developed dictionaries of all words and their word counts relating to the above categories stemming from all 10Ks filed from 1994-2008. They then carefully examined all words that occurred in at least 5% of all documents and created final word lists based upon the top 5% most used terms in the financial documents. These word lists have been widely used throughout the literature for word count sentiment analysis approaches.

They show their newly created dictionaries' robustness over the general Harvard dictionary by comparing both dictionaries negative word lists against filing period returns. The results show that the Harvard negative word list is not significantly related to file excess returns but for the authors' new financial negative word list there is a significantly negative coefficient (t-statistic of -2.64). Hence, higher levels of negative financial words contained within the specific dictionary translates into lower excess returns.

Mao and Bollen (2011), evaluate various sentiment sources and indicators to understand the predictive value of each. Specifically, they consider surveys, news headlines, search engine data and Twitter feeds as sources to draw sentiment with Twitter Investor Sentiment (TIS),[56] Tweet Volume of Financial Search Terms (TV-FST),[57] Negative News Sentiment (NNS),[58] and Daily Sentiment Index (DSI).[59] The authors look to evaluate the sentiment measures in relation to market indices such as the DJIA, trading volumes, and market volatility (VIX), as well as gold prices.

They show that TIS returns a positive correlation with market log returns (0.267) and is negatively correlated with VIX (-0.314). The DSI is positively correlated with both DJIA closing prices (0.277) and log returns (0.181). Although it exhibits negative relationships with trading volume (-0.341) and VIX (-0.832). As market volatility is commonly known as an investors' fear gauge, negative relationships with DSI and TIS would suggest that said measures relate to positive sentiment or lower risk perception among market participants. NNS returns a positive correlation with VIX (0.237), this is in line with expectations as a negative news sentiment indicator would relate to heightened market fear and increased riskiness. In line with NNS, TV-FST also returns a positive relationship with VIX (0.183) and hence has the same characteristics. The results suggest that increases in search terms may indicate investor uncertainty surrounding investment opportunities and hence greater risk and the perception of

[56] TIS defines a tweet as bullish if it contains the term "bullish", and bearish if it contains the "bearish" then calculates overall sentiment for that day.

[57] TV-FST calculates daily tweet volumes that contain one of 26 financial terms.

[58] NNS is Loughran and McDonald's (2011) negative word lexicon applied to financial news headlines.

[59] Daily Sentiment Index (DSI) provides daily market sentiment readings on all active US markets since 1987.

higher possible losses. Furthermore, the results of Granger Causality tests for this set of sentiment indicators shows statistically significant Granger causation in both directions for log returns and TIS, NNS and TV-FST. DSI however, returned no statistically significant causation with log returns.

These results show that the presumed negative sentiment indicators (confirmed through negative correlations with VIX) possess higher forecasting ability over non-negative indicators. This finding is in line with Loughran and McDonald's (2011) evidence surrounding the success of negative word lists' in capturing financial sentiment and return significant relations with market variables. It also reiterates the robust performance of domain specific dictionaries in determining relationships with market characteristics.[60]

Continuing to use the finance specific dictionary created by Loughran and McDonald (2011), Garcia (2013) evaluates the relationship between sentiment conveyed in media articles (specifically, the New York Times' 'financial markets' and 'topics in wall street' columns) with DJIA index returns over expansionary and recessionary periods.[61] For the expansionary period, a one standard deviation change in the pessimism factor incites a market movement of 0.035% in DJIA returns. In relation to the recessionary period, a one standard deviation increase in the pessimism factor increases the DJIA by 0.012%. All tests return significant results indicating that sentiment helps predict next day stock returns. A comparison of the two periods suggests that expansionary periods are statistically different and return large economic differences - roughly three to four times stronger.

In his final analysis, Garcia (2013) tested whether the results returned were driven by information or sentiment. The results thus far are consistent with the theory that media content proxies for investor sentiment (noise traders). Garcia (2013) states that in line with psychology literature (Bless et al. 1996; Forgas, 1998; Lerner and Keltner, 2000; Park and Banaji, 2000; Gino, Wood, and Schweitzer, 2009) reactions to news will be more pronounced during periods of anxiety and fear. The recessionary analysis shows heightened market movements and hence falls in line with this statement. However, on the contrary one could interpret the metrics created by the author as new information. Thus, the new information is then incorporated into stock prices and hence induces movement.

To evaluate the potential information channel hypothesis, he looks at the returns of the DJIA after the opening of the NYSE. No predictability should be found if the results are driven by information and markets are processing said information quickly. Garcia (2013) shows that positive word counts have predictive power in recessionary periods but not in expansionary periods. The positive word counts have statistically and economically robust results – a one standard deviation increase in positive words

[60] Mao and Bollen (2011) in their last sub-study of the paper forecast the DJIA, trading volumes and VIX. They find that adding sentiment measures increases the direction accuracy of forecasting in comparison to baseline forecasts that only consider historical price as input for each: DJIA (0.5 to 0.63), VIX (0.6 to 0.67) and trading volume (0.47 to 0.60). However, the overall results are not highly significant.

[61] Expansion and Recession periods are taking from the National Bureau of Economic Research (NBER) Business Cycle Dating Committee. The NBER define a recession as the period between a peak of economic activity and its subsequent trough. An expansion is defined between trough and peak.

in expansionary periods increases DJIA returns by 0.36 standard deviations. He concludes based on these results that the informational hypothesis cannot be ruled out when information is slowly incorporated into prices but does dismiss theories where prices fully and quickly adjust to new information. To confirm the driver of his results, Garcia (2013) evaluates the relationship between media and trading volume. A sentiment theory would suggest that extreme positive or extreme negative news incites disagreement among noise and rational traders and hence induces higher levels of trading. He shows that the pessimism factor can predict trading volume. Therefore, a behavioural story in which noise traders "follow the printed word" naturally generates more volume.

Jegadeesh and Wu (2012) create a new word weighting scheme for textual analysis on 10K documents and compare it with the commonly used Term Frequency-Inverse Document Frequency (tf-idf)[62] weighting scheme. They compare their new word weighting scheme with that used in Loughran and McDonald (2011) to understand whether the weighting schemes generate similar results. They find very low coefficients with both negative (-0.052) and positively (0.138) ranked words suggesting that the way words are weighed within textual analysis critically affects measured tone.[63]

The authors compare sentiment scores compiled using their new weight scheme to filling date returns. They split their sample into deciles of firms – decile one containing firms with the highest sentiment scores to decile ten with the lowest sentiment scores. They find that for positive sentiment scores the filing period returns are 1.84% for decile one to -1.40% for decile ten. The returns for negative scores decline through each decile from 1.23% in decile one to -1.37% in decile 10. Overall, Jegadeesh and Wu (2012) correct the issue surrounding the non-comprehensive underlying lexicon's ability to fully capture tone by creating a robust term weighting scheme. They further show that creating more accurate methods for defining sentiment is beneficial for understanding relationships with market characteristics.

Ferguson et al. (2015) evaluate 264,647 firm specific UK news media articles (from The Financial Times, The Times, Gaurdian and The Mirror) concerning FTSE100 firms to assess if such media articles contain relevant informtion about future stock returns They use The Stock Sonar (TSS) to measure the sentiment of news media articles, which leverages a dictionary approach using the Loughran and McDonald (2011) word lists. The authors show statistically significant evidence that both positve and negative sentiment conveyed in UK news predicts returns on the same day as the publication. A one standard deviation increase in positive (negative) words increases (decreases)

[62] Tf-idf is a word weighting scheme that uses statistical measures to assess a word's importance within a document or a collection of documents. This process was used by Loughran and McDonald (2011) and hence has been adopted in a large proportion of the literature.

[63] They infer that their model is superior to the general model as they can define relationships with a positive word list (and subsequently positive tone) with stock market returns which academics have failed to achieve in past studies. They posit that a comprehensive lexicon is almost impossible to ensure due to the vast nature of the English language. They create a test, randomly removing 50% of the lexicon and run the regressions again. Finding that the results are not statistically different and are able to reliably quantify tone. This further gives evidence to the robustness of their measure.

abnormal returns by 0.049% (0.023%). Investigating whether firm specific news stories that receive higher levels of attention intensifies investor reactions, Ferguson et al. (2015) find that high levels of media attention[64] surrounding a positve firm specific news publication has a significant impact on next period abnormal returns. Furthermore, the authors show that the significant predictive relationship between media and next period abnormal returns is driven by less visable firms. Thus highly visable firms within the FTSE100 experience less pronoucned effecets from positive and negative words in news stories.

Extending the analysis to determine the economic significance of their sentiment varible, Ferguson et al. (2015) look at market level events. The negative sentiment variabe tracks market shocks that materialised over the sample.[65] In relation to the overall market, a one standard deviation increase in positive (negative) news increases (decreases) returns by 0.034% (0.058%). These results evidence the predictive power of media content for returns in the UK, reiterating the importance of chosing a relevent sample that is highly used by traders to make decisions to create a robust sentiment variable that can accurately track the market. Furthermore, the evidence provided in this paper gives strong statistical and economic backing to the useability of sentiment as a tool for forecasting market returns.[66]

Bannier et al. (2017) analyse performance of the Deutscher Aktien Index (DAX) in reaction to sentiment conveyed in CEO speeches given at the AGMs of German firms.[67] The authors find no significant reaction in terms of cumulative abnormal returns (CARs) around the time of AGMs. They also show that the proportion of negative words (1.03%) detected in AGMs outweights that of positive words (0.485%) in line with prior literature on english financial information sources (Loughran and McDonald, 2011).

The authors find that changes in negative and positive sentiment have a strongly significant relationship with CARs, calculated from the day before the AGM to 30 days following. An increase in negative (positive) sentiment of 0.749 (0.353) corresponds with a decrease (increase) in CARs of 2.77% (3.14%). The findings shift however when considering the immediate market reaction,[68] Bannier et al. (2017) show that negative sentiment has no significant relationship with CARs. Positive sentiment shows statisitcal significance in the immediate term, although it is shown to have quite a small association with returns in economic terms. Bannier et al. (2017) conclude that the non-significant and

[64] High media attention is defined as more than three media articles published surrounding a specific firm on a given day.

[65] The variable peaked in correspondence to the high uncertainty period in 1986, when the UK withdrew from the European exchange rate mechanism, in 2002/03 with the impeding war in Iraq and the 2007/08 financial crisis.

[66] Ferguson et al. (2015) create a news-based trading strategy using positive and negative measures of sentiment as buy and sell signals. They create a long portfolio consisting of firms from the FTSE100 that have average net positive sentiment and a short portfolio comprised of firms that have net negative sentiment. Over the period 2003-2010 they returned 0.012% per day, resulting in a significant alpha when all transaction costs were ignored.

[67] To calcualte sentiment in the German language the authors translate the Loughran and McDonald (2011) word lists and focus them towards speeches given at AGMs.

[68] Evalauted over the three days surrounding the AGM (t-1 to t+1).

economically weak relation around the time of the AGM suggests an initial underreaction to qualitative information compared to the strong results seen in the longer period.

Adding to the research of Tetlock (2007) and Ferguson (2015), Johnman, Vanstone and Gepp (2018) evaluate sentiment conveyed in articles published in The Guardian newspaper in the UK, where sentiment is defined by Loughran and McDonald's (2011) finance-specific word dictionaries to investigate its impact on FTSE100 stock returns and volatility. The authors find that sentiment has no significant relationship with daily excess returns but appears to influence daily volatility. Negative (positive) sentiment corresponds with increases (decrease) in volatility. From these findings, the authors suggest that retail investors (defined as readers of The Guardian) do not have any meaningful effect on returns but do impact volatility. These findings agree with the model of De Long et al. (1990) and suggest that noise traders and rational arbitrageurs have different effects on a financial market – noise traders introduce noise to the markets.

From the above results Johnman, Vanstone and Gepp (2018), take a long position on a stock if the previous days negative sentiment value is greater than the 70th percentile of last year's average negative sentiment value. Excluding transaction costs (from 2002-2016) the sentiment strategy yields a greater return (0.061%) and Sharpe ratio (0.330) than a basic buy and hold strategy return (-0.007) and Sharpe ratio (-0.034). Overall, the findings demonstrate that sentiment measures drawn from sources that are more likely to be read by retail investors,[69] in comparison to sophisticated investors,[70] impact asset prices differently. The authors state this may be because retail investors do not add information to the market but increase volatility, hence deviating prices from their fundamental values.

The last paper in this review to incorporate Loughran and McDonald's (2011) word dictionaries is Jiang et al. (2019) who leverage the word lists to create a manager sentiment index[71] to forecast future aggregated S&P 500 index market returns. Initially, the authors show that their sentiment measure has greater predictive power over S&P 500 returns in comparison to multiple macroeconomic variables.[72] Jiang et al. (2019) show that their manager sentiment index has a negative and significant relationship with index returns – a one standard deviation increase in sentiment relates to a 1.26 standard deviation decrease in S&P 500 returns. A high manager sentiment is associated with low excess aggregate market returns in the next month. The authors hypothesize that when the manager sentiment index is high, market wide overvaluation occurs consequently leading to low future stock returns.

[69] Retail investors, also known as "noise traders", are market participants that hold random beliefs about future dividends.

[70] Sophisticated investors are market participants who hold Bayesian beliefs – assess future dividends based upon probability theory.

[71] The monthly manager sentiment index is created by aggregating manager sentiment from 10Ks, 10Qs and conference call transcripts from 2003-2014.

[72] See Goyal and Welch (2008: 1457) for the 14 macro-economic indicators used and their definitions.

The authors show that an overall manager sentiment index consistently beats all individual sentiment measures.[73] Further comparing their manager sentiment index to investor sentiment indexes the authors find that manager sentiment does not lead investor sentiment and vice versa. These findings indicate that manager sentiment and investor sentiment capture different subsets of sentiment information, and they are complementary in measuring market sentiment. Thus, manager sentiment has strong negative forecasting power for stock market returns. Jiang et al. (2019) conclude that the predictability found holds both in and out of sample showing its potential to generate economic value for investors.

The literature mentioned thus far has discussed studies employing early sentiment analysis techniques, from general dictionaries to more specific dictionaries capturing context-specific sentiment. Research in Natural Language Processing (NLP), linguistics and the wider computer science field however has consistently shown from the 1950s onwards that it is possible to use Machine Learning (ML) algorithms and statistical procedures to measure the properties of text and extract information automatically (El-Haj et al. 2018). ML algorithms learn from a training set of thousands of examples to find relationships, develop understanding and make decisions on future unseen cases. El-Haj et al. (2018) identify that the field of accounting and finance falls behind that of NLP studies in classification of sentiment using ML best practice. The studies contained in the following subsection are those that have applied ML for sentiment classification in a financial context.

2.2.2 Machine Learning Approach

Evaluating the relationship between internet message board contents impact on the DJIA Index and Dow Jones Internet Commerce Index (XLK), Antweiler and Frank (2004) stands as one of the first studies to return a significant result for forecasting market variables using sentiment metrics not calculated using a dictionary approach. Antweiler and Frank (2004) adopt a naive approach to assessing whether message boards contain relevant financial information. In particular, the authors address (i) whether the number of messages posted, or the bullishness of messages help to predict returns/volatility and (ii) whether disagreement among messages is associated with higher levels of trading. Using contemporaneous regressions, Antweiler and Frank (2004) find that a one standard deviation increase in bullishness translates into a 1.75 standard deviation increase in returns. This result is insignificant when time is lagged. They also show that higher levels of message posting and/or bullishness one day translate into significantly higher levels of market volatility the next day for both measures. These effects are shown to flow in both directions; however, the flow is stronger from messages to volatility.

[73] Jiang et al. (2019) compare their monthly manager sentiment index (footnote 36) with: Baker and Wurgler (2006) investor sentiment index, Huang et al. (2015) aligned investor sentiment index, University of Michigan consumer sentiment index, conference board consumer confidence index and Da et al. (2015) Financial and Economic Attitudes Revealed by Search (FEARS) investor sentiment index. See Jiang et al. (2019: 131) for definitions of each index.

Traditional analysis (Hirshleifer, 1977 and Harris and Raviv, 1993) and sentiment analysis (Tetlock, 2007; Siganos et al. 2017; Garcia, 2013) studies infer that higher levels of disagreement induce higher levels of trading volume. Antweiler and Frank's (2004) results show that greater levels of agreement in a period result in fewer trades in the same period. The authors provide evidence showing that a one standard deviation increase in agreement on a given day leads to a 0.142 standard deviation increase in trades on the next day. This finding is consistent with Cao et al. (2020), who suggest that individuals are more likely to trade when they know others have received the same signal as them. As this agreement takes time to be revealed, they note that higher levels of agreement today will result in more trades the following day.

Sprenger et al. (2013) also use a Naive Bayes classifier to determine Twitter sentiment and compare it with S&P 100 index returns, trading volume and volatility. The authors initially show a strong correlation between Twitter message volume and trading volume. Specifically, a 1% increase in message volume translates into a 10% increase in trading volume. This would suggest that individuals discussing opportunities often pursue these opportunities and potentially are able to convince others to invest. Furthermore, Sprenger et al. (2013) find a strong relationship between sentiment and S&P 100 returns. They show that increased positive sentiment in tweets is associated with rising stock prices. They note that no relationship between message volume and returns is found. However, an increase in volatility is observed as message volume rises, suggesting that uncertain investors may exchange information and consult their peers more than those who are less uncertain. Though Sprenger et al. (2013) find no support for the argument that disagreement amongst tweets drives market volatility, a negative correlation between agreement and trading volume is reported (-0.113).

Sprenger et al. (2013) also show that their sentiment measures cannot be used to predict returns. However, the effect of returns on sentiment is positive and significant. Hence, returns affect sentiment, but not vice-versa. They further show that message volume one day and two days prior are good indicators of trading volume with significant regression coefficients of 0.189 and 0.120 respectively. At the same time high trading volume triggers increased message volume over the following days.

Grob-Klubmann and Hautsch (2010) adopt the Reuters NewsScope Sentiment Engine (RNSE) to retrieve 29,497 news headlines with accompanying sentiment and relevance indicators. The authors look at the unconditional effects of news items - particularly the impact on volatility and liquidity and identify significant upward movements in money value traded, average trade sizes and volatility surrounding the release of news items. that volatility and trading activity increase when news items are published.

Blume et al. (1994) argue that higher volumes of media reflect a higher quality of news signal. Grob-Klubmann and Hautsch (2010) find that machine-indicated relevance of news is supported by market reactions. Hence, there is a significantly stronger reaction to news if the news has been ranked with high relevance. This implies that selection of relevant news (or more generally relevant information) is crucial in understanding market responses. Evaluating the difference in reaction to initial news and subsequent updates they find that trading on updated news is much more pronounced than trading on initial news, supporting the notion of news clustering and showing reiteration and reinforcement of news creates stronger signals which translate into stronger market reactions. Grob-Klubmann and Hautsch (2010) evaluate the economic relevance of sentiment by creating a trading strategy that buys at the best ask price and sells later at the bid for positive news items. For negative news items, the asset is sold at the best bid price and re-bought later at the ask price. They observe that sentiment indicators of news items have some predictability for future price movements.[74] However, they find abnormal returns to be mostly insignificant.

In a similar fashion to Grob-Klubmann and Hautsch (2010), Sun, Najand and Shen (2016) adopt the RNSE and evaluate sentiment at the intraday level for S&P 500 index returns. They use a dataset that not only considers news sentiment but also SEC filings, social media and earnings calls sentiment. Initially, Sun et al. (2016) show evidence that their lagged sentiment measure (split into half hour periods across the day) is a robust predictor of last half hour intraday returns. A one standard deviation increase in sentiment results in a 0.269 standard deviation increase in returns in the last half hour. However, when the authors limit their focus to recessionary periods[75] the strength of lagged sentient being a predictor of last half hour returns weakens.[76] A one standard deviation increase in sentiment over recessions results in a 0.216 standard deviation increase in returns.

Testing the economic relevance of their sentiment measure, Sun et al. (2016) take a long position in the S&P500 ETF if the regression results for returns in a particular half hour are positive and a short position otherwise. They find that their purely sentiment-driven model (Sharpe ratio of 1.28 and mean return of 8.34%) slightly outperforms a model only considering lagged returns (Sharpe ratio of 1.26 and mean returns of 8.17%). Both outperform the benchmark model[78] used by Sun et al. (2016). Investigating the driving force behind their sentiment variables predictability the authors find it is driven by the actions of noise traders who are more susceptible to shifts in sentiment.

Audrino and Tetereva (2019) evaluate sentiment spillover effects using graphical granger causality, focussing specifically on whether news sentiment, defined using RNSE, has cross-industry effects for the S&P 500 and Euro Stoxx 50 indexes. US and European firms are disaggregated according to the primary industry in which they operate (using a list of ten industries). The authors note that the relevance of news stemming from differing sectors shows fluctuating effects on returns that are spread evenly among industries. However, there is evidence of finance and energy news holding a greater

[74] Their trading strategy returns 0.035% annually but is not strong enough to return economic gains.

[75] National Bureau of Economic Research date recessions from March 2001 to November 2001, and from December 2007 to June 2009.

[76] Even still, during the recessionary period the variable is significant at the 10% level. Sun et al. (2016), returned a 3.6% R^2 with the inclusion of their sentiment variable.

[78] The benchmark model is based on the sample mean. Sun et al. (2016: 157) note that the benchmark strategy is equivalent to the case where only the constant term is included in the predictive regression. Thus, from a Bayesian perspective, investors who have a dogmatic prior belief that none of the predictors are useful should implement this benchmark strategy.

influence across all sectors. These influential sectors, have spillover effects that seem to be at least as important as the direct effects of their sentiment.

Furthermore, over periods of economic instability the impact of these spillover effects is intensified, which is broadly supportive of the earlier findings of Garcia (2013) who show that periods of heightened anxiety make investors more receptive to advice. Hence, any information disseminated by the media is more likely to be acted upon over these periods even if the information is in relation to a different sector. These results suggest that selection of relevant information to create a robust sentiment indicator may come from sources not directly linked to a specific company or index.[79]

Azar and Lo (2016) apply machine learning techniques utilising De Smedt and Daelemans (2012) 'pattern' python package.[80] The researchers focus on a dataset of Tweets that mention terms that are related to the Federal Open Markets Committee (FOMC), such as 'FOMC', 'Federal Reserve', 'Bernake' or 'Yellen' on the basis that decisions made by the FOMC are popular among the investment community and significantly affect asset prices (Bernanke and Kuttner, 2005; Cieslak, Morse, and Vissing-Jorgensen, 2014; Lucca and Moench, 2015). Azar and Lo (2016) create a sentiment index to understand the influence of sentiment conveyed in social media discussion of FOMC on asset prices.[81] The authors find that tweet sentiment can be used to predict a day ahead returns, with the effect intensifying on days when the FOMC meet. This implies that on days when the FOMC meet, tweets in relation to the meeting have predictive power for the next day's returns, controlling for Fama-French variables. A one standard deviation increase in tweet sentiment on FOMC days results in an increase of 0.58% in returns the following day. However, tweet sentiment on days that the FOMC do not meet become negligible for forecasting when Fama-French factors are included. Creating a trading strategy that uses tweet sentiment on days in which the FOMC meet, Azar and Lo (2016) show that their model passively tracks the CRSP value-weighted index except for eight days a year (when the FOMC meets) and significantly outperforms the market benchmark with a one-year period return of 22%.

Gu and Kurov (2020) set out to investigate Bloomberg's Twitter sentiment forecasting ability of Russell 3000 index market characteristics.[83] They show that Twitter sentiment has a statistically significant contemporaneous correlation of 0.14 on average with stock returns, further showing that the lowest 10% of sentiment firms have the highest average volatility, abnormal volume, bid ask spread and firm size than the top 10%. These findings fall in line with Engle and Ng (1993) who document that bad news tends to have more effect on stock return volatility than good news. Forecasting stock returns Gu and Kurov (2020) find that, on average, the stock return over the next 24 hours for firms

[79] Grob-Klubmann and Hautsch (2010) and Ferguson (2015) both also note the selection of relevant information.

[80] This package assigns a polarity score to a given text input based upon the SentiWordnet annotated dictionary. Each word within the SentiWordnet dictionary is assigned a triplet of numbers measuring its positivity, negativity and objectivity (Azar and Lo, 2016).

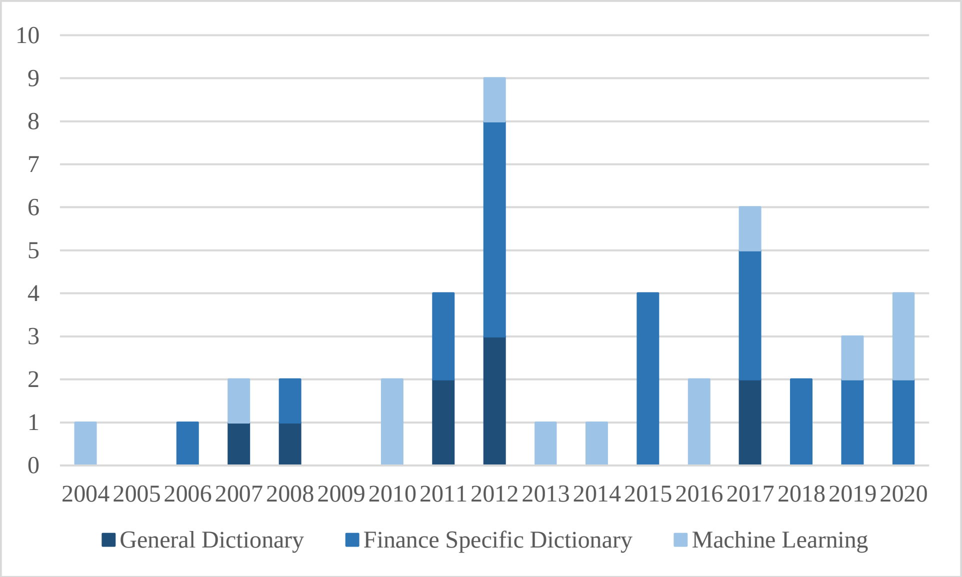
[81] They highlight that anyone can participate in conversations surrounding asset prices online and hence evaluation of said data may provide little to no information.

[83] Stock returns, trading volume, volatility, market capitalisation and the bid as spread.

with positive sentiment is roughly 0.027% higher than the return for firms with the most negative sentiment. Analysing the impact of sentiment on next day returns for a value-weighted index and an equal weighted index the authors show that the coefficient estimate for sentiment is much smaller for the equal weighted index (0.048) in comparison to that of the value weighted index (0.136), hence identifying that Twitter sentiment does have more predictive power for returns of small firms relative to large firms.

They then evaluate whether twitter sentiment has a predominant effect on stock returns. If the sentiment measure contains insightful fundamental information about stocks, its effect on returns should be permanent. However, if there are reversals the information provided could be produced by uninformed traders. Controlling for different time lags the sentiment measure contains some new useful fundamental information not yet incorporated into prices. The coefficient estimates on the lags of the sentiment measure are small and statistically insignificant. These findings indicate that Twitter sentiment contains value relevant information that has not yet been incorporated into stock prices. Finally, Gu and Kurov (2020) create two portfolios at the start of each trading day, going long on firms with high positive sentiment and short on firms with negative sentiment - rebalancing at the beginning of every day. Ignoring transaction costs this strategy returns an average of 0.086% daily which translates into 21.5% annual return with a Sharpe ratio of 3.17.

Figure 2.1: Frequency of Published Studies Applying Sentiment Analysis Methods to Financial Data



Notes: This figure shows the number of published studies within academic finance that have utilised sentiment analysis techniques to investigate associations between financial sentiment and trading activity. The annual frequency is broken down into three categories, based on the specific technique used to derive sentiment.

As discussed, financial disclosures have been subject to a steady stream of sentiment analysis literature within recent years. Figure 2.1 illustrates this trend over time by providing an overview of the number of published studies utilising different sentiment analysis methods, disaggregated by publication year for all articles referenced within this review, with finance-specific dictionaries and machine learning methods gaining in popularity in recent years.

2.2.3 State-of-the-Art Natural Language Processing Approaches

The previous sections of this paper discuss the application of dictionary and machine learning approaches to the financial domain and highlight that, as the techniques used to define financial sentiment increase in complexity, so too does the accuracy of the captured sentiment. However, El-Haj et al. (2018) identify that the field of accounting and finance falls behind that of NLP studies in the classification of sentiment using state-of-the-art methods. They note that there is a scarcity of advanced NLP techniques being applied in the financial domain.[84] While the ML techniques discussed in the previous section have been shown to classify financial sentiment better than more rudimentary approaches, alternative approaches such as transformer architecture (Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Alamoudi and Alghamdi, 2021) and multimodal analysis (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021) have been demonstrated as having greater abilities in accurately capturing sentiment.

2.2.3.1 Transformer Architecture

Before the introduction of the transformer by Vaswani et al. (2017), the authors highlighted that state-of-the-art results across various NLP tasks were dominated by Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) (Hochreiter and Schmidhuber, 1997; Chung et al. 2014). Instead of attempting to push state-of-the-art results by improving on previous RNN language models or Encoder-Decoder architecture (Jozefowicz et al. 2016; Wu et al. 2016; Luong et al. 2015), Vaswani et al. (2017) introduced the transformer, which is a model that is solely based on attention mechanisms and does not require recurrence or convolutions. Due to the complexities of transformer architecture, we do not provide a comprehensive overview of model architecture within this paper. However, due to the common implementation of this computation model, multiple in-depth descriptions of the model exists such as, Vaswani et al. (2017).[85]

Many models since the introduction of transformer architecture have returned state-of-the-art results in various tasks by adopting and building upon the initial method. For example, Raffel et al. (2020) introduced a text-to-text transfer transformer (T5), which has achieved state-of-the-art performance on the SQuAD question and answering task.[86] Brown et al. (2020) train an autoregressive language model (GPT3) on 175 billion parameters,[87] which returned the highest accuracy of 86.4% on

[84] El-Haj et al. (2018) cites that a potential reason for the lack of use of advanced NLP financial sentiment analysis models is the lack of substantial domain relevant datasets for training and testing.

[85]For a more detailed overview of transformer architecture, also see “The annotated Transformer” by Vaswani et al. (2022), available at http://nlp.seas.harvard.edu/annotated-transformer/.

[86] The SQuAD Q&A task presents a model a paragraph with a question about the paragraph. The goal of the model is to effectively answer the question posed. The answers to the questions give insight to how well a model can understand text.

[87] ChatGPT is built upon a variant of this model.

the LAMBADA language modelling task.[88] A third model which effectively utilises the transformer is Bidirectional Encoder Representations from Transformers (BERT). Devlin et al. (2019) introduce BERT and compare it to various other advanced models across multiple datasets. The authors show that the general BERT model, not pretrained on any specific data only finetuned towards the specific tasks performed competitively (80.5% accuracy, representing a 7.7% absolute improvement on GLUE).[89]

Across the three models discussed above, BERT performs particularly well on sentiment classification tasks (Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Alamoudi and Alghamdi, 2021). However, the only paper to the authors knowledge to use BERT in the financial domain is Hiew (2019), who applies BERT to posts on the Chinese social media platform Weibo relating to three listed firms on the Hong Kong Stock Exchange (HKSE) – Tencent, Ping An and CCB. The author compares their method with commonly used machine learning methods within the established literature,[90] finding that BERT vastly outperforms each of the comparison models on three key criteria.[91] These findings support the suggestion that BERT demonstrates stronger capabilities of financial sentiment classification in the Chinese language over commonly ML models.

Similar to traditional dictionary approaches, Howard and Ruder (2018) show that transformer model performance for text classification can be significantly improved when further pretrained on a domain-specific corpus. Yang et al. (2020) created a financial version of BERT named FinBERT that is pretrained on 4.9 billion tokens from three financial corpora: earnings conference call transcripts, annual reports and analyst reports. They compare FinBERT to BERT across three different financial sentiment analysis tasks; Financial Phrase Bank,[92] AnalystTone[93] and FiQA.[94] Finding that the model pretrained on general language, BERT, does not perform as well as the FinBERT model pretrained on financial language.

Transformer architecture is pushing the capabilities of machines in understanding text, however, as with all methods discussed previously in this review, it has limitations. Khan et al. (2022) provide a survey of the application of transformers for computer vision. They highlight the main limitations of transformers being high computational cost due to their size and complexity, large data requirements

[88] The LAMBADA dataset tests a model’s ability to handle long-range dependencies in text. The task requires a model to predict the last word of a sentence based upon a context paragraph as input.

[89] The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training and evaluating NLP models.

[90] They compare with a RNN based Bidirectional LSTM, BiLSTM (see Hochreiter and Schmidhuber, 1997), the Multichannel Convolutional Neural Network (CNN) (see Kim, 2014), the CPU-efficient FastText (see Joulin et al. 2016) that is adopted by Facebook, and the Transformer with attention mechanism (see Vaswani et al. 2017).

[91] Precision score is the number of positive class predictions that belong to the positive class. Recall score is the number of positive class predictions from all positive examples in the dataset. F1 score is a measure of balance that concerns both precision and recall in one number.

[92] A publicly available financial sentiment dataset consisting of 4,840 sentences from financial news.

[93] A dataset consisting of 10,000 sentences labelled with positive, negative or neutral sentiment taken from analyst reports.

[94] An open challenge dataset consisting of 1,111 sentences annotated for financial sentiment.

due to their need for a substantial amount of good quality data for training and the models' poor interpretability due to their complex architectures.

2.2.3.2 Multimodal Analysis

Most sentiment analysis techniques, across all subject fields of academic research, have mainly used singular modality-based models - in the most part text-based classifiers, which have been shown to be useful for various tasks such as forecasting box office revenues (Asur and Huberman, 2010), election outcome prediction (Tumasjan et al. 2010), classifying customer reviews (Grabner et al. 2014) and – as the afore-mentioned literature suggests – stock market prediction.

More recently, audio and visual data have also been used singularly to identify sentiment. The analysis of speech for emotion classification has been researched extensively (Koolagudi and Rao, 2012). Multiple studies have evaluated the ability of vocal cues to define sentiment alone, showing that audio cues can successfully define sentiment (Mairesse, Polifroni and Fabbrizio, 2012; Mayew and Venkatachalam, 2012; Pérez-Rosas and Mihalcea, 2013; Kaushik, Sangwan and Hansen, 2013; Pereira, Luque and Anguera, 2014). Soleymani et al. (2017) note that the field of sentiment analysis using visual data alone has not been fully researched. Indeed, the sole research to the author's knowledge in this area is Borth et al. (2013).

Soleymani et al. (2017) highlight that recent developments in this branch of natural language understanding have started to consider the combination of modalities.[95] Multimodal sentiment analysis can be defined as the inclusion of additional modalities (audio and/or visual) to compliment text-based models to improve sentiment classification. Extending the input data of sentiment classifiers is gaining traction due to its usefulness in assessing sentiment on a plethora of publicly accessible multimodal data platforms such as Facebook, YouTube, Reddit, and Twitter (Ghandi et al. 2023). The gold standard for multimodal sentiment analysis is considered to be the combination of all three communication modalities – text, audio and visual. Various studies have used the combination of all three modalities to define sentiment, showing that the use of a tri-modality model is more robust at classifying sentiment over bi-modal and singular modality models (Morency, Mihalcea and Doshi, 2011; Houjeij et al. 2012; Wollmer et al. 2013; Bhaskar, Sruthi and Nedungadi, 2014; Poria, Cambria and Gelbukh, 2015; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021).[96]

The main advantage of using multimodal classifiers for sentiment classification is the additional behavioural cues provided by the visual and audio data. The insights that vocal and visual data provide are substantial and allow for a more robust sentiment to be captured. However, there are limitations of

[95] However, the authors do note that this branch of sentiment analysis, although promising, is still in its infancy.

[96] Poria et al. (2015) further evaluate the accuracy of combinations of two modalities against the trimodal model and each modality on its own. The authors find that all dual combinations of data outperform all singular modality models. They highlight that, absent the trimodal model, a combination of visual and audio data performs the best then textual and visual data next best and finally textual and audio data performing the worst out of all pairs of modalities but still better than any singular modality model.

multimodal sentiment analysis, particularly in its application to the financial domain. The limitations come in the form of access to multimodal data, the adhesion of the different modalities into a successful classifier, and the generalisation of multimodal models. The application of multimodal analysis in finance poses issues due to the lack of substantiated data sources that contain more than one modality – the only dual modality reliable data source in finance to the authors' knowledge is earnings conference calls.

Ghandi et al. (2023) highlight there are various methods to fuse together the different modalities in a multimodal classifier. In their review of sentiment literature, it is evident that there is a substantial lack of analysis of text-audio multimodal classifiers and subsequently a lack of consensus on the best way to extract and fuse together these two modalities – this is evidently an obstacle when evaluating earnings calls as the two modalities stemming from said disclosure are text and audio. Finally, multimodal sentiment analysis models do not generalise well. If a model has been trained on a specific person/or set of people and learned their behavioural cues, the results of the model on another set of people do not scale to true generalisation.

2.2.4 Comparison of Approaches

Above I have discussed papers employing dictionary, machine learning and deep learning approaches to sentiment classification across various financial information sources. However, thus far this literature review has not focussed on the comparative performance of discussed methods. There have been multiple papers that compare methods and sources to assess the most optimal way of defining sentiment. For example, Kearney and Liu (2014) survey the various information sources and analysis methods that have been utilised in relevant literature. Assessing the usefulness of financial disclosures, media content and online media content, Kearney and Liu (2014) state that each have their own distinct advantages and disadvantages. For example, financial disclosures are valuable information sources as information comes directly from insiders who know most about their firms. They also not only produce content relating to past matters but forward-looking content. However, the authors note that management may not always 'tell the truth, the whole truth and nothing but the truth' (p.174).

The authors note that media articles that discuss general economic or market wide events are an appropriate choice from which to develop a measure of sentiment for studying market level characteristics. However, unlike disclosures, media articles mostly discuss information in hindsight which potentially limits its forecasting power. Moreover, as online media is unregulated and open to all, sentiment drawn from these sources is likely to contain little new information that is relevant to the market. Kearney and Liu (2014) state that a substantial amount of the information produced online is from noise traders. Information posted to unregulated online media is less likely to be accurate, reliable or contain new value relevant information therefore having a higher noise to information ratio. Hence, testing for market efficiency with this source that is more aligned with small investor sentiment is not ideal. However, the use of this source for behavioural finance may be beneficial as behavioural finance aims to understand the financial decision-making processes of investors at the individual level to interpret the overall market environment. Overall, Kearney and Liu (2014) suggest that a combined approach using multiple sources could provide the most robust sentiment indicator.

Kearney and Liu (2014) then discuss the benefits and drawbacks of the dictionary-based approach and machine learning approaches. The authors suggest that for ease of use and cost benefit the dictionary approach is most optimal. However, they note that general dictionaries are not robust in a financial context. They further remark that the choice of a specific weighting scheme dictates the accuracy of this process. In comparison, ML models are generally more accurate in classifying sentiment, although the ML approach is more time consuming and costly in comparison.

McGurk, Nowak and Hall (2020) build upon these findings by comparing the robustness of a sentiment index created through Loughran and McDonald’s (2011) dictionary approach with an index created from Taddy’s (2013) tokenization approach.[97] They show the tokenization index sentiment[98] and dictionary produced sentiment[99] are related to abnormal returns. McGurk et al. (2020) further show that their index is more capable of forecasting returns for the following day than the Loughran and MacDonald (2011) index. A one standard deviation increase in their unigram positive, neutral and negative sentiment measures translates to an abnormal return change of 0.013%, -0.428% and 0.337% respectively. A one standard deviation increase in their bigram positive, neutral and negative sentiment measures translates into an abnormal return change of 0.061%, -0.442% and 0.177% respectively. However, the dictionary approach does not return any significant results.

Renault (2017) explore online investor sentiments and intraday stock returns relationship by looking at the ability of first half-hour market sentiments to forecast the last half hour stock returns. Drawing sentiment from stocktwits they compare their own specific dictionary[100] with a default parameter machine learning classifier (M1). They find that the accuracy for classifying sentiment for their weighted lexicon (L1) is 74% and for the non-weighted lexicon (L2) is 76%. However, when looking at the percentage of classified messages the weighted lexicon classifies 90% of messages in comparison to the substantially lower 61% of the non-weighted lexicon. The LM and Harvard dictionaries only return accuracies of 63.06% and 58.29% respectively. A more striking figure is that LM only classifies 26% of messages. The machine learning approach (M1) returns 75% classification accuracy which is slightly higher than that of the weighted dictionary and classifies all messages.

[97] This is a statistical sentiment analysis approach that unlike the dictionary approach does not require specific words to be categorised as positive or negative. Rather it uses manually labelled text to identify relevant tokens McGurk et al. (2020). The authors analyse both unigrams (one word) and bigrams (two words at a time) to identify sentiment.

[98] The regression coefficients for the unigram sentiment variables positive, neutral and negative estimating abnormal returns are 0.621, 0.684, 0.691. The regression coefficients for the bigram sentiment variables positive, neutral and negative estimating abnormal returns are 0.308, 0.797, 0.120.

[99] The regression coefficient for the dictionary-based approach estimating abnormal returns is 0.003.

[100] Renault (2017) create their own stocktwitts specific dictionary following a process like how Loughran and McDonald (2011) created their dictionary. This dictionary is used in two different ways; a weighted dictionary approach (L1) and an equally weighted dictionary approach (L2).

Findings show that the first half-hour change in investor sentiment predicts the last half-hour stock market return. A one standard deviation increase in the ML model, L1 and L2 dictionaries sentiment in the first half hour of the day predicts an increase in last half hour returns of 0.0273%, 0.0274% and 0.0227% respectively.[101] No predictability was found when using the LM and the Harvard dictionaries.

In his next study Renault (2020) notes that more complex and time-consuming machine learning methods do not necessarily increase the performance of sentiment classification. The author concludes that simple machine learning algorithms such as Naive Bayes and Maximum Entropy, may be sufficient in deriving textual sentiment from online sources. Hence, suggesting that more complex algorithms do not necessarily equate to more accurate results for online sources.

In an earlier paper, Das and Chen (2007) evaluate the accuracy of five different algorithms for classifying messages drawn from the Yahoo! Message board. They incorporate the rainbow algorithm of McCallum (1996) as a benchmark for classification. This model works particularly well in sample as it does not have a fixed lexicon. Instead, it creates a lexicon based upon words in the training set that are the discriminants.[102] Looking at in-sample accuracies the Rainbow algorithm performs best with 97% accuracy with the following returning weaker results: Naive Classifier (92.25%), Vector Distance Classifier (49.20%), Discriminant-Based Classifier (45.72%), Adjective-Adverb Phrase Classifier (63.37%), and Bayesian Classifier (60.70%). However, when looking at out-of-sample data the abovementioned algorithms[103] perform better, particularly the Discriminant-Based Classifier which returned the strongest result with 40.6049. However, and more importantly as highlighted by the authors, the machine learning models return lower numbers of false positives which are the errors that impact results the most.[104]

The relationship between a sentiment measure[105] and constituent firms of the Morgan Stanley High-Tech Index (MSH) primarily operating in the Technology sector. The results show that the sentiment index had a correlation of 0.48 with the MSH returns, implying their sentiment measure is not excessively noisy.

[101] Renault (2017) creates a trading strategy based upon the first half hour sentiment being predictive of the last half hour returns buying (selling) the S&P 500 ETF at 3.30 p.m. on days with an increase in novice investor sentiment during the first half-hour of that day, and selling (buying) at 4:00 p.m. They find that the average annualized return of a strategy using half-hour change in novice investor sentiment as a trading signal is equal to 4.55%, with a Sharpe ratio of 1.496.

[102] Terms that are influential in defining sentiment.

[103] Naive Classifier, Vector Distance Classifier, Discriminant-Based Classifier, Adjective-Adverb Phrase Classifier, and Bayesian Classifier.

[104] The authors note that false positives 'doubly' (p.1381) impact results because sentiment is incremented by the wrong sign. Hence, false positives are more costly than other errors.

[105] This sentiment measure is calculated using a voting classifier that considers all of the models used in their comparison of models' section - noted in footnote 64.

Similarly, Guo, Shi and Tu (2016) evaluate many of the machine learning models used in Das and Chen (2007) neural network[106] along with the LM dictionary. They find on the Thomas Reuters News Archive database that neural networks perform the best with in-sample classification accuracies of 99.85%. In an out-of-sample analysis the authors also show that their neural network outperforms a Naive Bayes algorithm by 20% and an LM dictionary approach by 5% with a 79.6% accuracy. These results using a neural network are higher than that of any comparison study discussed within this review. Therefore, the application of neural networks seems to be more robust at classifying news sentiment in comparison other machine learning methods used within the literature.

In a more recent comparison, Jayaraman and Dennis (2020) use earnings call sentiment from 1,200 NASDAQ firms across an 18-month period beginning in 2017 to predict stock price change post earnings announcement. To classify the sentiment of these calls, the authors deploy Loughran and McDonald's (2011) positive, negative and uncertain categories separately and then use each of the three sentiment indicators as inputs for three machine learning classifiers.[107] The authors find that the logistic regression (used as a baseline classifier) returned an accuracy[108] of 55%, whilst the SVM and RF models performed better with classification accuracies of 66% and 73% respectively. In a further analysis they find that earnings call sentiment has similar predictive power in comparison to earnings surprise and revenue surprise variables. However, the authors note that the combination of the sentiment and surprise features increases next day share price movement classification to 78%. These findings add to the consensus of prior literature which indicates that earnings call sentiment has predictive power in forecasting share prices.

Highlighting the fact that most of the extant finance literature uses basic sentiment analysis approaches and that researchers have not yet implemented recent advancements in NLP, Hiew et al. (2019) adopt a state-of-the-art technique developed by Google named Bidirectional Encoder Representations from Transformers (BERT). Specifically, the authors use BERT to create a financial sentiment index that's aim is to predict future stock returns. They apply the BERT model to posts on the Chinese social media platform Weibo relating to three listed firms on the Hong Kong Stock Exchange (HKSE) – Tencent, Ping An and CCB. The authors compare their deep learning method with commonly used machine learning methods within financial sentiment literature.[109] BERT far outperforms each of the comparison models on three key criteria, scoring 79.3 on precision,[110] 75.4 on

[106] A back propagation neural network (BPN) containing one hidden layer is used by the authors. For a breakdown of the optimised weights for each node see p.165.

[107] Random Forrest (RF), Support Vector Machine (SVM) and Logistic Regression.

[108] Accuracy in Jayaraman and Dennis (2020) paper is defined as a models ability to predict the direction of movement of stock prices.

[109] They compare with a Recurrent Neural Network (RNN) based Bidirectional Long Short-Term Memory (BiLSTM) (see Hochreiter and Schmidhuber, 1997), the Multichannel Convolutional Neural Network (CNN) (see Kim, 2014), the CPU-efficient FastText (see Joulin et al. 2016) that is adopted by Facebook, and the Transformer with attention mechanism (see Vaswani et al. 2017).

[110] Precision score is the number of positive class predictions that belong to the positive class.

recall[111] and 78.5 on F1 score.[112] All other models return results no more than 77.6 on precision, 64.8 on recall and 71.3 on F1. Hiew et al. (2019) note that these findings give confirmation that BERT demonstrates stronger capabilities of financial sentiment classification in the Chinese language over commonly ML models.

Evaluating the BERT sentiment index (BSI) return predictability, Hiew et al. (2019) compare their index with the option implied financial sentiment index of Han (2009) (OSI)[113] and the market implied financial sentiment of Baker and Wurgler (2006) (MSI).[114] Adding these sentiment indicators to fundamental market factors[115] and forecasting for stock returns the authors find that, for each stock, a combination of all three sentiment indexes leads to a better prediction as almost all have lower Mean Squared Errors (MSE)[116] than each index on their own. Not considering the combination model of all three, BSI outperforms the other two indexes for the majority of the 2016-2018 period and across the three firms used in the study. Overall, Hiew et al. (2019) provide evidence to suggest that BERT, for the authors small sample of three firms on the HKSE, returns better classification accuracy than most popular ML methods. This analysis suggests that an evaluation of BERT against commonly used methods is required on a larger dataset to assess whether these results hold.

Overall, the research conducted using sentiment analysis has a clearly defined path, one that is following the footsteps of the NLP and linguistics literature: techniques used to define sentiment are increasing in complexity to capture the most robust sentiment measures possible. The literature discussed in this subsection shows the improvement over time in classification of more advanced ML models in comparison to basic dictionary approach measures. In the computational linguistic literature, transformers are now state-of-the art with BERT[117] returning state of the art results in a plethora of NLP tasks (Liu et al. 2019; Sun et al. 2020; Nogueria and Cho, 2020). The finance sentiment analysis literature to date has shown economically small, but statistically significant relationships between sentiment measures and market characteristics through techniques such as the dictionary approach, specific finance dictionaries and machine learning. It is necessary however to implement the techniques that are achieving state-of-the-art results within NLP to understand whether these measures are more adept at classifying financial sentiment and can return more robust and economically significant

[111] Recall score is the number of positive class predictions from all positive examples in the dataset.

[112] F1 score is a measure of balance that concerns both precision and recall in one number.

[113] Han (2008) creates an option-implied sentiment proxy based upon the implied skewness of option information.

[114] Baker and Wurgler (2006) define a set of market data that they believe is driven by investor sentiment to create a market sentiment index.

[115] Following Verma and Soydemir (2009), the authors select eight fundamental factors as control variables, including one-month interest rate, economic risk premium defined by the difference between three-month and one-month interest rates, inflation rate, the return on portfolio of winning stocks over past twelve months minus those losing stocks, the currency fluctuation of Hong Kong dollar, and the Fama-French three factors: the excess market portfolio return, the return on portfolio of small companies minus big ones, and the return on portfolio of high book to market value companies minus low book to market value ones.

[116] MSE measures the average squared difference of estimated values an actual value.

[117] Google's BERT model adopts transformer architecture in a bidirectional manner.

relationships with market variables. The following section extends the review of sentiment analysis discussed within this section particularly focusing on sentiment analysis for Earnings Conference Calls.

2.3. Applications of Sentiment Analysis on Earnings Conference Calls

Earnings Conference calls (henceforth referred to in this literature review as 'earnings calls') are a channel of communication whereby company managers, commonly Chief Executive Officers (CEOs) and Chief Financial Officers (CFOs), provide a statement surrounding past, present and future firm performance and answer questions posed by interested parties - analysts, institutional investors and individual investors. Frankel et al. (1999) cite the rapid growth of firms utilising earnings calls, which today has transcended into a widespread adoption of the communication channel - some 92% of companies represented by the National Investor Relations Institue (NIRI) members today actively run earnings calls. They conclude that the contents of conference calls provide additional information to the market. Matsumoto, Pronk, and Roelofsen (2011), Doran et al. (2012) and McKay Price et al. (2012) provide support for this statement, finding that earnings call participants are actively engaged to the extent that new and meaningful information comes to light. This additional information is the product of analysts and institutional investors' continued participation and probing for information alongside the supplementary insights that managers sometimes provide above that contained within the press release. Earnings calls are structured in a different format compared to other qualitative data communications used within the industry and these differences allow for new information to be unearthed.

Earnings calls are typically characterised by two elements (i) discussion of firm performance by firm executives and (ii) a question-and-answer session between the firm executives and market participants. Expanding on each of these sessions in more detail, the introductory statement at the start of the call given by firm executives, mimics the above formats of other financial disclosures in that it is predefined and constitutes a one-way channel of information (Blau et al. 2015). This scripted section offers managers the ability to explain prior results, such as highlighting successes over the previous financial quarter, or suppressing the anxiety among investors surrounding poor performance. Furthermore, it allows executives to give insight into future firm prospects. The question-and-answer session follows the opening statements, allowing analysts and investors the opportunity to ask managers questions surrounding firm performance; thus, allowing participants to further understand the reasons behind the results obtained in the previous financial quarter and gain insight into the future directions of a firm. Frankel et al. (1999) note that the potential reason for these calls adding additional information to the market stems from their structure. First, having outside parties on the call, with an ability to question company management, potentially provides further information to that provided in pre-written statements. Second, managers are more likely to respond to questions by producing more forward-looking statements than disclosed through other channels, such as annual or quarterly disclosures. Consistent with Frankel et al. (1999), Li (2010) finds that the sentiment of forward-looking statements in the Management Discussion and Answer (MD&A) section of 10Ks has explanatory power incremental to other variables and is positively correlated to future performance.[119]

In recent studies the distinct set up of earnings calls - particularly the two differing sections and the nature of participants on the call - has provided opportunity for varied research. Authors have produced research evaluating associations between specific manager (Larcker and Zakolyukina, 2012; Mayew and Venkatachalam, 2012; Davis and Tama-Sweet, 2012; Davis et al. 2015) or analyst (Millan and Smith, 2017) sentiment and returns. Comparisons between manager and analyst sentiment (Brockman, Li and McKay Price, 2015; Borochin et al. 2017; Chen, Nagar and Schoenfeld, 2018), and the actions of investors in response (Mayew and Venkatachalam, 2012; Blau, DeLisle and McKay Price, 2015; Amoozegar et al. 2020; Bochkay et al. 2020; Chen, Han, and Zhou, 2023), have also been examined. However, research in this area primarily focuses on the overall sentiment of a call – sentiment calculated and aggregated based on all call participants (McKay Price et al. 2012; Doran et al. 2012; Wang and Hua 2014; Borochin et al. 2017; Fu et al. 2019).

To the author’s knowledge, McKay Price et al. (2012) conduct the first investigation of associations between the sentiment of earnings calls and market variables. Controlling for the numerical representation of the earnings surprise, the results show that positive/negative earnings call sentiment is significantly related to the initial earnings announcement window[120] abnormal stock returns, the post-earnings announcement drift, and abnormal trading volume. Further, the researchers find that qualitative information in the form of earnings calls has greater explanatory power on subsequent returns over long horizons,[121] in comparison to the actual earnings number. More succinctly, the market may find it easier to incorporate numerical data, but qualitative data provided in the earnings call format is shown to provide additional value relevant information. Particularly, the Q&A section of the call holds significant ability in predicting CARs, post earnings drift and abnormal trading volume, when controlling for numerical earnings surprise and the sentiment of the prepared remarks. Thus, showing that the only financial disclosure to contain natural language conversations surrounding firm performance, returns a rich source of information. Combined, these findings, as the first of their kind show the incremental informativeness quarterly earnings calls have on market reactions.

Doran et al. (2012) focus specifically on Real Estate Investment Trusts (REITs).[122] Their analysis focuses on the extent to which linguistic sentiment produced in earnings calls concerning REITs is associated with future fluctuations in market value. Consistent with McKay Price et al. (2012), earnings

[119] The return on assets (ROA) in the following quarter for firm with a positive MD&A is 5 percentage points higher than a firm with a negative MD&A. Furthermore, a change in interquartile range for MD&A sentiment is found to move ROA by 1 percentage point.

[120] The three-day window from the day before to the day after an earnings call.

[121] From two days after a call up to sixty days after a call.

[122] The authors note that REITs are constantly involved in asset acquisition and disposition activities. Hence, the underlying revenue generating asset bases are constantly changing for REITs. These unique characteristics of REITs provide a natural setting in which to study the relation between stock returns and conference call content.

call sentiment is found to have significant explanatory power over abnormal returns at the market level. Interestingly, the authors note that firms whose earnings calls contain substantial positive (negative) sentiment have higher (lower) abnormal returns. Furthermore, their analysis produces findings that indicate positive call sentiment can completely offset negative earnings surprises for low earnings surprise firms.[123] These results suggest that managers have the potential to improve firm performance by using positive linguistic terminology during calls.

Borochin et al. (2017) also identify earnings calls as an important medium for disseminating information to the market. Their analysis differs from the studies mentioned previously in this subsection as it focuses on a measure of risk: value uncertainty.[124] Their results indicate that higher levels of pessimism lead to greater pricing uncertainty, with higher levels of optimism creating the opposite effect. Further, the authors separate earnings call sentiment into three distinct aspects: the manager's sentiment during the call introduction, the manager sentiment during the Q&A, and analyst sentiment during the Q&A. The authors find that both participants impact investor uncertainty but with differing magnitudes. Negative coefficients for the managerial introductory element of the call were identified.[125] This suggests that value perceptions of investors for the upcoming quarter are slightly impacted by a managers' introductory statement. However, no relationship for manager Q&A sentiment and returns were significant thus implying their sentiment within this section of the call is less influential. In comparison to managers, the expression of negative analyst Q&A sentiment is shown to heavily influence investor uncertainty.[126] This suggests that analysts are perhaps viewed in a more trusted and objective light on the call. Hence, analyst sentiment receives more market attention and thus holds greater influence on share price.

Borochin et al. (2017) partially addresses the extent to which managers can manipulate sentiment during earnings calls, finding that management's attempt to conceal bad news (answering analysts' questions with prepared scripts emulating the sentiment produced in their introduction) increases value uncertainty. Moreover, a large discrepancy in sentiment between managers and analysts leads to value uncertainty, but a closer aligned sentiment lowers the risk metric. This suggests that effectively mitigating poor performance through sentiment is an extremely difficult - verging on impossible - task, as even a well-informed manager who can adequately produce a more optimistic sentiment surrounding poor results will reveal a larger sentiment spread.

[123] Low earnings surprise is when a firms' reported profits are significantly below its earnings estimate.

[124] The authors evaluate an options market instead of the commonly assessed equities markets. A share price in an equities market reflects the current value of the firm. However, the implied volatilities in an options market reflect investors' uncertainty surrounding a firms' future value. Hence, Borochin et al. (2017) evaluate value uncertainty.

[125] A one standard deviation movement in managerial introductory tone reflects a -0.010 movement in the value uncertainty measure.

[126] A one standard deviation movement in analyst Q&A tone incites a -0.028 movement in the value uncertainty measure.

Most recently, Fu et al. (2019) analyse associations between earnings call sentiment and a different risk metric: stock price crash risk.[127] Higher levels of optimism on the quarterly calls are found to negatively predict stock price crash risk with statistical significance.[128] Thus, higher optimism reduces stock price crash risk. This conclusion is somewhat anticipated given that optimism is typically associated with positively performing companies and stock price crash risk is associated with struggling institutions. Consistent with McKay Price et al. (2012), the Q&A section of the call is found to have greater predictive power over market pricing than the introduction section, reinforcing that the market pays closer attention to the Q&A section of the call. However, there is a lack of consensus as to the most informational aspect of earnings calls. For example, Fu et al. (2019) provide conflicting evidence suggesting that managers' sentiment throughout the Q&A section has stronger and more statistical prediction power. This conflicts with Borochin et al. (2017), who conclude that "managers, as corporate insiders, possess private information and engage in truthful communication during conference calls”. Thus, indicating that managers generally remain truthful on conference calls and do not attempt to mislead participants to improve their performance, even in the face of extreme downside risk.

A recurring theme throughout the above papers is the impact that manager-specific sentiment and style has on the market reaction to earnings calls. The first paper to look in-depth at managerial sentiment is Larcker and Zakolyukina (2012), who evaluate whether the language of executives can assist in unearthing financial reporting manipulation or misstatements. In their analysis of individuals occupying managerial positions, they find that deceptive CEOs and CFOs have distinctive traits in common. For example, both: (i) use more references to general knowledge, (ii) use fewer non-extreme positive words, and (iii) limit their discussions surrounding shareholder value. These findings infer that the consideration of managerial linguistic features offers a valuable tool in understanding the quality of financial reporting. Finally, the researchers assert that linguistic models applied in this setting dominate, or are at least equate to, models that are based on purely accounting and financial information.

Mayew and Venkatachalam (2012) evaluate nonverbal communication on earnings calls. Unlike Larcker and Zakolyukina (2012), these authors looked at managerial affective states[129] in relation to future firm performance. Kappa, Hess and Scherner (1991) suggest that indicators of emotion through vocal cues are accurately detected and often as good as or better than facial cues and expressions to interpret information. This analysis suggests that a manager’s vocal cues allow analysts to learn about a manager’s affective state, and in turn about the firm's financial future. The way the manager

[127] Extreme downside risk in returns. It can be defined as the conditional skewness of the returns distribution which captures the information asymmetry, between inside managers and outside investors, in the risk associated with a stock.

[128] Regression results for the predictive power of call tone on stock price crash risk indicate that an increase of one standard deviation in call tone results in a decrease in stock price crash risk of 0.092 standard deviations.

[129] The underlying emotional state. An example of positive affective states is defined by the author as, happiness, excitement, and enjoyment. Examples of negative affective states are fear, tension, and anxiety.

communicates, surrounding his/her firm's performance, will contain cues which further interpret the situation because emotions are drawn from situations for the most part.

The research identifies three main findings. First, vocal cues that reflect managerial affective states predict future unexpected earnings. The authors show that a one standard deviation in a positive affective state defined by vocal cues increases unexpected earnings by 0.069% and 0.075% for two period and three periods ahead unexpected earnings.[130] Similarly, a one standard deviation increase in negative affective states decreases unexpected earnings by 0.031% and 0.043% for two period and three period ahead unexpected earnings. These findings suggest that affective states contain incremental information for predicting future unexpected earnings, which is particularly useful for predicting two periods ahead. The figures given above are in relation to the high scrutiny partition of Mayew and Venkatachalam's (2012) study.[132] Conversely, the low scrutiny results for the same regressions do not produce any statistically significant results implying that the market pays more attention to a manager's response when analysts are interrogating them. In this situation managers are forced to produce natural answers and cannot rely on scripted responses, hence the market's strong reaction to information produced in these situations.

Secondly, investors and analysts respond to vocal cues at least partially. Mayew and Venkatachalam (2012) show that investors react to both positive and negative affective states, with their market reaction being more pronounced to negative affective states in high scrutiny situations. On the contrary, positive affective states in high scrutiny conditions is significantly related to revisions in analyst recommendations. Hence, implying that on average analysts respond to and incorporate positive affective states when making changes to their recommendations. Finally, negative managerial affective states predict future returns.[133] Thus, communication by managers in conference calls does appear to hold informational content, particularly when evaluating future firm performance. Perhaps the greatest contribution to the literature however is the use of nonverbal communication in a capital market setting. In previous sections the successfulness of future firm predictions has been shown via textual analysis. This study has since provided the foundations for numerous subsequent studies employing this different modality, though this is an area still in its infancy.

In the same vein, Davis (2012) and subsequently Davis (2015), assess the effect that managers other than the chief executive have on the sentiment of earnings conference calls, seeking to identify the extent to which manager-specific optimism impacts the language used in firms' conference calls. The initial results show that managers' sentiment and language choice is strongly and positively related

[130] The length of a period is 90 days.

[132] High scrutiny is defined by Mayew and Venkatachalam (2012) as firms who return results lower than analyst forecasts and vice versa for the low scrutiny partition. The authors split the data into these two categories to assess whether there is any difference in managers affective state when being 'interrogated' by analysts (p.2) due to returning results lower than forecasted (high scrutiny) or not.

[133] A one standard deviation increase in negative affective states relates to a decrease of 6.5% cumulative abnormal returns for the period 2 days – 180 days after a call.

with market reaction at the 5% significance level. Evidence of a manager effect on market reaction is shown through an increase in adjusted R^2 from 9.95% in the base regression (no manager variables) to 10.6% with the addition of managerial fixed effects.[134] Finally, the overall tone of the call is influenced by the arrival of an optimistic or pessimistic manager to the firm, thus suggesting that market reactions can be impacted by individual managers.

The first two of Davis' (2012) results, mentioned above, fall in line with previous research outlining that informational content can be gleaned through evaluation of managerial sentiment on conference calls. The third and final finding aligns with (Larcker and Zakolyukina, 2012; Mayew and Venkatachalam, 2012; Fu et al. 2019) in terms of the predictive power of managerial sentiment in earnings calls on the investment decision of investors. This paper is particularly important as it is perhaps the first to demonstrate that managerial sentiment on a conference call does not only reflect the private knowledge a manager has surrounding his/her firm, but that it is also a product of manager-specific tendencies towards optimism or pessimism. The authors show multiple factors which are associated with these tendencies (i.e. manager age, volunteerism, work experience, gender).

Building on the manager-specific findings in Davis (2012), Davis (2015) evaluates (i) whether each manager has a specific, consistent sentiment,[135] (ii) whether this sentiment remains constant across different firms they work for and, (iii) if the sentiment can be measured. It is found that manager-specific style can be detected and measured, and that it does stay constant even when the manager moves to a new firm. Further, observable managerial characteristics are identified which play a strong role in explaining the sentiment outputs generated. For instance, managers who started their career in a recession use less positive language, along with managers who have investment banking experience. On the other hand, managers with charitable involvement tend to be more positive. Drawing upon psychology literature such as Scheier and Carver (1993) to evaluate an individual's optimistic tendencies and how they translate into assessment of current events, implications of future performance and impact of language choice has proven beneficial in this study. Having these characteristics defined in Davis (2012) and Davis (2015), future analysis could potentially use these as control variables to reveal the true sentiment of a call, possibly creating a more informative variable for forecasting market metrics.

In a more recent analysis of managerial emotional states, using more advanced methods, Hajek and Munk (2023) analyse 1,278 earnings conference calls from the 40 largest U.S. firms, incorporating both textual and audio information to enhance financial distress prediction. While they acknowledge the widespread use of text-based sentiment analysis in financial forecasting, they highlight the

[134] Managers' sentiment as defined by the Henry (2006) and Loughran and McDonald (2011) wordlists are robust in predicting future operating performance.

[135] Manager specific sentiment can be thought of like a personality. Everyone has a personality specific to them and it does not change no matter the job or situation they are in. Davis (2015) evaluates whether managers have a specific sentiment, much like their personality, and whether this can be identified throughout different roles in their career.

underexplored role of managerial emotional states. To address this, they leverage a model combining FinBERT-generated text sentiment and speech-derived emotional states from the audio of earnings calls. Their initial analysis shows that while the FinBERT model alone performs well in predicting financial distress,[136] the inclusion of both text and audio features yields highly competitive results, significantly outperforming traditional dictionary-based approaches like Loughran and McDonald's (2011). In a subsequent analysis, the authors compare a baseline model using financial indicators alone with a model created using deep learning architecture that fuses text, audio sentiment, and financial data for one-year-ahead distress prediction. Their findings reveal that as richer representations of earnings call content are integrated, forecasting accuracy improves, with the multimodal deep learning model achieving an impressive 0.954% in predicting financial distress. This research underscores the value of incorporating multimodal data to enhance the predictive power of financial models.[137]

Evidently, managerial sentiment and style have been shown as informative variables for analysis surrounding various market metrics (Larcker and Zakolyukina, 2012; Mayew and Venkatachalam, 2012; Davis and Tama-Sweet, 2012; Davis et al. 2015). However, specific research into analyst attitude on calls remains understudied in the existing literature with perhaps the sole investigation being that conducted by Millan and Smith (2017). In their paper, the authors create a list of compliments used commonly by analysts on earnings calls, through an extensive reading of call transcripts. With this corpus at their disposal, they investigated analyst compliments, denoted as praise, on earnings calls to unearth potential relationships with the market. The underlying logic for this argument is thus: analysts may compliment managers to curry favour and build relationships, to gain a better position in accessing private information. Alternatively, they may compliment in an unbiased manner based upon the merits disclosed in the earnings announcement.

Millan and Smith (2017) find that praise is significantly and positively associated with abnormal earnings announcement stock returns, showing that a one standard deviation increase in praise coincides with a positive abnormal return of 1.34%. In comparison to traditional sentiment measures, such as Loughran and McDonald's (2011) wordlist, the authors find that their 'praise' dictionary is 3.5 times stronger in predicting the magnitude of abnormal returns. Furthermore, these results indicate that praise given by analysts is given accordingly, and not excessively produced when not merited, thus dismissing the idea that analysts compliment managers to curry favour. Curiously, praise is found to be statistically significant and related to future stock returns while both sentiment measures are not (overall sentiment and analyst sentiment). Thus, analyst compliments, defined as praise, looks to be a robust variable in understanding positive firm performance.

[136] The authors use Altman's z-score as their indicator of financial distress.

[137] This is a recent study within the domain of financial sentiment analysis which was conducted after the start of this thesis. The incorporation of both modalities on earnings calls to define sentiment is gradually increasing which provides strong validation of the direction of work undertook throughout this thesis.

So far in this review three sections of earnings calls literature have been addressed, namely: overall tone, managerial tone, and analyst tone. An evident gap stemming from the latter two sections is the comparison of managerial sentiment to analyst sentiment. Brockman, Li and McKay Price (2015) compare the sentiments of managers and analysts. The researchers were successful in identifying the magnitude of market reaction to different participants. Managers are found to speak with significantly greater optimism, at greater length and use less complex language in comparison to their call counterparts. This is somewhat expected, given that managers are disseminating information and in doing so want to communicate clearly and in a positive manner surrounding their firm, whereas analysts attend to gain further information surrounding earnings figures and future performance. Thus, differences in communication are expected.

Brockman, Li and McKay Price (2015) produce three key findings. Firstly, at the time of the call managerial and analyst sentiment is significantly associated with stock prices. Splitting managerial sentiment into three terciles - low, medium and high - the authors identify differences between more pessimistic managerial sentiment in comparison to more optimistic managerial sentiment. Their results in relation to stock price show that low managerial (analyst) sentiment returns a negative market reaction of -0.78% (-1.60%). Regarding the medium sentiment tercile, the market again reacts although the magnitude is comparatively smaller at -0.23% (0.10%). The most optimistic sentiment tercile incites a positive and highly significant market reaction of 0.81% (1.30%). This suggests that, in terms of market reactions, managerial and analyst sentiment form a similar pattern throughout the terciles with analysts evoking a stronger reaction. Secondly, overall positive (negative) sentiments are related to positive (negative) abnormal returns. Thirdly, both managerial and analyst sentiment is quickly incorporated into stock prices, with analyst sentiment gaining a stronger market reaction. The results also suggest that market participants attach more credence to variance in analyst sentiment in comparison to that of managerment. This seemingly confirms that the difference between both parties is not trivial.[138]

In related research, Chen, Nagar and Schoenfeld (2018) analyse manager-analyst conversations on earnings calls. Their research evaluates the impact, if any, that manager-analyst sentiment has on intraday stock prices. In their exploration of communication exchange these authors identify a similar theme to many prior papers: analyst sentiment carries more weight in market reactions. Based on the considerable evidence showing large fluctuations in stock prices over the interactive section of the call (Frankel et al. 1999; Matsumoto et al. 2011; Brockman, Li and McKay Price, 2015), the authors seek to establish which features of a call drive these movements. The results suggest that intraday share

[138] Based on these findings, the researchers attempt to create a portfolio based upon analyst sentiment measures. They create the portfolio by ‘going long’ on stocks with high analyst sentiment and ‘going short’ on stocks with low analyst sentiment. This produced a significant abnormal return of 1.32% over a six-month period. Thus, the authors conclude that a portfolio based upon this strategy is a good idea, however, may not be economically significant considering the inclusion of transaction costs.

prices significantly respond to analyst sentiment with evidence suggesting that the effect strengthens when analyst sentiment is relatively negative.[139]

Two accompanying findings may help to understand this relationship. Firstly, in comparison to management, analysts speak in a more neutral fashion. Secondly, both participants' sentiment moves away from optimistic as the call progresses. Thus, if analysts speak in a neutral fashion and are not biased, changes in sentiment thus reveal further information surrounding performance. Furthermore, call sentiment begins quite optimisticly due to managerial introductions and begins to move towards a sentiment which fits with the firm's performance of the previous quarter. Analysts seem to begin the call closer to this level of sentiment, thus investors utilising analyst sentiment may identify the informational content of the call quicker. Combined, these findings point strongly in favour of analyst sentiment being more influential in the market setting, and thus more useful from a commercial perspective.[140]

Thus far most of the literature has focused on the overall markets and has typically tested hypotheses that the market responds to sentiment information produced on earnings calls. Blau, DeLisle and McKay Price (2015) were the first to consider the perception of sentiment. Specifically, their paper evaluates whether short sellers incorporate "soft information"[141] (p.203) into their forecasts, thus allowing them to understand if and how investors gauge and use sentiment. In doing so, they look to understand whether an abnormal sentiment measure[142] is used by short sellers. The findings show that short sellers are sophisticated investors, and they can and do use soft information from earnings calls when valuing their stocks, proving that short sellers trade against firms with positive earnings surprises and high abnormal sentiment.

Most recently, Bochkay et al. (2020) evaluate the impact of extreme words[143] on market digestion. Creating a corpus of extreme words, the authors first find that abnormal returns are much more strongly correlated to extreme language relative to moderate words. A one standard deviation increase in extreme language results in a 6.9% increase in abnormal trading volume. They further show that over a 60-day period there is no indication of reversals or price drifts. The authors use analyst revisions to evaluate analyst reaction to extreme language. From the investigation over a ten-day period, the findings show that language extremity is strongly associated with analyst revisions and that analysts

[139] This effect is not seen with managerial sentiment.

[140] This does not mean that managerial sentiment should be disregarded however as it is still influential.

[141] Qualitative information.

[142] This measure is the difference between introductory statement sentiment and Q&A sentiment. The authors note that it measures inflated talk by managers who often speak overly optimistically in the introductory statements in comparison to more objective talk from analysts in the Q&A section. Kartik et al. (2007) identify that inflated talk should be considered bad news – hence the authors are evaluating whether short sellers are sophisticated enough to process inflated talk information in their forecasts as bad news.

[143] Bochkay et al. (2020) create a corpus of extreme words by deploying a Human Intelligence Task (HIT) on Amazon's Mechanical Turk service (MTurk). This task asked "highly qualified English-speaking workers" to rank 50 randomly selected words from the author's dictionary on a scale from -5 (extremely negative) to +5 (extremely positive). The average score from all participants was used to rank extreme words.

react more strongly to extreme positive language. Both findings are consistent with the idea that sentiment is an effective measure of market characteristics and show that the incorporation of sentiment measures is being used by investors.

In a similar investigation, Chen, Han, and Zhou (2023) assessed how emotional sentiment variables extracted from 848 earnings calls by Chinese companies (2012-2019) influenced analysts' follow-up behaviour. They segmented the audio files at the speaker level, trained a CNN-based model on the CASIA Chinese emotion database,[144] and applied it to the segmented audio, producing probabilities for 12 emotion categories based on six predefined dimensions: joy, panic, anger, surprise, sadness, and neutrality, for both genders. These probabilities were consolidated into three sentiment categories—positive, negative, and neutral—and compared to an analyst follow-up behaviour metric, calculating the ratio of analysts who renewed their ratings reports after a call. Results showed that analysts leveraged emotional cues from the audio when deciding to issue rating reports, with sentiments in the introductory and Q&A sections significantly impacting decisions. Specifically, positive and neutral sentiments from questioners led to upward rating revisions, while negative sentiments from management or peer analysts resulted in downward adjustments.

Chen, Han, and Zhou (2023) further analysed the impact of adding text and audio data from earnings calls to a base model for EPS prediction. Both modalities improved accuracy when added to the model singularly, with the audio modality increasing predictive ability by 32%, while the text modality only by 9%. These results indicate that audio modality of earnings calls provide more incremental information. Furthermore, the authors show that combining both modalities enhanced EPS forecasting by 39%, outperforming the use of each modality individually.

Overall, the analysis conducted so far on earnings calls has shown the importance of sentiment measures in further understanding market movements. All studies to date have found a degree of association between earnings call sentiment and market activity that is statistically significant. However, to increase the power of sentiment measures on market forecasts, it may be the case that more robust sentiment analysis models need to be utilised. Recent techniques in computational linguistics offer a potential opportunity to create more accurate measures and potentially return more significant results in the accounting and finance domain. Furthermore, leveraging the vocal modality on earnings calls, as shown by Mayew and Venkatachalam (2012) to be informative, alongside traditional textual measures is an avenue not yet explored by the literature. Mehrabian (1968) introduces the 7-38-55% rule, inferring that the communication of human emotion is as follows: 7% is communicated through the semantic contents of a message, 38% through a message’s vocal attributes and 55% via facial expression. Hence, with prior literature showing both textual and vocal characteristics on earnings calls

[144] Bao et al. (2015) created an annotated multimodal emotion database that classifies spontaneous emotional segments from films, TV shows and talk shows stemming from 219 native Chinese speakers.

being informative of market characteristics and following Mehrabian (1968) rule the adhesion of both measures is an observable future direction for the literature.

2.4. Chapter Summary

This chapter focused on Earnings conference call sentiment and its relationship with market characteristics. Following a brief introduction in Section 3.1, Section 3.2 evaluated the sources of financial information sentiment, ranging from financial disclosures to social media, assessing the association with market variables. Section 3.2 placed considerable emphasis on the techniques used to determine sentiment within the extant literature. As this thesis primarily focusses on analysis of earnings calls, Section 3.3 provides insight into the sentiment analysis literature using these calls. Specifically, it focusses on whether different variations of earnings call sentiment induces market characteristics[145] to fluctuate.

El-Haj et al. (2018) identifies that the field of accounting and finance falls behind that of NLP studies in classification of sentiment using state-of-the-art methods. They note that there is a scarcity of ML applications in the financial domain with the majority who do employ ML using basic approaches such as Naive Bayes (Antweiler and Frank, 2004; Li, 2010; Sprenger et al. 2013). While Naive Bayes and other similar algorithms are simple and have been shown to classify financial sentiment, alternative approaches such as deep learning have been demonstrated as having greater abilities in classifying sentiment in different domains (Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Alamoudi and Alghamdi, 2021). Hence, further investigation into the application of more advanced models is required to understand if they are better suited to classify financial text over and above that of the more basic models discussed throughout Section 3.2.

Section 3.3 identifies relationships between various sentiment measures stemming from earnings calls and market characteristics – particularly market returns. The literature is fully focused on the textual modality with the only exception being Mayew and Venkatachalam (2012) who leverage vocal cues to assess managers' affective states and in turn market reaction. They note that vocal cues are good indicators of emotion and are accurately detected and used to interpret information. The combination of both textual and vocal modalities has been used within the computer science literature to create more robust measures of sentiment (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021) but not yet leveraged within finance. Besides the potential of creating a more robust sentiment indicator utilizing both textual and vocal modalities, the further use of vocal modalities can be assessed like that of Mayew and Venkatachalam (2012) to further assess their impact on market characteristics.

Therefore, from the research into extant literature there exists an opportunity to make additions to both the finance and computer science academic fields. Utilising developments in multimodal sentiment analysis is something that has not yet been considered which may provide deeper insights

[145] Commonly used market characteristics are returns, trading volume and volatility.

into drivers of financial market behaviour. From a finance perspective the question whether advanced multimodal models used within linguistics are appropriate for classifying financial sentiment must be asked to determine if the sentiment drawn from these models is related to market characteristics. Furthermore, the assessment of vocal modalities, combined with the commonly assessed textual modalities, impact on the decision-making process of investors and subsequently the impact on market characteristics will give evidence towards one of the competing theories. Vocal cues impact on decision making is well documented within the psychology literature[146] and hence gives backing to the assessment of such questions in a financial setting.[147]

3. Methods

3.1 Introduction

This chapter aims to provide a comprehensive overview of the methods used in Chapters 4, 5 and 6. The techniques used within all three studies are directed towards the overarching question outlined within this thesis: to further understand the decision-making behaviours of market participants in response to the release of information from earnings conference calls. To evaluate this question from a new perspective this thesis leveraged methods from the psychology, NLP and finance domains. The combination of approaches generated the creation of a multimodal classifier that leverages the textual and paralinguistic components of earnings conference calls to define the sentiment on said calls. The multimodal classifier deploys FinBERT to generate numerical representations of the textual content of earnings calls and combines these representations with paralinguistic features from the audio content of these calls. The combination of both text and paralinguistic features are used as inputs in a deep neural network to classify the overall sentiment of the earning calls within the sample used in this thesis. Creating the dataset, discussed throughout this chapter, that allowed for a multimodal sentiment study to be completed creates a significant contribution to the area of finance sentiment analysis in terms of the novelty and scale of the dataset developed. Furthermore, the creation of this multimodal dataset has allowed for a combination of more sophisticated techniques to be used in classifying the sentiment of earnings calls and thus allows for a more robust sentiment indicator to be used when investigating the relationship between call sentiments and market characteristics.

El Haj et al. (2018) notes that one reason for the limited application of advanced methods in finance sentiment analysis is the scarcity of suitable datasets. This thesis makes a primary contribution to the literature by utilising the largest known financial multimodal dataset to date, comprising 4,860 earnings calls and 637,220 sentences, with fully aligned paralinguistic features. The process of aligning calls and generating paralinguistic features is highly complex and time-consuming, requiring significant

[146] See Chapter 1.4 Theoretical Justification.

[147] A version of this literature review has been published in the Journal of Intelligent Systems in Accounting, Finance and Management.

technical precision as discussed in the following section 3.5. In comparison, prior studies have analysed much smaller datasets, such as Mayew and Venkatachalam's (2012) 466 earnings calls, Chen, Hand, and Zhou's (2023) 848 calls, and Li et al. (2020) dataset of 3,443 calls. This thesis, however, leverages a unique dataset exclusively comprising earnings call data from S&P 100 companies, unprecedented in both its size and scope. The creation of this extensive dataset enabled an index level analysis of the impact both of textual and paralinguistic information communicated during earnings calls on market behaviours.

To fully explain the methods used within this thesis, this chapter is set up as follows, section 3.2 gives a brief introduction to the methods used within this thesis. Section 3.3 discusses the data being used for this thesis, identifying the choice of financial disclosure and the market being evaluated throughout. Section 3.4 then gives a visual representation of the process used in the creation of the dataset, revealing the methods being used at each step of the process. As this thesis evaluates the content of earnings calls on various market variables the following two sections 3.5 and 3.6 both go into depth of the treatment of both respective data sources, earnings calls and market data respectively, how they are used and put into a final workable format. Section 3.7 goes into more depth surrounding the specific calculation of the dependent variables used within the regression analysis in chapters 5 and 6 of this thesis. Section 3.8 then gives an insight into the underlying techniques used within the sentiment analysis process of these studies, explaining how transformer architecture works and how it is being applied here. Finally, section 3.9 concludes this chapter.

3.2 Research Methods

The methods underpinning the work conducted within this thesis are of high importance as it is one of the determining factors that sets it aside from the bulk of work conducted in prior studies involving sentiment analysis. Ball and Brown (1968) began evaluating the utility of financial disclosure content for investors. Early studies in this area focused on the informational value of numerical data in financial disclosures (Fama, 1970; Black and Scholes, 1973; Merton, 1973) but as the literature has progressed there has been a venture into the usefulness of qualitative data for investors. Gentzkow et al. (2019) in their review of the textual data literature, highlight the substantial progress achieved in the methods used to understand qualitative data over the past 20 years. At the forefront of this specific strand of qualitative content analysis which this thesis focuses on (sentiment analysis), there are two main techniques that have gained superiority in classification accuracy – transformer architecture[149] and multimodal analysis.[150] However, the implementation of these techniques has not yet ventured fully into financial research. This thesis leverages both techniques in a combined fashion to progress the methods

[149] Introduced by Vaswani et al. (2017) transformer architecture is a neural network that fully depends on an attention mechanism to draw global dependencies between input and output.

[150] Multimodal analysis refers to an investigation in the NLP domain which uses all three modalities of communication or a combination of two of these modalities (verbal, non-verbal and visual). This thesis creates a multimodal model which leverages the two modalities of communication on earnings calls – textual information and audio information.

used to calculate financial sentiment. The following section highlights the data used to create and evaluate a multimodal sentiment classifier that leverages transformer architecture whilst giving detailed explanations of the inner workings of both methods separately.

3.3 Dataset

Chapter 2 identifies that previous literature has applied sentiment analysis to various financial datasets to unearth relationships with various financial markets. Sentiment has been calculated from financial disclosures, social media and news media to evaluate its relationship with disparate financial markets. As this thesis aims to evaluate the accuracy of a multimodal sentiment classifier in defining financial sentiment and explaining market behaviours, in comparison to the most used financial sentiment classifiers in previous analysis it must be applied to a homogenous financial dataset. Hence, allowing for a robust comparison to be made and intelligible conclusions to be drawn. The National Investors Relations Institute (NIRI, 2004) highlight that earnings calls are the second most used disclosure for communicating corporate information to financial market participants behind the earnings press release. Earnings calls are extremely popular amongst financial market participants, subsequently making them a prominent source of information for financial research. A potential reason for the popularity of this disclosure is the finding that return variances and trading volumes are heightened during earnings calls suggesting that the disclosure provides addition information above and beyond what is disclosed in the press release alone (Frankel et al. 1999; Irani, 2004). This finding along with the perpetual presence of analysts' and investors on these calls implies that the calls likely contain relevant incremental information that investors use to trade in real time. Frankel et al. (1999) builds on the above point by discussing two reasons that give earnings calls the capacity to bring additional information to the market. Firstly, detailed segment data not in the press release is sometimes disclosed on these calls. Secondly, in their review of a sample of calls they find anecdotal evidence that managers, often in response to analysts' questions, make forward looking statements which again are usually not included within the press release. This point is consistent with Kimbrough (2005) who alluded to the fact these forward-looking statements are a product of the somewhat informal exchange of performance details produced in the conversational question and answer portion of the call.

Matsumoto et al. (2011) highlights that the unknown drivers of the incremental informativeness of earnings calls over the accompanying earnings press release is one of the reasons the financial disclosure is popular for financial research. Earnings calls are particularly popular in the branch of financial sentiment analysis. A stated reason for their popularity is due to earnings calls being the only financial disclosure in which participants use natural language to communicate (Core, 2001 and McKay Price et al. 2012). Both authors highlight that through the application of sentiment analysis to these natural language conversations it may be possible to advance our understanding of the impact of call content on market participants' decision making. This indeed has been the case with scholars in the financial sentiment analysis area studying multiple facets of sentiment on these calls and indeed finding relevant information.

Previous research evaluating the sentiment of earnings calls mainly focuses on estimating the relationship between overall call sentiment and market index characteristics (McKay Price et al. 2012; Doran et al. 2012; Wang and Hua 2014; Borochin et al. 2017; Fu et al. 2019). However, comparisons between manager and analyst sentiment have been evaluated (Brockman, Li and McKay Price, 2015; Borochin et al. 2017; Chen, Nagar and Schoenfeld, 2018), along with specific manager sentiment (Larcker and Zakolyukina, 2012; Mayew and Venkatachalam, 2012; Davis et al. 2012; Davis et al. 2015) and analyst sentiment each in relation to market returns. Moreover, the actions of investors in response to sentiment has been examined (Mayew and Venkatachalam, 2012; Blau, DeLisle and McKay Price, 2015; Amoozegar et al. 2020; Bochkay et al. 2020). These citations highlight the interest of earnings call sentiment and provide a solid basis to compare this newly introduced sentiment analysis method, defined within this chapter, to a well-established domain.

An equally, if not more, important reason as to why earnings calls were selected as the medium in which this thesis explores sentiment is due to their multimodal nature. Sentiment analysis within the financial domain has evaluated content derived from a plethora of financial information sources. Research has assessed sentiment from financial news media[151] (Tetlock, 2007; Tetlock, Saar-Tsechansky and Macskassy, 2008; Garcia, 2013; Ferguson et al. 2015; Johnman, Vanstone and Gepp, 2018; Grob-Klubmann and Hautsch, 2010; Sun, Najand and Shen, 2016; Audrino and Tetereva, 2019), 10K and 10Q statements (Loughran and McDonald, 2011; Davis and Tama-Sweet, 2012; Jegadeesh and Wu, 2012; Jiang et al. 2019), Earnings Press Releases (Henry, 2006:2008; Davis and Tama-Sweet, 2012), Analyst reports (Twedt and Rees, 2012), CEO speeches (Bannier et al. 2017), Twitter (Bollen, Mao and Zeng, 2011; Mao and Bollen, 2011; Sprenger et al. 2013; Azar and Lo, 2016; Gu and Kurov, 2020) and Facebook (Siganos, Vagenas-Nanos and Verwijmeren, 2014; 2017). All of these information sources that have been leveraged for financial sentiment only contain textual data. There is potential for CEO speeches to contain text, audio and visual data but, to the author's knowledge, there are no data vendors that provide such details. Hence, to answer the initial question that this thesis poses in Chapter 4 – whether a multimodal sentiment classifier outperforms more basic approaches commonly used in extant literature – earnings calls provide the most suitable financial information set.

Once the decision had been made to analyse earnings calls sentiment, an appropriate set of earnings calls had to be chosen. Chapter 2 identifies that the majority of previous research has been conducted evaluating earnings call sentiment stemming from companies listed within the US. The US capital market is one of the largest, most liquid and most dynamic markets in the world with firms being subject to comprehensive and transparent regulatory frameworks creating an ideal setting for the

[151] The Wall Street Journal, Dow Jones News Stories, New York Times, The Financial Times, The Times, Gaurdian, The Mirror and Reuters News Scope Engine.

investigations conducted throughout this thesis (U.S. Securities and Exchange Commission, 2023). Furthermore, falling in line with the rationale used to select earnings calls, which were selected as they provided a medium in which comparisons can be made between the sentiment classifier introduced within this thesis and previous financial sentiment classifiers, the decision was made to evaluate calls from US based firms. The Standard and Poor's (S&P) 100 index is a market index that represents the 100 biggest US listed publicly traded companies. The S&P 100 is a subset of the S&P500 index and was introduced in 1983. To gain entry into the S&P 100 from the larger 500 index a firm must have a market capitalisation of at least $5.3 billion and trade at least 250,000 shares in each of the six months prior to entering the index (Leetham, 2023). The index's aim is to provide a measure of performance for large capitalisation companies based in the US. The index contains firms from a plethora of industries such as technology, healthcare, finance and energy and hence makes it a good indicator of the overall US stock market, as well as a good performance measure of large capitalisation firms.[152]

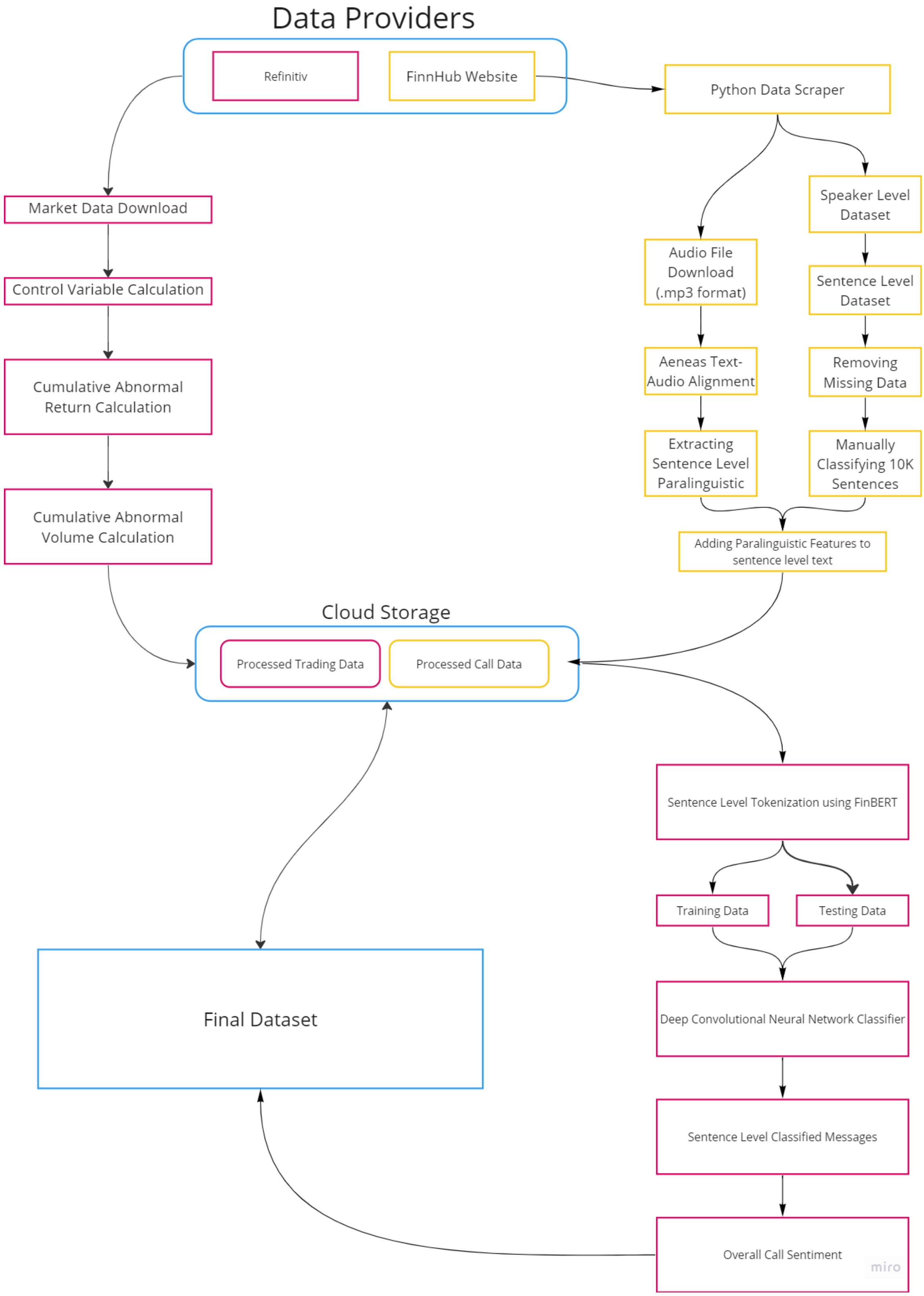
Furthermore, accessing textual and audio data of earnings calls was difficult however data vendors seemed to have more data surrounding such information for the highly visible S&P 100 than any other markets from the author's experience. Hence, due to the substantial amount of financial sentiment research conducted on US based firms, the representativeness of the S&P 100 and the ease of access to earnings calls data from said market made it an ideal choice for this research.

3.4 Data Collection and Processing Schematic

This section presents a data collection and processing schematic to visualise the quantitative techniques and processes used to create the final dataset that is analysed in Chapters 4, 5 and 6. The schematic shown in Figure 3.1 outlines the workflow, beginning from the acquisition of data from data providers ending with the final dataset. The tasks shown in this schematic are the retrieval, pre-processing and analysis of downloaded earnings call and market data to create the final dataset. This chapter focuses on the right-hand side of this schematic, discussing the boxes outlined in yellow, with the left-hand side of the schematic being discussed in Chapters 5 and 6. The first of the three main steps relates to the acquisition and transformation of earnings conference call data. This process can be seen in the top right-hand corner of Figure 3.1 and the specific details of how this section of the schematic was conducted is explained in detail within section 3.5.The last step in creating the final dataset involves defining sentiment for each sentence from each call on the dataset. The process of training, testing and validating a classifier model that is then used to classify all sentences can be seen in the bottom right-hand corner of the data collection and processing schematic. An overview of these sentiment analysis techniques is provided within section 3.6.

[152] See Appendix 3.1 for a full list of the S&P 100 firms used within this thesis along with their market cap, number of calls used, industry, sector and location.

Figure 3.1: Data Collection and Processing Schematic


Data Providers
Refinitiv
FinnHub Website
Python Data Scraper
Market Data Download
Control Variable Calculation
Cumulative Abnormal Return Calculation
Cumulative Abnormal Volume Calculation
Speaker Level Dataset
Audio File Download (.mp3 format)
Sentence Level Dataset
Aeneas Text-Audio Alignment
Removing Missing Data
Extracting Sentence Level Paralinguistic
Manually Classifying 10K Sentences
Adding Paralinguistic Features to sentence level text
Cloud Storage
Processed Trading Data
Processed Call Data
Sentence Level Tokenization using FinBERT
Training Data
Testing Data
Final Dataset
Deep Convolutional Neural Network Classifier
Sentence Level Classified Messages
Overall Call Sentiment
miro

Notes: This figure is a visual representation of the processes used to collect, transform and analyse the data used for Chapters 4, 5 and 6. The top of the figure highlights the two data providers used to download market data and earnings calls – Refinitiv and FinnHub. As the figure moves through the processes downloading, transforming and analysing the data it eventually reaches the final dataset which hosts a collection of sentiment data relating to each earnings call and market data across the full timeframe evaluated in this thesis 2005-2021.

3.5 Earnings Call Data

S&P100 Earnings Conference Call data from 2005-2021 was downloaded from FinnHub.[154] This data was scraped using a custom python script that accessed Finnhub's Application Programming Interface (API),[155] returning 7,047 transcripts. For each transcript, the original dataset came in the form of nine variables: name, speech, session, time, ID, quarter, symbol, year and audio.[156] In this initial format, each row corresponded to the text associated with the uninterrupted speech of each speaker, hence containing multiple sentences per row. Once transcripts were combined into one dataset, this equated to 524,478 rows. From the original 7,047 transcripts, 1,684 calls were removed as they had missing data[157] leaving 5,305 call transcripts. Following this process, the final dataset contained 442,782 rows of transcript data.

As the transcript dataset only included links to audio files hosted online, another python script was then created to download these audio files. The speaker level dataset was then split at the sentence level (i.e., each row corresponded to one sentence of text spoken by each individual on the calls) which expanded the dataset to 2.6 million rows. The sentence level text audio alignment process requires text to be at the sentence level. Hence, this step was impertinent to the successful alignment of the text and audio files.

Before the alignment process could begin, each conference call was given an individual directory where files were stored.[158] The generation of paralinguistic features begins with sorting each calls transcript (a .txt file where each line relates to a sentence), and the overall audio of the call (.mp3 file) into their associated individual directory. Two processes are then applied to the files. Firstly, a text-audio alignment process is applied to the calls to generate sentence level audio clips and, secondly, paralinguistic features are calculated from the sentence level audio clips. The first step in this research follows a similar method to Li et al. (2020) by leveraging the iterative forced alignment python package Aeneas.[159] Iterative forced alignment takes an audio file and a group of text files and automatically associates each text file with timings in the audio file. Aeneas takes a mathematical approach to aligning text and audio files by adopting a Dynamic Time Warping (DTW) algorithm. DTW is an established approach that has the capability of finding the optimal alignment between two time-dependent sequences as can be seen by example given in Figure 3.2 (Muller, 2007).

[154] FinnHub is a financial data vendor, and it can be found at https://finnhub.io/.

[155] An API is a set of rules, protocols and tools that allow different software applications to interact and communicate with each other.

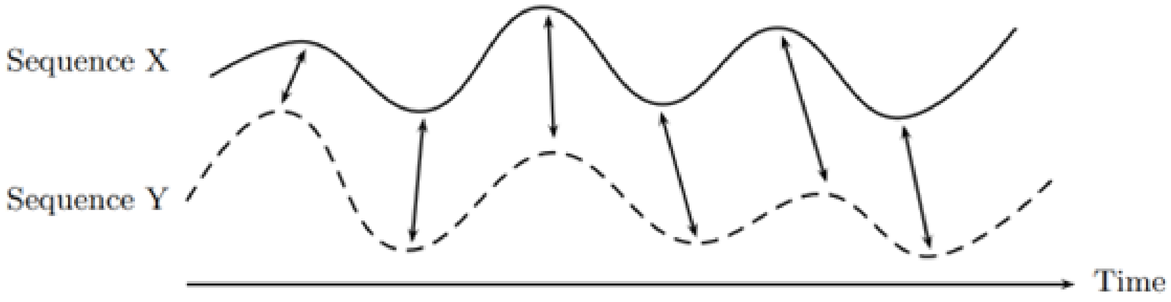
[156] Definitions of the variables from first to last: Name of the speaker, Text corresponding to that speaker, Session indicates if the text was in the introductory statement or the Q&A session of the call, Time gives the date and time of the call, A specific call ID provided by FinnHub, Quarter defines what quarter the call took place in, the company Ticker, the Year of the call and a link to the full phone call in audio format stored online.

[157] Missing data refers to a company call either having a missing transcript or audio file.

[158] This format is a requirement so that the alignment process aligns the correct text file with the correct audio file.

[159] https://github.com/readbeyond/aeneas.

Figure 3.2: Dynamic Time Warping (DTW) Example

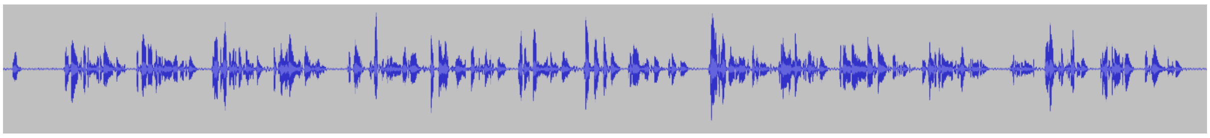


Notes: This figure gives a visual representation of DTW aligning two time-dependent series. The arrows represent the alignment points on both time-series (Muller, 2007:70).

Muller (2007) highlight that DTW is commonly used to compare speech patterns in automatic speech recognition. The Aeneas package uses DTW by aligning the Mel-Frequency Cepstral Coefficients (MFCCs) representations of audio files – the real audio and a group of synthesized audio files. The real audio file in this analysis is the overall earnings call audio. The synthesized audio files are synthetic audio files relating to each sentence from the textual transcript file of the overall call that have been created using a text-to-speech (TSS) engine.

The first step Aeneas takes is converting the real audio file into a useable wave file format. A depiction of what the real audio file looks like after it has been converted into the mono wave format is shown in Figure 3.3:

Figure 3.3: Example depiction of a mono wave audio file



Notes: This wave is being portrayed as a real audio file in wave file format for this analysis. In this research the real audio file is an earnings conference call i.e., the above figure represents a mono wave file of one of the earnings calls used in this research.

Following the initial step that converts the real audio file, the next steps are related to creating and adapting synthesized audio files from a transcript. In this step Aeneas creates a synthesized audio file in a wave format for each sentence using TTS,[160] creating a time map containing each individual synthesized audios length as a measure of time and then joining each individual synthesized audio clip together to create an overall synthesized wave file. Depictions of these can be seen in each of the four panels in Figure 3.4:

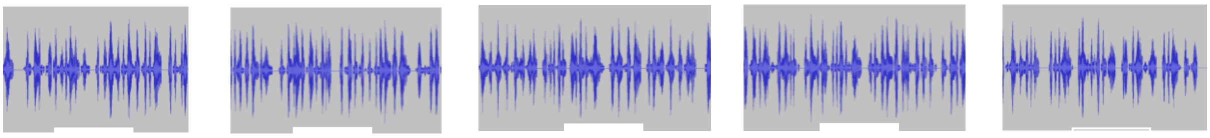
[160] The audio produced using TTS needs to be intelligible audio for the alignment to work but it does not need to sound natural.

Figure 3.4: Process of creating a full synthesized wave file

Panel A: Five example sentences from an Earnings Call transcript

Transcript ID	Sentence
S1	"Good afternoon and thank you for joining us."
S2	"Speaking first today is Apples CEO, Tim Cook, and he’ll be followed by CFO, Luca Maestri."
S3	"After that, well open the call to questions from analysts."
S4	"Please note that some of the information youll hear during our discussion today will consist of forward-looking statements."
S5	"These statements involve risks and uncertainties that may cause actual results or trends to differ materially from our forecast."

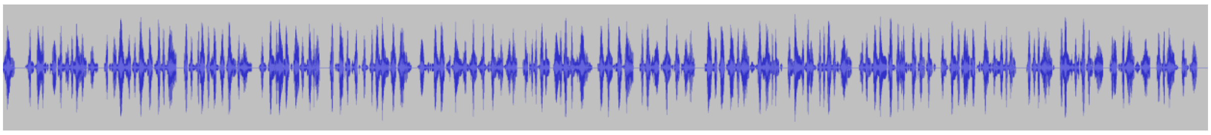
Panel B: Example of producing synthesized wave files for sentences in Panel A



Panel C: Example time map created from synthesized audio waves in Panel C

Transcript ID > Map ID	Time
S1 > M1	0 - 0.1569
S2 > M2	0.1560 - 2.3210
S3 > M3	2.321 - 3.8769
S4 > M4	3.8769 - 5.1123
S5 > M5	5.1123 - 8.0000

Panel D: Full synthesized wave file containing all synthesized sentences from the transcript

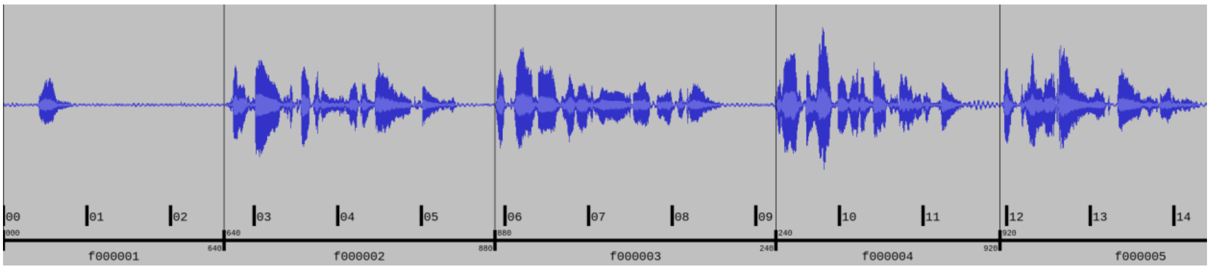


Notes: This figure provides a visual aid to the process Aeneas uses to create a synthesized wave file from transcript data.

Next abstract matrix representations are created for both the real wave file and the synthesized wave file; these representations are commonly known as Mel-Frequency Cepstral Coefficients (MFCCs). In the above figures 3.3 and 3.4 each spectrum depiction of a wave can be considered a Mel-Frequency Cepstrum (MFC), which is a representation of the short-term power spectrum of a wave. To create MFCCs each MFC is compared to cosine wave shapes, with the MFCC being a numerical representation of how similar the MFC is to a particular cosine wave shape.[161] For each MFCC matrix, each column represents one frame or sub-interval of the audio file i.e. if one frame was denoted as being 1 second then the example synthesized wave shown in figure 3.4 panel C would contain 8 frames and subsequently the matrix would contain 8 columns. As both the real wave and the synthesized wave will have different lengths, each matrix will have a different number of columns. The rows in both matrices represent one MFCC coefficient of each frame. For both the real and synthesized audio files the algorithm can then map a time in the audio file corresponding to a frame in either the real or synthesized MFCC matrices.

The following step then uses the DTW algorithm to find the minimum cost path to transform the synthesized audio into the real audio. To do so Aeneas takes the dot product of both MFCC matrices to create a cost matrix and then runs the DTW algorithm over this cost matrix to find an approximation of the minimum cost path. The output from the DTW is a time map which relates the columns, or time frames, in the real MFCC matrix to the synthesized MFCC matrix. This allows for the intervals of audio produced in the synthesized audio to be mapped back onto the real audio which outputs the times in which each sentence in the transcript is spoken in the real audio file. The real audio file is then split into multiple separate audio files that relate to each sentence. For instance, in the below example five sentences were spoken, so at the end of the process five smaller audio clips relating to each sentence would be generated.

Figure 3.5: Real audio file



Notes: This is a depiction of the final form of the real audio wave that contains time staps relating to each sentence spoken. F000001 relates to the time in which the first sentence on the transcript is spoken and so forth.

The second step in the process of calculating paralinguistic features for each sentence on every earnings call, is the step of generating the numerical representation of the paralinguistic features. This step is relatively simpler in comparison to the forced iterative alignment step. To generate the numerical representation of each sentence level audio file a custom python script was created using PRAATs Parselmouth python library.[162] This script systematically accessed each sentence level audio file and produced paralinguistic features for each sentence.[163] The paralinguistic features for each call were stored in a .csv file and were then incorporated into the main earnings call dataset using row indexing.

[161] MFCCs return the overall shape of the audio signal but do not contain specific details such as the tone of voice.

[162] https://parselmouth.readthedocs.io/en/stable/.

[163] A table of the paralinguistic features extracted from each sentence are described in Table 4.2.

Chapter 4 creates and trains the multimodal classifier that is used throughout the following two empirical chapters. Due to the time constraints associated with this research and the time-consuming process of forcibly aligning the earnings calls transcripts and their accompanying audio files, the first empirical chapter was completed on twenty firms from the overall S&P100 dataset. These firms were specifically selected at the beginning of the process to provide a robust representation of the overall index to mitigate the possibility of poor generalisation of call classification in chapters 5 and 6. The subset of twenty firms for chapter 4 contained 595,074 sentences. The accuracy of the alignment was roughly 30% which left 164,632 sentences that contained a full repository of paralinguistic features.

Chapters 5 and 6 of this research use all firms in the S&P 100 index which translates to 2.6 million original sentences. The alignment process for the full dataset returned 733,031 sentences that contained a full list of paralinguistic features. Before this data could be used to classify calls, and then implemented into a final dataset with financial market data, one last step to clean the data was taken. The 733,031 sentences relate to all participants on earnings calls which include operators. An operator on these calls is someone who opens the call and introduces the management team of the firm hosting the earnings call. Evidently, the speech relating to operators on the call doesn't provide any incremental information surrounding firm performance and subsequently is not used by market participants as an indication of future firm performance. Hence, all sentences relating to operators were removed from the dataset. This translates into 637,220 sentences containing a full inventory of paralinguistic features pertaining to managers and analysts on these calls. These 637,220 sentences translate into a full clean dataset of 5,172 earnings conference calls. Interestingly, the original number of calls used at the start of the alignment process was 5,305. Therefore, even though the accuracy of forced alignment process was only 30%,[164] it still managed to produce aligned sentences on 98% of calls. This would suggest that all calls in the final clean dataset of 5,172 do not contain all sentences from each call and implies there is a reasonable number of sentences missing from each call.

3.6 Sentiment Analysis Techniques

Sentiment Analysis is the computational study of people's opinions, attitudes and emotions toward an entity (Medhat, Hassan and Korashy, 2014). In academic finance, a number of researchers have utilised sentiment analysis to understand the underlying meaning of financial disclosures in attempts to gain an information advantage within financial markets. Chapter 2 of this thesis highlights that the techniques used to define sentiment within the financial domain fall behind that of the techniques used in the more advanced computer science literature. The authors show that many papers within finance use general dictionaries, finance specific dictionaries or machine learning methods to define sentiment. However, only a select few use more adept transformer architecture and deep learning

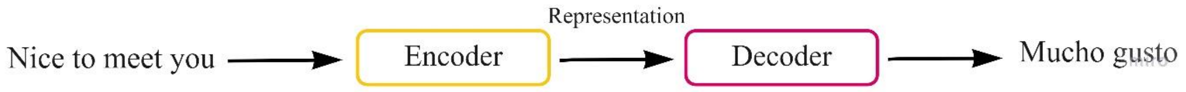
[164] We define the accuracy of the text-audio alignment process as the number of sentences successfully returned with a full list of paralinguistic features (733,031) compared to the initial number of sentences used in the process (2.6M).

techniques (Hiew, 2019). This thesis adopts a financially trained version of the Large Language Model (LLM) Bidirectional Encoder Representations from Transformers (BERT) to create numerical representations of earnings call transcripts. FinBERT, the financially pretrained BERT model, was introduced by Yang et al. (2020) which is trained on 4.9 billion tokens from corporate reports, earnings conference call transcripts and analyst reports. BERT, FinBERT and the majority of state-of-the-art language models today all rely on transformer architecture that was introduced by Vaswani et al. (2017).

To understand how BERT works[175] you first must understand the transformer. Chapter 2 highlights that prior to the introduction of the transformer, NLP tasks were dominated by Recurrernt Neural Networks (RNNs), Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). The transformer was introduced by Vaswani et al. (2017) to create a model that did not attempt to improve on previous results by using similar methods to the models mentioned above but instead introduced a model that did not require recurrence or convolutions. The transformer is a deep learning model that fully depends on an attention mechanism to draw global dependencies between input and output. Vaswani et al. (2017) return state of the art results in translation quality using the transformer architecture with many state-of-the-art models in various other NLP tasks adopting similar architectures. For example, Raffel et al. (2020) introduced a text-to-text transfer transformer (T5), which has achieved state-of-the-art performance on the SQuAD question and answering task.[176] Brown et al. (2020) train an autoregressive language model (GPT3) on 175 billion parameters,[177] which returned the highest accuracy of 86.4% on the LAMBADA language modelling task.[178]

The transformer architecture that was introduced in Vaswani et al. (2017) leverages an encoder and a decoder. The encoder is fed an input sentence and creates a representation of that sentence, which is then sent to the decoder that takes the representation and creates an output. An example of this process is provided in figure 3.6, which demonstrates how a transformer may be used to perform a language translation task.

Figure 3.6: A visual depiction of a basic transformer



Notes: This figure depicts the inner workings of a transformer completing a machine translation task. It shows a transformer translating a sentence from English to Spanish. In this example both the encoder and decoder are shown to give the reader a basic idea of how a transformer functions.

[175] And subsequently FinBERT.

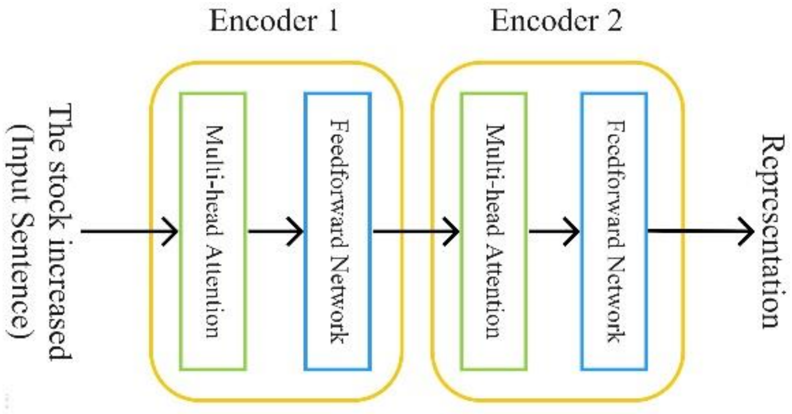
[176] The SQuAD Q&A task presents a model a paragraph with a question about the paragraph. The goal of the model is to effectively answer the question posed. The answers to the questions give insight to how well a model can understand text.

[177] ChatGPT is built upon a variant of this model.

[178] The LAMBADA dataset tests a model’s ability to handle long-range dependencies in text. The task requires a model to predict the last word of a sentence based upon a context paragraph as input.

However, BERT does not use the decoder layer of a transformer. It makes use of the encoder layer of the transformer to create a numerical representation of its textual input.[179] The transformer architecture outlined in Vaswani et al. (2017) stacks six encoders on top of each other. The first encoder receives the input sentence, the next receives the output from the first encoder, and so on until the last encoder returns the representation of the input sentence. The number of encoders stacked upon one another is not set and different values can be used. For instance, BERT-Base uses 12 encoders whilst BERT-Large uses 24 encoders.[180] Figure 3.7 is an example of a two-encoder stack showing the sublayers contained within each encoder.

Figure 3.7: An Example of BERT depicting multiple Encoders



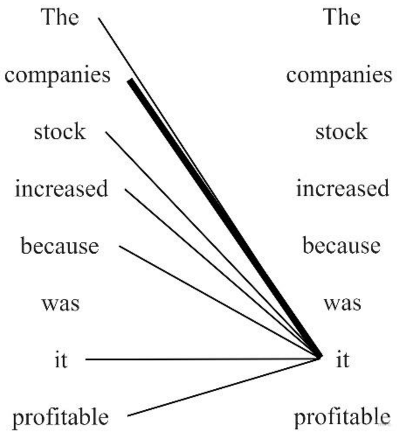
Notes: This figure gives an example of how BERT stacks encoders on top of one another to create a robust numerical representation of input sentences. This example only contains two encoders due to spatial reasons however it identifies the two sublayers contained within an encoder and the process of moving from an input sentence to the desired numerical representation.

Within each encoder there are two sublayers – a multi-head attention layer and a feedforward network. Multi-head attention is based upon the self-attention mechanism. The self-attention mechanism allows the machine to understand which words within a sentence relate to each other. It first takes each word within a sentence and computes its representation and whilst computing these representations the model takes into consideration how each word in the sentence relates to each other. In the following figure 3.8 we can see that to compute the representation for the word "***it"*** the model relates the word "***it"*** to all other words in the sentence to understand more about the word. In doing so it understands that the word ***it*** is related to the word "***company"*** and not the word "***stock"*** as can be seen from the thick line.

[179] Evidently, computers cannot understand text the same way humans process and understand text, however they are very adept at working with numbers. Hence, the job of an encoder is to create a robust numerical representation of a sentence so that the computer can understand the sematic contents of text.

[180] BERT-Base using 12 encoders has fewer parameters (110 million) than BERT-Large (340 million) which uses 24 encoders. The extra trainable parameters of BERT-Large allow it to return more accurate results in comparison to BERT-Base. However, BERT-Base having fewer parameters allows it to be used with less computational power and memory.

Figure 3.8: An Example of the Self-Attention Mechanism



Notes: This figure shows how the self-attention mechanism contained within the multi-head attention sublayer of an encoder relates each word within a sentence with one another to gain an understanding of which words are more related than others. In the above example we can see the self-attention mechanism identifies that 'it' in this example refers to the company.

In practice the self-attention mechanism begins by generating embeddings[181] for each word in an input sentence. A word embedding is the vector representation of a word, and the values of the embeddings are learned during training. Learning embeddings during training is an advantage in comparison to alternative, fixed representations models (such as Word2Vec and Naïve Bayes) as BERT produces dynamically informed word representations based upon the words around them and their relationships.[182] Take the example sentence: "***the stock increased"***. Let X_1 be the word embedding for the word ***the***, X_2 be the embedding for the word ***stock*** and X_3 be the embedding for ***increased***.[183]

Figure 3.9: Examples of Word Embeddings for the Example Sentence

- X_1 = [0.87, 6.41, …, 4.32]
- X_2 = [5.63, 1.24, …, 3.89]
- X_3 = [12.14, 9.21, …, 5.56]

Notes: This figure gives a numerical representation of the word embeddings associated with the example sentence.

These embeddings can be represented as an input matrix X as shown below:

[181] Word embeddings are representation of words that come in the form of a vector that contains the meaning of the word. If two-word embeddings lie close together in the vector space, they can be considered to share a similar meaning.

[182] For instance, take the two sentences 'the dog had a loud bark' and 'the tree bark was brown'. In both sentences the meaning of the word bark is different. For a fixed representation model the word embeddings for bark in both cases will be the same as they are predetermined. However, using self-attention the model will create two different word embeddings for both 'bark' as it learns from the words surrounding the word.

[183] The numbers here and throughout the following examples are arbitrary. They are being used to show the process used in self-attention.

Figure 3.10: Example Input Matrix

$$\begin{array}{l} \text{The} \\ \text{Stock} \\ \text{Increased} \end{array} \begin{bmatrix} 0.87 & 6.41 & \dots & 4.32 \\ 5.63 & 1.24 & \dots & 3.89 \\ 12.14 & 9.21 & \dots & 5.56 \end{bmatrix} \begin{array}{l} x1 \\ x2 \\ x3 \end{array}$$

X
(Input Matrix)

Notes: This figure gives a numerical representation of what the Input Matrix, created from the word embeddings, for the example sentence would look like.

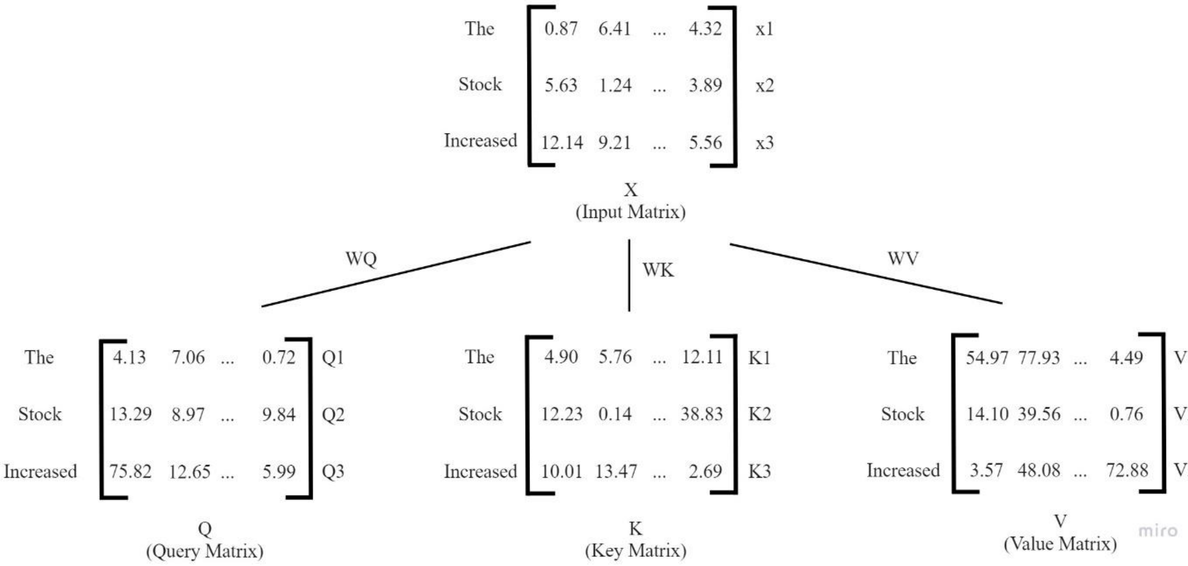
From the above input matrix X, the first row of the matrix represents the word ***the***, the second row represents the word ***stock*** and the third row represents the word ***increased***. The dimensions of an input matrix are defined as:

Input Matrix Dimensions = [sentence length * embedding dimension]

Therefore, if the embedding dimension of this model is 100 then the dimensions of matrix X would be [3,100]. The word embeddings and sentence lengths change model to model, with BERT having an embedding dimension of 768 and being able to handle pieces of text containing up to 512 tokens. After the input matrix has been created the next step in self-attention is creating three new matrices: Query matrix (Q), Key matrix (K), and Value matrix (V). These three matrices are created by multiplying the input matrix X by three weights: W_Q, W_K and W_V.[184] From figure 3.11 the first rows of Q,K,W relate to the word ***the***, the second rows relate to the word ***stock*** and the third rows relate to the word ***increased***. The dimensionality of the Q, K, W vectors is [3,64]

[184] The values of W_Q, W_K and W_V are initially random, and their optimal values will be learned in the training phase. As the optimal weights are learned, more accurate Q, K and V matrices will be created.

Figure 3.11: An Example of Creating Query, Key and Value Matrices from the Input Matrix



Notes: This figure shows the step of creating query, key and value matrices from the input matrix. This is one of the first steps in the self-attention process that is completed within the multi-head attention layer of each encoder.

Now with these matrices set up; the self-attention mechanism relates all words to each word in the sentence using four steps. The first step is computing the dot product between the query matrix and the transpose of the key matrix, the second step is then to obtain stable gradients within the matrix calculated with the first step, the next step normalizes this matrix and the final step is creating the attention matrix. Following the above steps the encoder returns a matrix that initially shows us how similar each word is to each other. For instance, in the below figure 3.12 the word ***the*** is more related to itself in this example as it has the highest coefficient compared to the others in that row. In essence step one gives us the similarity score of each word to the rest of the words in the sentence.

Figure 3.12: Depiction of the first step in the four step self-attention process

$$QK^{T} = \begin{array}{c|ccc|} & \text{The} & \text{stock} & \text{increased} \\ \text{The} & q_1.k_1 & q_1.k_2 & q_1.k_3 \\ \text{stock} & q_2.k_1 & q_2.k_2 & q_2.k_3 \\ \text{increased} & q_3.k_1 & q_3.k_2 & q_3.k_3 \end{array} = \begin{array}{c|ccc|} & \text{The} & \text{stock} & \text{increased} \\ \text{The} & 150 & 90 & 70 \\ \text{stock} & 65 & 120 & 99 \\ \text{increased} & 87 & 96 & 110 \end{array}$$

Notes: This figure shows the matrix multiplication of the query matrix and the transpose of the key matrix.

From the above figure we can see that the word ***the*** is more related to itself than it is to the other words in the sentence as the first row shows that the dot product for $q_1.k_1$ is higher (150) than that for $q_1.k_2$ (90) and $q_1.k_3$ (70). We then can understand how much each word is related to each other based upon the dot product values in each row. In this case ***stock*** and ***increased*** are also more related to themselves than to the other words in the sentence as seen by their dot product values in the matrix.

The second step of the process is used to obtain stable gradients within the matrix. This is done by dividing the matrix QK^T by the square roof of the dimension of the key vector. As noted earlier the embedding dimension of the key matrix is 64. Therefore, in this case the QK^T is divided by 8.

Figure 3.13: Creating Stable Gradients in the QK^T matrix

$$\frac{QK^T}{8} = \begin{array}{r|ccc|} & \text{The} & \text{stock} & \text{increased} \\ \text{The} & 18.75 & 11.25 & 8.75 \\ \text{stock} & 8.125 & 15 & 12.375 \\ \text{increased} & 10.875 & 12 & 13.75 \end{array}$$

Notes: This figure shows the stabilisation of gradients by dividing the QK^T matrix by the square root of its embedding dimension.

Step three then normalises the above matrix calculated in step two as it is in an unnormalised format. Applying a softmax function to the above matrix will create a normalised matrix with values ranging between zero and one, hence turning the vector values into probabilities. This newly created normalised matrix shown below is denoted as the score matrix. It gives us a better understanding of how each word is related to each other. In the score matrix below, we can see that the word ***the*** is 99% related to itself. The word ***stock*** is 93% related to itself, 6% related to ***increased*** and 1% related to ***the***. Moreover, ***increased*** is 81% related to itself 14% related to ***stock*** and 5% related to ***the***.

Figure 3.14: The Score Matrix

$$\text{Softmax}\left(\frac{QK^T}{8}\right) = \begin{array}{r|ccc|} & \text{The} & \text{stock} & \text{increased} \\ \text{The} & 0.99 & 0.01 & 0 \\ \text{stock} & 0.01 & 0.93 & 0.06 \\ \text{increased} & 0.05 & 0.14 & 0.81 \end{array}$$

Notes: This is a normalised version of the calcautled in figure 3.8.8 which shows how related each word is to all other words in the sentence in a normalised format.

After normalising the matrix using the softmax function to create the score matrix the final step computes the attention matrix, Z. The attention matrix contains the attention values for each word in the sentence. It is calculated by multiplying the score matrix by the value matrix as shown in figure 3.15:

Figure 3.15: The Attention Matrix

$$Z = \underbrace{\begin{array}{cccc} & \text{The} & \text{stock} & \text{increased} \\ \text{The} & 0.99 & 0.01 & 0 \\ \text{stock} & 0.01 & 0.93 & 0.06 \\ \text{increased} & 0.05 & 0.14 & 0.81 \end{array}}_{\text{Softmax}\left(\frac{QK^{T}}{8}\right)} \quad \underbrace{\begin{array}{ccccc} \text{The} & 54.97 & 77.93 & \dots & 4.49 \\ \text{stock} & 14.10 & 39.56 & \dots & 0.76 \\ \text{increased} & 3.57 & 48.08 & \dots & 72.88 \end{array}}_{\substack{V \\ \text{(Value matrix)}}} = \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix}$$

Notes: The attention matrix is computed by multiplying the score matrix by the value matrix.

Where Z_1, Z_2 and Z_3 are the sum of the value vectors weighted by the score matrix. For example, the self-attention for the word ***increased***, Z_3, will contain 5% of the values from the value vector V_1 that represents the word ***the***, 14% of the values from the value vector V_2 that represents the word ***stock*** and 81% of the values from the value vector V_3 that represents the word ***increased***. This computation can be visually represented as:

Figure 3.16: Example of Calculating the Values Contained within the Attention Matrix

$$Z_3 = \underbrace{0.05\ (54.97\ 77.93\ \dots\ 4.49)}_{\text{The}} + \underbrace{0.14\ (14.10\ 39.56\ \dots\ 0.76)}_{\text{stock}} + \underbrace{0.81\ (3.57\ 48.08\ \dots\ 72.88)}_{\text{increased}}$$

Notes: Visual representation of how the value for Z3 within the attention matrix is calculated. This same process is used to calculate the values for both Z1 and Z2 also.

The sentence used earlier "the companies stock increased because it was profitable" is used to illustrate the process further highlighting its usefulness. In this example the word ***it*** relates to the company and not the stock. Hence, the self-attention for the word ***it***, Z_{it}, would be computed using the steps outlined above, with the following output:

Figure 3.17: Example of Calculating the Values in the Attention Matrix

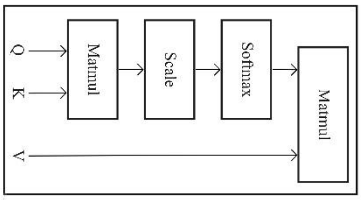
$$Z_{it} = \underbrace{0.00\ (27.93\ 3.90\ \dots)}_{\text{The}} + \underbrace{1.00\ (8.89\ 43.16\ \dots)}_{\text{companies}} + \dots + \underbrace{0.00(53.56\ 11.12\ \dots)}_{\text{profitable}}$$

Notes: This is the same visual representation as provided in figure 3.16 however the input sentence has been changed to the example used earlier in this chapter "the companies stock increased because it was profitable" and the attention matrix row associated with the word "it" within said sentence.

It can be seen from the example that the self-attention value of the word ***it*** contains 100% of the values from the value vector $V_{companies}$, therefore allowing the model to understand that the word ***it*** refers to ***companies*** and not ***stock***.

Figure 3.18: Attention Matrix Equation and Graphical Depiction of the Process

$$Z = softmax(\frac{QK^T}{\sqrt(d_k)})V$$



Notes: This figure provides the equation used to calculate the attention matrix (top) and a visual process of creating the final attention matrix from the initial Query, Key and Value matrices calculated from the input matrix.

Figure 3.11 through to figure 3.17, demonstrates how the self-attention process works and the preceding figure 3.18 shows how the attention matrix is calculated in equation form and in a graphical form respectively.

Multi-head attention follows on from the self-attention mechanism explained above. Above a single attention matrix (Z) was calculated. However, in multi-head attention instead of computing a single attention matrix multiple attention matrices are computed. To explain why multi-head attention is used, take the following sentence as an example ***how are you***. Say the similarity score has been calculated for this sentence following the process discussed above and now calculating the self-attention for the word ***are***, we have the following:

Figure 3.19: Self-Attention Calculation Example

$$Z_{\text{are}} = \underset{\text{how}}{0.7(12.45\ 5.78\ \ldots)} + \underset{\text{are}}{0.25(45.63\ 89.01\ \ldots)} + \underset{\text{you}}{0.05(27.87\ 64.91\ \ldots)}$$

Notes: This is the same visual representation as provided in figure 3.8.11 however the input sentence has been changed to "how are you" and the attention matrix row associated with the word "are" within said sentence.

Looking at the attention for the word ***are*** above, we can see that it is dominated by the word ***how***. In other words, the attention value for the word ***are*** contains 70% of the word ***how***'s value vector, 25% of its own value vector and 5% of ***you***'s value vector. Domination of a word like this is only useful if the meaning of a word is ambiguous. For instance, Z_{it} in figure 3.17 uses 100% of the values from the word ***companies*** value vector. In this instance the domination of the word ***it*** is acceptable as it is ambiguous. The meaning of ***it*** could be referring to companies or stock in this example**.** Unless we have an ambiguous case then the dominance of a word is not useful – like the ***how are you*** case. To make sure results are accurate multiple attention matrices are calculated and then their results are concatenated to create a robust final attention matrix. For instance, if five attention matrices have been calculated, to create the final multi-head attention matrix we concatenate all attention heads and then multiple by a new weight matrix, W_0, as shown:

Figure 3.20: Multi-head Attention Matrix

$$\text{Multi-head attention} = \text{Concatenate}(Z_1, Z_2, Z_3, Z_4, Z_5)\ W0$$

Notes: This figure shows how multiple attention matrices are concatenated to create a multi-head attention matrix.

Understanding the above multi-head attention matrix is the first step in understanding how an encoder works and how BERT works. However, before feeding the input sentence to the encoder layer there first needs to be a positional encoding applied to the input. The positional encoding step of BERT is extremely important, as unlike other models that feed input in word-by-word BERT receives a whole sentence at a time. Therefore, the positional encoding step gives an understanding of the position of each word within a sentence which in turn allows a better understanding of the meaning of the sentence. The following example gives a good understanding of how the positional encoding step functions. First let's take the sentence used previously ***the stock increased*** and get the embeddings for each word in the sentence. Our input matrix dimension (the matrix representation of our input sentence) will be [sentence length * embedding dimension]. We know our sentence length is 3 and let's say our embedding dimension is 4 for the example. Then we have the following input matrix X.

Figure 3.21: Example Input Matrix

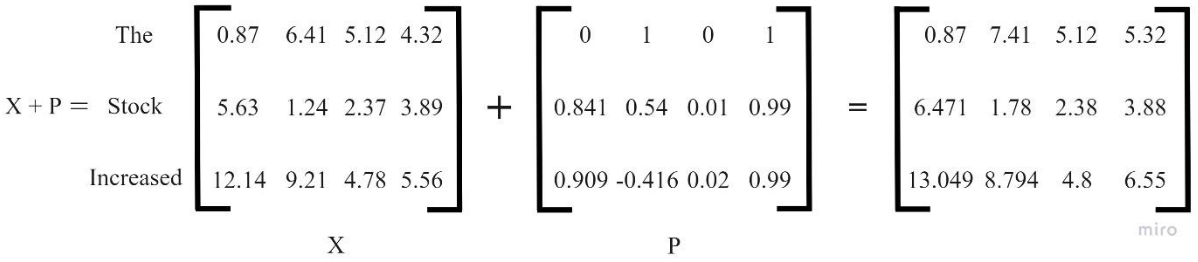
$$\begin{array}{l} \text{The} \\ \text{Stock} \\ \text{Increased} \end{array} \left[\begin{array}{cccc} 0.87 & 6.41 & 5.12 & 4.32 \\ 5.63 & 1.24 & 2.37 & 3.89 \\ 12.14 & 9.21 & 4.78 & 5.56 \end{array}\right] \begin{array}{l} x1 \\ x2 \\ x3 \end{array}$$

$$\underset{\text{(Input Matrix)}}{X}$$

Notes: This figure gives an example input matrix that is used in the following discussion to provide an explanation of positional encoding. The example matrix has a sentence length of three and an embedding dimension of four.

If this input matrix is given to the encoder with just word embeddings and no positional encoding the transformer cannot understand the word order. Hence, we need to add in some information regarding the position of each word. To do so we add a positional encoding matrix, P, to the input matrix X.

Figure 3.22: Adding Positional Encoding to the Input Matrix



$$X + P = \begin{array}{r} \text{The} \\ \text{Stock} \\ \text{Increased} \end{array} \underbrace{\begin{bmatrix} 0.87 & 6.41 & 5.12 & 4.32 \\ 5.63 & 1.24 & 2.37 & 3.89 \\ 12.14 & 9.21 & 4.78 & 5.56 \end{bmatrix}}_{X} + \underbrace{\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0.841 & 0.54 & 0.01 & 0.99 \\ 0.909 & -0.416 & 0.02 & 0.99 \end{bmatrix}}_{P} = \begin{bmatrix} 0.87 & 7.41 & 5.12 & 5.32 \\ 6.471 & 1.78 & 2.38 & 3.88 \\ 13.049 & 8.794 & 4.8 & 6.55 \end{bmatrix}$$

Notes: This figure shows the matrix addition process of adding the positional encoding matrix to the input matrix. This step allows the model to understand where each word is in a sentence.

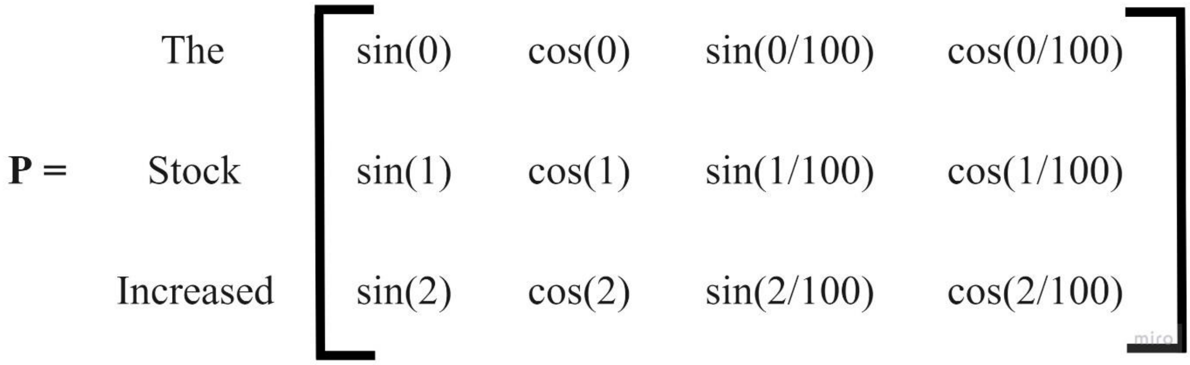
The positional encoding matrix is calculated using the following sinusoidal function:

$$P(pos, 2i) = \sin\left(\frac{pos}{100^{\frac{2i}{d_{model}}}}\right)$$

$$P(pos, 2i+1) = \cos\left(\frac{pos}{100^{\frac{2i}{d_{model}}}}\right)$$

Where, **pos** relates to the position of a word in a sentence and **i** relates to the position of the embedding. Furthermore, we use the sine function when **i** is even and the cosine function when **i** is odd. Therefore, for the example giving above we can compute the positional encoding matrix as follows:

Figure 3.23 : Calculating the Positional Encoding Matrix



$$\mathbf{P} = \begin{array}{r} \text{The} \\ \text{Stock} \\ \text{Increased} \end{array} \begin{bmatrix} \sin(0) & \cos(0) & \sin(0/100) & \cos(0/100) \\ \sin(1) & \cos(1) & \sin(1/100) & \cos(1/100) \\ \sin(2) & \cos(2) & \sin(2/100) & \cos(2/100) \end{bmatrix}$$

Notes: This figure shows how each value within the positional encoding matrix is calculated using the sine and cosine functions provided immediately above this figure.

We know that in our example **the** is in the 0th position, **stock** is in the 1st position and **increased** is in the 2nd position. In the below matrix, P, I have substituted in each of the correct values that would be used in this instance:

Figure 3.24 : Positional Encoding Matrix Containing Numerical Values

$$P = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0.841 & 0.54 & 0.01 & 0.99 \\ 0.909 & -0.416 & 0.02 & 0.99 \end{bmatrix}$$

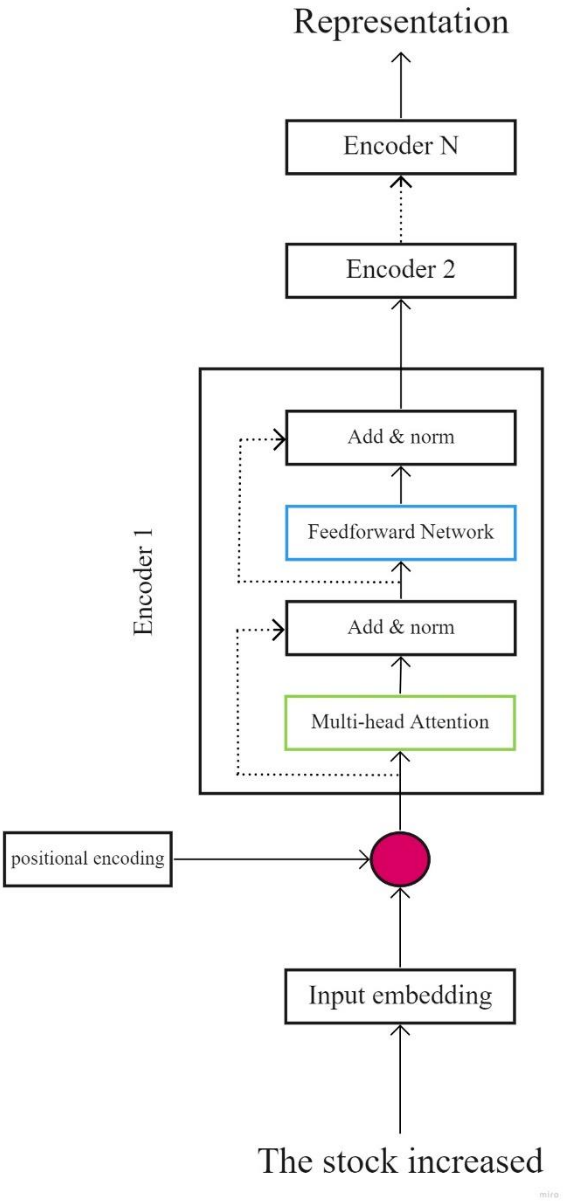
Notes: This figure provides an example of a positional encoding matrix that would be used in a BERT model. It contains the values that were calculated using the equations within figure 3.23.

Then element-wise addition is performed with the embedding matrix X as seen in the X+P. This modified input matrix is then fed to the encoder allowing the transformer to understand the word positioning of each word within the sentence before completing multi-head attention.

The last two sublayers that require an understanding to fully interpret how the encoder layer of a transformer works are the feed forward network sublayer and the add and norm component sublayer. The feedforward layer within the encoder is used to process the output from one attention layer and process it in a way to better fit the input for the next attention layer. It consists of two dense layers with ReLU activations. The parameters of the feedforward network are the same over the different positions of the sentence and different over the encoder blocks. Furthermore, the add and norm component layer connects the input and the output of a sublayer using a residual connection followed by layer normalization. Its main job is to connect the input of the multi-head attention sublayer to its output and connect the input of the feedforward sublayer to its output.

Therefore, from the layers and sublayers discussed above we can piece together how a transformer's encoder works. Firstly, input embeddings are generated for a given input sentence. Then a positional embedding matrix is added to the initial input embeddings matrix to give the model a sense of word positioning. Once the positional encoding has been added to the original embeddings, the new matrix is fed into the first encoders multi-head attention sublayer. The steps above calculating multi-head attention are completed to allow the model to understand word relationships and create dynamically informed representations. From here the add and norm component is used to connect the multi-head attention's input to its output and feed into the feedforward layer. Again, the add and norm component is used to feed the output of the first encoder to the input of the second encoder. This process continues for each encoder until the final encoder outputs the final representation of the input sentence. Tying all the above components together we then have the encoder of a transformer architecture:

Figure 3.25 : Full BERT Process



Notes: This figure gives a visual representation of the full process used within the BERT model starting from an input sentence all the way through to the final robust numerical representation of said input sentence.

BERT follows the exact same method as above with the only difference being that it reads the input bidirectionally – it reads the sentence forwards and backwards. There are multiple configurations of BERT created by the google language team however the model that I shall be adopting is FinBERT. Yang et al. (2020) introduce FinBERT as a financially pretrained BERT model that adopts the same configuration as BERT-Base that is further pre-trained on 4.9 billion tokens from financial corpora, including corporate reports, earnings conference call transcripts and analyst reports. Thus, giving it a more context specific understanding of the language used in the financial domain.

3.7 Summary

This chapter gives a detailed insight into the selection, manipulation and application of the data and sentiment analysis methods used within this thesis. Section 3.3 discusses the selection process used to identify a suitable financial disclosure and financial market in which the multimodal sentiment classifier can be applied to. A visual depiction of the overall system relating to data collection, manipulation and creation is then created and discussed in section 3.4. Section 3.5 then gives an in-depth explanation of how paralinguistic features were generated from our earnings call dataset using text-audio alignment and speech analysis software. Finally, section 3.6 gives a comprehensive explanation of the inner workings of the sentiment analysis model, FinBERT used to create numerical representations of the earnings call textual data. In doing so section 3.6 extensively discussed transformer architecture and particularly identifies how the encoder half of a transformer works and allows all models that use it to be so effective.

4. A Multimodal Sentiment Classifier for Financial Decision Making

4.1. Introduction

Over the last two decades, researchers have used various sources of textual data to determine statistically and economically significant drivers of financial market activity, including the use of sentiment analysis methods to uncover the underlying attitudes held towards an entity (Soleymani et al. 2017). Early applications of sentiment analysis methods, for example Antweiler and Frank (2004), have inspired a wide-ranging field of analysis which incorporates various sources of qualitative information, such as company filings, public news and social media. The literature now encompasses several sentiment analysis methods ranging from traditional dictionary approaches to more computationally demanding machine learning (ML) and deep learning (DL) methods, with recent evidence suggesting that advanced ML and DL approaches are more robust at capturing financial sentiment than the commonly used dictionary approach methods (Kearney and Liu, 2014; Guo, Shi and Tu, 2016; Renault, 2017).

Despite the recent emergence of ML and DL methods, El-Haj et al. (2018) identify that the application of Natural Language Processing (NLP) for sentiment classification to support financial decision making is far less advanced than in other disciplines. Specifically, the authors identify a scarcity of ML application in financial decision support mechanisms, with the majority who do apply ML techniques using comparatively basic classifiers, such as Naive Bayes (Antweiler and Frank, 2004; Li, 2010; Sprenger et al. 2013). While Naive Bayes and other similar algorithms are less computationally demanding and have been shown to classify financial sentiment with a considerable degree of accuracy, alternative approaches such as DL have been demonstrated as having greater abilities in classifying sentiment across different domains (Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Alamoudi and Alghamdi, 2021).

A key limitation of existing sentiment analysis methods in support of financial decision making is the near-exclusive focus on the textual modality which omits potentially important audio cues. The sole exception in this respect is Mayew and Venkatachalam (2012), who leverage vocal cues to assess managers' affective states and evaluate associations with securities pricing behaviour, finding vocal cues to be effective indicators of emotion. This is to some extent consistent with Mehrabian's (1968) 7-38-55% rule inferring that – where words, vocal tone and non-verbal cues indicate inconsistent emotions or attitudes – the proportion of communication attributed to non-verbal cues outweighs vocal tone and words. A panacea of sentiment analysis would therefore involve the incorporation all three modalities, allowing for classification of sentiment in the same fashion as humans.

Though the feasibility of conducting sentiment analysis which incorporates words, vocal cues and non-verbal cues is limited by a lack of relevant financial disclosures conveying information across all three modalities, this chapter presents a significant step forward through the treatment of corporate earnings calls as a dual modality (audio-textual) disclosure. Specifically, in a unique contribution to the literature, this chapter establishes a DL-enabled multimodal sentiment analysis classifier, trained on a sample of corporate earnings calls concerning S&P 100 firms, and measures the incremental effectiveness of incorporating the audio modality into financial sentiment analysis through a comparison with well-established sentiment analysis classifiers. With prior literature showing both textual and vocal characteristics on earnings calls being informative of market characteristics, and following Mehrabian's (1968) rule, the model applied here merges both modalities to create more accurate classification methods, which could allow for a greater understanding of the dynamics between sentiment and market characteristics, and thus represents an observable future direction for the literature. The combination of both textual and vocal modalities has been used within the computer science literature to create more robust measures of sentiment (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021) but – to the authors' knowledge – this research represents a first application for financial decision support.

This chapter directly addresses an existing gap by adopting state-of-the-art techniques from NLP to define financial sentiment. Specifically, introducing a multimodal sentiment analysis method that leverages a financial version of Bidirectional Encoder Representation Transformations (BERT),[185] *FinBERT*, alongside paralinguistic features and a DL classifier. This chapter establishes comparisons between this newly established classifier and methods that are already established within the previous literature, namely the general dictionary approach, the specific dictionary approach, and the ML approach, as well as more recently adopted methods such as BERT. In this respect, this chapter builds

[185] See Devlin et al. (2019) for an in-depth explanation of this transformer model and Yang et al. (2020) for an explanation of FinBERT - the financial version of BERT.

on the disclosure sentiment comparison of dictionary and machine learning methods established by Frankel et al. (2022).

The results are promising, and shed new light on the specific aspects of earnings conference calls for which the various classifiers perform well, or poorly. Firstly, it is found that the multimodal classifier achieves an out of sample accuracy rate of 74.88%, where more traditional classifiers achieve accuracy of between 42.18% and 56.40%. This figure is particularly impressive when compared to more recent approaches such as BERT (65.17%) and FinBERT (73.46%). When isolating the audio modality, it is found the accuracy rate to be comparatively poor (34.12%), suggesting that accuracy is primarily determined by the textual modality, but is enhanced by the inclusion of audio characteristics. At a more granular level, findings indicate that the multimodal classifier is particularly adept at detecting positive and negative sentiments relative to existing alternative methods, although classification accuracy of the neutral class is more in line with modern alternatives such as BERT. The classifier also performs comparatively well at identifying sentiment within the Q&A section of the earnings call, specifically concerning managers' sentiment. However, the multimodal classifier is dominated by BERT and FinBERT at classifying sentiment during the management discussion aspect of the call. Comparatively, traditional classifiers underperform across all sentiment categories and earnings call sections.

This chapter progresses the sentiment analysis methods through the identification that the addition of paralinguistic data to create a multimodal DL-based sentiment classifier is more advantageous than a text only classifier. The findings have implications for the research area of financial sentiment analysis, and potentially pave the way for the advancement of multimodal techniques. The remainder of the Chapter is organised as follows. Section 4.2 identifies the stages involved in the development of the aligned audio-textual dataset and provides an overview of the data used for this comparative analysis. Section 4.3 describes the multimodal classifier design, and outlines each of the benchmark sentiment analysis methods used for comparison. Section 4.4 discusses the accuracy results of the multimodal classifier, with additional testing, before Section 4.5 concludes.[186]

4.2. Data

For the analysis, this chapter takes a snapshot of S&P 100 constituents in 2021 and uses the financial database *Finnhub* to obtain corporate earnings call transcripts and accompanying audio recordings for each firm. Company earnings calls were downloaded for the sample period beginning in 2005 and ending in 2021.[187] Due to computational intensity associated with the alignment of textual and

[186] A version of this chapter has been published in the International Review of Financial Analysis titled "A Multimodal Sentiment Classifier for Financial Decision Making".

[187] Analysing earnings calls over a 17-year period introduces certain challenges, including shifts in macroeconomic conditions, regulatory environments, communication norms, and audio quality. These factors can introduce heterogeneity in both content and delivery. However, this thesis develops a general-purpose multimodal sentiment classifier designed to operate robustly across varying contexts. As a proof-of-concept analysis, the focus is on demonstrating the viability and potential of multimodal sentiment classification in financial settings. Accordingly, while temporal differences exist, they are secondary to the broader objective of evaluating the effectiveness of multimodal approaches in capturing market-relevant sentiment.

audio data (to be described), it is necessary to proceed with a sample of twenty firms from the initial list, selecting four firms at random from each quintile when ranking all companies by market capitalisation. This ensures the sample selection is conducted across the firm size distribution. Thus, the final dataset consists of text and audio files corresponding to corporate earnings call transcripts for twenty sample firms across the seventeen-year sample period.

Next, the chapter focuses on the generation of paralinguistic features, such as tone and pitch, associated with each earnings call sentence. This process consists of two stages. Firstly, aligning earnings call transcripts with the corresponding audio file to generate sentence level audio clips through a process of "iterative forced alignment" (Chapter 3.5). Secondly, generation of paralinguistic features from the sentence level audio clips using a widely used phonetics tool (Chapter 3.5). For the iterative forced alignment process, this research follows the method previously established by Li et al. (2020), using Python to leverage the iterative forced alignment package *Aeneas*.[188] However, this analysis differentiates the approach by focusing on all earnings call participants (managers and analysts) and directly aligning calls at the sentence level. Comparatively, Li et al. (2020) narrow their focus to the "Management Discussion" section of the earnings call, and adopt a two-stage alignment process, aligning first at a paragraph before subsequently aligning at the sentence level.

4.2.1. Dataset Characteristics

The final dataset consists of 711,031 text-audio aligned sentences taken from corporate earnings call transcripts for twenty constituent firms of the S&P100 index (as of 2021) between 2005 and 2021. From this dataset, a randomly sampled subset of 2,106 manually classified (as positive, negative or neutral) messages containing an equal number of sentences from all three sentiment categories (702 sentences per category) is selected.[195] This approach is adopted as ML and DL algorithms are known to be less predictive on unseen data when trained on imbalanced dependent variable sets, and often become biased towards the overrepresented data category (Rezapour, 2020).

Table 4.1: Summary of Earnings Call Sentences

Panel A: By Year and Call Section

Year	No. Sentences	MD (%)	Q&A (%)	Avg. Sentiment	Avg. Audio (s)
2005	8	100.00	0.00	0.50	12.75
2006	46	36.96	63.04	0.33	19.86
2007	116	41.38	58.62	-0.03	14.00

[188] https://github.com/readbeyond/aeneas.

[195] For robustness, this chapter employed alternative training sets, such as (i) an unequally weighted random sample of 10,000 earnings call sentences and (ii) a random sample of 2,000 earnings calls sentences for which sentiment was independently assigned, cross-referenced and agreed upon (within an x% tolerance) between two students. Though the accuracy rates of the classifiers are found to differ to some extent dependent on the dataset used, the rankings of classifiers in our comparative analysis remain similar. Appendix 4.4 includes key results for sentiment classification using alternative approaches (i) and (ii) in the supporting material for the reader's perusal.

2008	176	22.73	77.27	-0.19	14.98
2009	173	8.67	91.33	-0.18	17.73
2010	135	14.81	85.19	-0.01	14.97
2011	140	1.43	98.57	-0.04	11.43
2012	152	0.66	99.34	-0.03	13.78
2013	136	1.47	98.53	0.07	15.12
2014	157	8.28	91.72	0.06	18.26
2015	153	0.00	100.00	-0.08	18.81
2016	147	1.36	98.64	-0.03	14.72
2017	114	1.75	98.25	0.17	17.58
2018	117	0.00	100.00	0.17	15.43
2019	155	0.00	100.00	0.03	19.06
2020	149	0.00	100.00	0.07	18.51
2021	32	0.00	100.00	0.13	15.72
Total	**2106**	**8.07**	**91.93**	**0.00**	**16.16**

Panel B: By Call Section and Speaker

		Call Section		**Speaker**	
Sentiment	**N**	**MD (%)**	**Q&A (%)**	**Mgmt (%)**	**Analyst (%)**
Positive	702	9.40	90.60	55.56	44.44
Negative	702	7.83	92.17	50.57	49.43
Neutral	702	6.98	93.02	54.99	45.01
Total	**2106**	**8.07**	**91.93**	**53.70**	**46.30**

Notes: This table disaggregates the extracted 2,106-sentence sample across the years from which the earnings call took place (Panel A), and across the type of speaker delivering the sentence (Panel B). Panels A and B show the number of sentences associated to each section of the call, where 'MD (%)' and 'Q&A (%)' show the proportion of call sentences that took place during the Management Discussion and Q&A sections of the calls, respectively. For Panel A, 'Avg Sentiment' shows to the arithmetic average sentence sentiment score, where a score of +1 (-1) corresponds to a positive (negative) score. 'Avg Audio' shows the average recording length, in seconds (s), for each sentence. For Panel B, 'Mgmt (%)' indicated the proportion of sentences delivered by firm managers on the call, and 'Analyst (%) shows the proportion delivered by analysts.

A breakdown of the 2,106 text-audio aligned sentences used within this chapter is provided in Table 4.1. The number of sentences per year remains quite consistent throughout the timeframe used, with the exception of the years at the beginning and end of the sample period (2005, 2006 and 2021). Furthermore, 91.93% of sentences are sampled from the Question-and-Answer (Q&A) section of the earnings call, suggesting that the majority of earnings call discussion centres on the discussion between managers and analysts. Though the proportion of Q&A sentences used in the dataset remains relatively consistent from 2009 onwards, the earliest years in the sample (2005, 2006 and 2007) are dominated by sentences occurring within the management discussion section of the call. The lowest average sentence sentiment per year is identified in 2008 (-0.19) followed by 2009 (-0.18), which is to some extent expected given that these earnings conference calls took place amidst the global financial crisis. The highest annual average sentiment (+0.50) occurs at the beginning of the sample period (2005).

To offer context surrounding the manual classification of sentences within the sample, messages were manually classified as containing “positive”, “neutral” or “negative” sentiment.[196] Sentences are classified as negative (positive) if they have unfavourable (favourable) connotations towards the performance of the firm engaged in the earnings call. Sentences are classified as “neutral” if they do not contain any significant information regarding beneficial or suboptimal firm performance.[197] Then an assessment of the proportions of positive, neutral and negative sentiment across each category was conducted.[198] The results indicate that proportions across categories are similar, however there is a slight skew towards positive messages in the MD portion of the call, and similarly a slight skew towards negative speech from analysts. This is consistent with prior studies finding that managers speak with significantly greater optimism than analysts on earnings call (Brockman, Li and McKay Price, 2015), perhaps due to a managers’ preference for positive framing when disseminating corporate performance information. Furthermore, Renault (2017) finds that call sentiment is more positive towards the beginning of an earnings call due to managerial introductions, before becoming more balanced later in the call, when financial analysts begin questioning managers.

The sample of text-audio aligned sentences are pre-processed and split into training and validation sets, where pre-processing includes the removal of special characters, and the transformation of audio data using a scaling function[199] that scales all features to within the range zero to one. Following pre-processing, sentences are randomly assigned to training and validation sets using a stratified train-validation split of 80% training and 20% validation data.[200]

4.2.2. Paralinguistic Data

A key contribution of this chapter in respect to multimodal analysis for financial decision making is the use of paralinguistic data in the novel sentiment classifier. The audio features used for the purposes of training and testing our classifier are a subset of audio features provided by the phonetics library *Praat*, namely: mean pitch, mean intensity, number of periods, fraction of unvoiced, number of voice breaks, jitter local, shimmer local, mean autocorrelation, mean noise-to-harmonics ratio and audio length. A definition for each feature is provided in Table 4.2. The features represent a subset of a larger set of audio features output by *Praat* that were selected based on multicollinearity tests for each

[196] Examples of positive, negative and neutral messages along with examples of hard to classify messages are provided in Appendix 4.1.

[197] The sentences were manually classified by two researchers. The final dataset of 2,106 sentences only included sentences where the sentiment was agreed upon separately and independently by both researchers.

[198] A breakdown of sentence sentiment relating to call sections and participants after manual labelling is also provided in the supporting material (Appendix 4.2).

[199] *MinMaxScaler* from python’s scikit-learn library.

[200] The dataset is divided into training and validation sets to evaluate the performance of the multimodal sentiment classifier. The training set is used to fit the model—allowing it to learn patterns from the combined text and audio features. The validation set, which the model has not seen during training, is then used to assess how well the model generalises to unseen data. This split helps prevent overfitting and provides a more realistic estimate of how the classifier would perform in practical applications. A stratified split was applied to ensure a balanced class distribution as stratification of the data preserves the same proportion of sentiment categories across both the training and testing set.

variable: features were only included in our classifier if they were not characterised by strong correlations with other variables, to reduce the amount of noise introduced to the multimodal model.[201]

Table 4.2: Definition of Paralinguistic Features

Feature	Description
Mean Pitch	Quality of sound, governed by the rate of vibrations produced; the degree of highness or lowness of a tone.
Mean Intensity	Acoustic power carried by sound waves per unit area in a direction perpendicular to that area.
No. of Periods	Frequency of vibration cycles per second.
Fraction of Unvoiced	Percentage of an audio segment which is unvoiced
No. of Voice breaks	Number of distances between consecutive pulses that are longer than 1.25, divided by the pitch floor.
Jitter Local	Average absolute difference between consecutive periods, divided by the average period.
Shimmer Local	Average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.
Mean Autocorrelation	The mean (over all analysed time points) of the autocorrelation, ranging between 0 (theoretical white noise) and 1 (perfectly periodic signal).
Mean NHR	A 'noise-to-harmonics' ratio between periodic and non-periodic components of speech.
Audio Length	The length of each audio clip in seconds

Notes: The above table offers concise definitions of each audio variable included in our multimodal sentiment classifier model. Justification for the inclusion of each variable can be found within this section.

Existing research suggests that the content of what we say matters, and that the way in which we communicate matters (Guyer, Fabrigar and Vaughan-Johnston, 2018). More succinctly, how we speak conveys substantial information beyond the content of communication. Indeed, there is a body of psychology literature that relates to vocal characteristics and their impact on persuasion and/or decision making. For example, evidence suggests that vocal pitch can impact upon listeners' perceptions of speakers' personal traits and qualities. Qualities such as credibility, tranquillity, persuasion, trustworthiness and maturity are associated with a lower level of vocal pitch. Conversely, high pitch voices are considered immature, nervous, informal, less credible and less persuasive (Chattopadhyay et al. 2003; Feinberg et al. 2005; Klofstad, Anderson and Peters, 2012; Martín-Santana et al. 2015; Wang et al. 2018; Chau et al. 2020; Song et al. 2020).

From a decision-making perspective Chua et al. (2020) suggest that higher pitched voices increase risk aversion in the listener whilst a lower pitched voice heightens risk tolerance. In a professional setting, Sorokowski et al. (2019) show that both men and woman demonstrate a tendency

[201] A list of all of all available paralinguistic features prior to multicollinearity testing is provided in the supporting material (Appendix 4.3).

to lower their mean pitch in an authoritative context, with this effect more pronounced for women. Gelinas-Chebat et al. (1996) define intonation as the variation of pitch which reflects a voice's melodic contour, with the authors noting that a higher intonation is associated to some degree with a lack of self-confidence, where lower intonation is thought to reflect self-confidence and competence (Brooke and Ng, 1986; Wallbott, 1982; Apple et al. 1979).

Prior studies also suggest that a higher intensity of vocal signal, which can also be defined as signal loudness (Gelinas-Chebat et al. 1996), creates a perception of credibility and trustworthiness and a perception to listeners that the speaker has a greater efficiency in articulating their arguments in comparison to softer spoken speakers (Erickson, Lind, Johnson and O'Barr, 1978; Conley, O'Barr and Lind, 1979; Brooke and Ng, 1986; Bradac, Mulac and House, 1988). Van Zant and Berger (2020) study the paralinguistic cues used by speakers in nonverbal persuasion attempts, finding evidence that speakers: (i) spoke at a higher volume (greater intensity); (ii) spoke at a higher pitch; (iii) varied their volume more; and (iv) spoke at a faster rate. The authors further suggest that speakers are more persuasive when speaking with greater intensity and more varied volume.

Jitter and shimmer audio features have also received some attention with regards to their impact on perceptions about a speaker, with studies incorporating these measures in relation to their characterisation of stress, and finding that jitter and shimmer features diminish during experimentally induced stress (Mendoza and Carballo, 1998; Park et al. 2011; Giddens et al. 2013). Furthermore, both metrics have been shown to be important variables for the analysis and classification of speaking style. For example, Li et al. (2007) show that the addition of both features to a classification model results in increased classification accuracy, in comparison to a model that only contained baseline spectral and energy features.

The remaining audio cues have received comparatively little coverage in the existing literature, in terms of speakers' perception and listener influence. However, prior literature has shown that, along with the features discussed above, the remaining features are beneficial for machine emotion classification as they increase classification rates. The "fraction of unvoiced" (and the number of voice breaks) reflecting the proportion (and number) of pauses within an audio clip, represent commonly used vocal characteristics in emotion classification literature. For example, Morrison, Wang and De Silva (2007) adopt seven vocal characteristics from calls received by call centres to classify the emotion displayed on the calls, which includes a "fraction of unvoiced" variable. The authors adopt multiple classification models to assess which model returns the highest accuracy for the task at hand. After performing baseline classifications with each model, feature selection techniques were applied to assess which audio features were optimal to include. For all models, the fraction of unvoiced variable was included after feature selection, highlighting the variable's robust correlation with accurate emotion classification.[202]

Using a sample of 10,000 video clips extracted from social media platforms, Morency, Mihalcea and Doshi (2011) classify the sentiment of each video using the fusion of text, audio and visual data. The authors created a proof of concept which suggests that a multimodal approach is effective in identifying online video sentiment. To do so, the authors draw textual and visual data from the videos alongside two audio features: namely pause duration and pitch. The findings suggest that classifying emotion using a tri-modality approach outperforms each of the three modalities in isolation. Furthermore, Poria, Cambria and Gelbukh (2015) incorporate pauses into their vocal modality data for multimodal classification, showing that the inclusion of the vocal modality increases classification accuracy. Indeed, even the worst results obtained using two modalities still outperformed the best unimodal accuracy.

The remaining variables, autocorrelation and noise-to-harmonics ratio (NHR), are quite commonly used in computer emotion and speech recognition tasks (Nwe, Foo and De Silv, 2003; Rong, Li and Chen, 2009; Lee, Kim and Kang, 2014). The benefits of including said variables are demonstrated by Noroozi et al. (2017), who use a similar range of paralinguistic variables for a vocal-based emotion recognition task. The authors improve on past accuracies using this set of paralinguistic variables, highlighting that their enhanced performance is in part due to their smaller set of features, which lowered model complexity whilst improving computational speed and thus emotion classification.

The literature discussed above accentuates the benefits of including each paralinguistic feature within a sentiment classifier as they have been shown in various studies to carry informative insights surrounding the emotional state of the speaker. Therefore, each of these paralinguistic cues have been incorporated into this finance-specific classifier, presenting an ability to assess the extent to which vocal characteristics that have been shown to be predictive of emotional states in other domains are still relevant in a financial setting.

4.3. Methods

In this section, the multimodal sentiment analysis model is first introduced, before discussing the various sentiment analysis classifiers that are employed to provide effective performance benchmarks. The benchmark models used in this case are amongst the most commonly used in past finance literature, ranging from long-standing popular methods such as dictionaries, to more recent, state-of-the-art transformer models. The following subsection 4.3.2 provides an overview of each model type, but also provides appropriate citations for those wishing to gain a greater understanding of the technical aspects

[202] The authors strongest result came from a voting classifier model that adopted forward selection, which returned an accuracy of 79.43%.

of the respective classifiers. For transparency, an overview of the settings and parameters used in each case are provided.

4.3.1 Multimodal Sentiment Classifier

The multimodal sentiment classifier introduced by this thesis leverages the FinBERT transformer model that has gained popularity in recent years, and for which a full overview is provided in the later discussion of benchmark models (Chapter 4.3.2) and the previous Chapter. The multimodal classifier, however, also incorporates the audio modality from corporate earnings calls. The inclusion of nonverbal cues in studies of financial decision making is almost non-existent (Mayew and Venkatachalam, 2012), despite a relatively high level of adoption in other academic domains. For example, studies within the psychology domain highlight the ability of vocal attributes to uncover the underlying meaning of a message, with Mehrabian's (1968) 7%-38%-55% rule accentuating the lack of information conveyed through the textual modality alone. Furthermore, existing evidence suggests that a combination of text and audio data improves classification accuracy, and consequently creates a more robust representation of sentiment (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021). Given that both textual and vocal characteristics of earnings calls have been found to be informative (Mayew and Venkatachalam, 2012), and that NLP literature finds a combination of text and audio to significantly increase classification accuracy, the merging of both measures represents an important innovation.

The multimodal model takes numerical representations created using FinBERT from the textual sentences[203] along with the sentence-level numerical representations of paralinguistic features (see Table 4.2) as input into a deep neural network classifier (DNN) to make sentiment predictions. To assist in the model building, the optimal hyperparameters of the DNN were identified using the Random Search approach (available in the 'Keras Tuner' Python library). Rather than using a trial-and-error method, the Random Search evaluates multiple configurations of layers and nodes to return the optimal set up for a specific DNN problem. The random search optimiser arrives at an optimal structure that has an input layer of 500 nodes, two hidden layers consisting of 250 and 125 nodes, respectively, and an output layer consisting of three nodes. The parameters of the DNN are the same as the neural network used for both BERT and FinBERT discussed below, namely a ReLu activation function in the input layer and hidden layers with a softmax activation function for the output layer. Similarly, the multimodal model deploys a he_uniform kernel initializer and an adam optimizer to evaluate the accuracy metric. The DNN uses a batch size of 150 epochs.

In order to add a degree of robustness to the reported accuracy of the new multimodal sentiment analysis model, two additional models to capture the information contained within paralinguistic data are employed: an 'audio only' classifier and a multimodal FinBERT neural network (NN) classifier. The audio classifier takes the audio attributes incorporated into the multimodal classifier in isolation,

[203] The calculation of numerical representations from textual data are explained in Chapter 3.6.

inputting them into a neural network. By doing so, this analysis provides an understanding of the extent to which audio features alone are reliable in predicting sentiment, in comparison to the multimodal model and the text-based benchmark classifiers introduced in Section 4.3.2. This should allow for interpretation as to whether audio features are more effectively considered on their own, or in combination with the textual modality. The second robustness model, the multimodal FinBERT classifier, uses the same neural network and textual features used for the FinBERT model, but includes the addition of paralinguistic features to ensure that any increase in performance is the result of audio feature inclusion, rather than the use of a DNN.

4.3.2 Benchmark Models

4.3.2.1. Dictionaries

Many of the earliest studies employing textual analysis techniques adopt a dictionary, also known as "word count", approach (Guo, Shi and Tu, 2016; Loughran and McDonald, 2016). The concept behind this approach is comparatively intuitive in light of more recent machine learning methods, in that the sentiment conveyed within a financial text is determined by a count of words within the text that also appear within pre-defined word lists (Li, 2010). There are two variations of the dictionary approach: general and domain specific. Bhonde et al. (2015) note that, for the general approach, word lists are created by first collecting a set of general sentiment words with known positive or negative implications. Once this initial list is created, it is then expanded upon by including synonyms and antonyms for the sentiment words. This iterative process of expanding the word lists ends when no new words can be found. After the process of collecting synonyms and antonyms ends, an inspection of the words is usually completed to clean up the lists. Kearney and Liu (2014) highlight the Harvard IV psychosocial dictionary as the most used general dictionary within a financial decision-making context, which was originally developed for content analysis in the social psychology domain.

The domain-specific approach was introduced to the literature to overcome the domain-specificity limitation. For example, Loughran and McDonald (2011) demonstrate that general dictionaries misclassify words used within a financial context, noting that 73.8% of negative words within the Harvard dictionary are not considered negative in a financial context. Gonzalez-Bailon and Patloglou (2015) and Ribeiro et al. (2016) also demonstrate the limitations of general dictionaries in classifying content in domain-specific settings by applying the general dictionary to text stemming from various domains and showing that the reliability and validity across these differing sets is low. A solution to this issue is the creation of domain-specific dictionaries, where adding words to an existing dictionary and deleting irrelevant words (or words with different meanings) within a specific context would be beneficial (Grimmer and Steward, 2013; Diesner and Evans, 2015). Loughran and McDonald (2011) create a finance-specific set of dictionaries that have been widely used to classify and analyse financial sentiment.

For both dictionary approaches the way to define sentiment is similar, with the only difference being the words contained within each dictionary type. Abirami and Gayathri (2016) highlight that the most basic way to define sentiment using these dictionaries is to count the number of positive and negative words within a body of text with reference to the dictionary categories. After this count is complete, a comparison of how many positive versus negative words in the text infers how positive or negative the text is.

Utilisation of dictionary methods presents advantages in comparison to machine learning techniques, including the comparatively lower computational power (or resource) required to create and apply dictionary methods. However, there are also considerable drawbacks to this approach. For example, the lexicons are characterized by a finite number of words and the sentiment orientation for each word is fixed, resulting in a lack of classification accuracy (Bhonde et al. 2015; D'Andrea et al. 2015; Abirami and Gayathri, 2016). For the purposes of this chapter, both a general and domain-specific dictionary were applied to the training and validation sets, namely the Harvard-IV4 (general) and Loughran and McDonald (domain-specific) dictionaries.

4.3.2.2 Machine Learning (Naïve Bayes)

Machine Learning (ML) algorithms aim to classify sentiment through the inspection of a set of numerical features that have been created to represent text. In this study, the ML algorithm attempts to assign a sentence to one of three classes (positive, negative, and neutral) based upon a set of numerical features representing each word in the associated sentence. The majority of ML algorithms for sentiment analysis create numerical features related to text based on a fixed representation approach, whereby each word is given a numerical representation, and every word is then populated with the associated representation. Renault (2017) highlights the stages of ML classifier use. Firstly, the classifier is trained on pre-classified data. Second, the model makes predictions on a test dataset using the understanding it has gained from the training phase, allowing users to understand the performance of the classifier.

A probabilistic Naïve Bayes classifier is used to evaluate the accuracy of ML sentiment classification of our dataset as it is amongst the most commonly adopted SL approaches for sentiment analysis. The Naïve Bayes method estimates the probability of a document's sentiment given its contents. Specifically, it estimates the probability of a word's sentiment by looking through a series of positive and negative texts and counting how often the word appears in each (Troussas et al. 2013). Pang, Lee and Vaithyanathan (2002) provide a comprehensive overview of the Naïve Bayes classifier, noting that even though the model is simplistic in process and is based upon the Naïve "bag of words" assumption of the model, Naïve Bayesian models generally perform well. Dey et al. (2016) note that a benefit of the Naïve Bayes classifier is that it only requires a small amount of training data to establish parameters necessary for classification, though Pang et al. (2002) highlight that more sophisticated algorithms have the potential to yield better results in sentiment classification tasks. For the purposes of this chapter, both the training and validation sets were tokenized and stop words were removed before using the Naïve Bayes classifier. Tokenized sentences were then transformed into an array data structure to be input into the SL algorithm for training and validation.

4.3.2.3. Deep Learning (BERT and FinBERT)

The deep learning (DL) approach to sentiment classification used for this chapter is based upon transformer architecture introduced by Vasawani et al. (2017). Though various DL approaches exist for sentiment classification, such as Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTM), and Convolution Neural Networks (CNNs), the architecture introduced by Vasawani et al. (2017) has been implemented within many models[204] that have achieved state-of-the-art results across a number of NLP tasks. This analysis employs BERT[205] and FinBERT[206] as benchmarks representing DL approaches to classifying financial sentiment. At a high level, the aim of both BERT and FinBERT is to create robust numerical representations of textual data to allow ML and DL models to understand text, and make classifications from the textual representations. The fundamental difference between BERT and FinBERT is the data that each model is pretrained on: BERT is pretrained on 800M words from BookCorpus and 2,500M words from Wikipedia (Devlin et al. 2019) whereas FinBERT is pretrained on 2.5B tokens from Corporate Reports, 1.3B tokens from earnings conference calls and 1.1B tokens from analyst reports (Yang, Christopher and Huang, 2020).[207]

In this analysis both BERT and FinBERT were applied to the training and validation data sets to generate sentence representations, where each textual sentence is returned as an array containing 768 features. The training representations were then input into a neural network classifier to recognise patterns between the representations and associated sentiment classification (positive, negative or neutral). The neural network was then applied to the validation set to assess its accuracy. The neural network used for both BERT and FinBERT contains an input layer of 768 nodes, mirroring the size of the features being fed into the model. The neural network adopts a ReLu activation function in the input layer, with a softmax activation function for the output layer. It deploys a he_uniform kernel initializer and an adam optimizer to evaluate the accuracy metric. Furthermore, the network uses a batch size of 110 epochs.

4.4. Results

4.4.1. Classifier Accuracy

This section begins by discussing the results gained from the comparative analysis between those classifiers introduced within earlier sections 4.3.1 to 4.3.2. The analysis first reflects upon the

[204] BERT, GPT3 and T5 all implement transformer architecture and return state-of-the-art results on a plethora of NLP tasks (Liu et al. 2019; Sun et al. 2020; Nogueira and Cho, 2020).

[205] https://huggingface.co/docs/transformers/model_doc/bert

[206] https://github.com/yya518/FinBERT

[207] For greater explanation of these models and how they generate numerical representations from textual data see Yang, Christopher and Huang (2020).

classifiers' ability to accurately predict sentiment across the full dataset of 2,106 earnings call sentences. Following the initial results, the validation data is disaggregated to assess each model's accuracy according to the speaker (managers or analysts) and the section of the call (management discussion versus Q&A). This should allow for a deeper understanding of the areas of corporate earnings call where the different classifiers perform best. The accuracy (AC) of each of the methods discussed in the previous section will be compared directly using the validation accuracy metric (*see* Renault, 2020), where TP represents the number of true positive classifications,[208] TN is the number of true negatives,[209] FP is the number of false positives[210], and FN is the number of false negatives.[211]

$$AC = \frac{TP + TN}{TP + TN + FP + FN}$$

Table 4.3 presents the recorded accuracy for each model. Generally, the results suggest that the classification accuracy increases with the complexity of the method used to define sentiment. Curiously, the audio only classifier is the least accurate method within this study,[212] which is potentially due to the training dataset having been assigned sentiment classifications using textual data. The dictionary approaches, Harvard IV4 and Loughran-McDonald (LM), return the lowest training and validation accuracy amongst textual classifiers, which is consistent with previous evidence suggesting that advanced techniques are more robust at defining financial sentiment than the commonly used dictionary approaches (Kearney and Liu, 2014; Guo, Shi and Tu, 2016; Renault, 2017).

Table 4.3: Overall Call Classification Accuracy Results

	Accuracy	
Model	Training	Validation
Audio Only	40.08%	34.12%
Harvard IV4	42.99%	42.18%
Loughran-McDonald (2011)	49.23%	47.39%
Naïve Bayes	90.14%	56.40%
BERT	74.76%	65.17%

[208] In this multiclass classification task using the One-Vs-Rest method TP is calculated as the number of classifications the model correctly predictions. For instance, there is three classes (Class A, B and C) in this classification tasks and a TP is when the model correctly predictions Class A when it truly is Class A.

[209] Following the definition provided in footnote 206, TN is then defined as the number of times the model predicts *not* Class A and the true class is either Class B or C.

[210] FP is calculated as the number of classifications the model makes as Class A but the true Class is B or C.

[211] FN is calculated as the number of classifications the model makes as *not* Class A when the true classification is Class A.

[212] This being said, it can be seen from Appendix 4.4 that the audio only classifier performs better on the alternative datasets mentioned. In the case of both alternative datasets, the audio only classifier outperforms both dictionary approaches. This is perhaps due to the imbalanced nature of the alternative datasets towards the neutral sentiment category. The audio neural network classifier looks to be heavily overtraining towards neutral sentences and thus returning a higher classification accuracy due to the imbalanced nature of the dataset.

FinBERT	97.80%	73.46%
Multimodal FinBERT NN	**99.76%**	74.64%
Multimodal FinBERT DNN	**99.76%**	**74.88%**

Notes: This table outlines the accuracy for each sentiment classifier when the classifier is used to classify the training dataset of 1,684 sentences (in-sample) and the validation dataset of 422 sentences (out-of-sample). The highest accuracy achieved in each case is signified using bold text.

The ML method, using a Naïve Bayes classifier performs particularly well when classifying the sentiment of training data that it has already seen, (90.14%) however this accuracy level drops considerably when classifying the unseen validation dataset (56.40%), potentially indicating an overfitting issue by the SL model when classifying training data. There is a considerable degree of consistency, in that the greatest accuracies are generated by the four DL methods, with the multimodal FinBERT DNN returning the highest in-sample accuracy (99.76%). This represents outperformance of 1.42% and 9.71% when compared to the single (text) modality FinBERT and BERT methods, respectively. That being said, the performance of all classifiers predictably drops when considering the out-of-sample dataset. Even then, however, the multimodal FinBERT DNN classifier (74.88%) dominates all other textual modality classifiers by between 1.42% (FinBERT) and 32.70% (Harvard IV4 Dictionary).

The FinBERT NN model, which uses text and audio modalities, outperforms the text only FinBERT model by 1.18% out-of-sample, suggesting that the inclusion of the audio modality does allow for a more accurate capture of earnings call sentiment, albeit on a marginal scale. This is potentially important, as it may suggest that multimodal methods can lead to a cleaner measure of sentiment when investigating relationships between financial disclosure and trading behaviour.

Table 4.4: Classifier Accuracy (Out-of-Sample) Disaggregated by Sentiment Category

	Sentiment Category		
Model	**Negative**	**Neutral**	**Positive**
Audio Only	27.66%	48.57%	23.40%
Harvard IV4	26.95%	35.71%	63.83%
Loughran-McDonald	40.43%	73.57%	28.37%
Naïve Bayes	58.87%	38.57%	71.63%
BERT	53.19%	67.14%	75.18%
FinBERT	71.43%	73.05%	75.89%
Multimodal FinBERT NN	70.21%	**75.00%**	**78.72%**
Multimodal FinBERT DNN	**75.18%**	70.71%	**78.72%**

Notes: This table identifies the validation (out-of-sample) accuracy across the three sentiment categories for each method being compared in this study. It highlights what method is most adept at classifying each individual sentiment category and subsequently diagnoses in what areas model fall short.

Table 4.4 reports the out-of-sample classification accuracy across each sentiment category in the validation set. This table allows the analysis to establish whether the multimodal classifier is particularly adept at predicting positive, negative, or neutral earnings call sentences, relative to the benchmark text-modality models. In other words, if the improved performance of the multimodal classifier is driven by one sentiment class – for example, positive sentiment – then this may suggest that audio cues are potentially more informative, or consistent. The results suggest that Multimodal FinBERT DNN returns the highest accuracy for positive and negative sentiment categories, whereas the highest neutral classification accuracy is achieved by Multimodal FinBERT NN. Combined, these results suggest that multimodal classification dominate across all sentiment categories, with recent computationally demanding text-modality models outperforming the traditional dictionary and Naïve Bayesian classifiers.

The results also return evidence suggesting that domain-specific models, on average, outperform general purpose ones, with finance-specific models (Multimodal FinBERT NN, Multimodal FinBERT DNN and FinBERT) achieving the highest classification accuracy amongst all classifiers. Furthermore, the Harvard Dictionary returns the lowest accuracy for negative classifications (26.96%), which is consistent with prior findings that general dictionaries misclassify words used within a financial context. For example, Loughran and McDonald (2011) tailor their dictionary approach to the negative sentiment category,[213] using the rationale that negative sentiment has more influence on trading activity. These results offer support to this approach, indicating that the finance-specific dictionary has a substantially higher negative (40.43%) than positive (28.37%) classification accuracy. The results also suggest that the finance-specific dictionary is better able to classify negative sentiment than the general dictionary (26.95%). Combined, these results highlight the importance of building models on the foundations of appropriate financial context.

Table 4.5: Breakdown of Accuracy across different sections and participants of the call

Model	MD	Q&A	Mgmt	Analyst
Harvard IV4	34.48%	42.75%	41.78%	42.64%
Loughran-McDonald	51.72%	47.07%	47.56%	47.21%
Audio Only	37.93%	33.84%	38.67%	28.93%
Naïve Bayes	65.52%	55.73%	58.22%	54.31%
BERT	72.41%	64.63%	66.22%	63.96%
FinBERT	**93.10%**	72.01%	69.04%	**77.33%**
Multimodal FinBERT NN	89.66%	73.54%	75.56%	73.60%

[213] The negative word list created by Loughran and McDonald (2011) has 2,337 words. When compared to their positive list of 353 words, the negative word list is substantially larger and highlights the authors' focus on the negative sentiment category.

Multimodal FinBERT DNN	89.66%	**73.79%**	**76.00%**	73.60%

Notes: This table identifies the validation accuracy across the different sections and participants on the call for each method being compared in this study. where 'MD' and 'Q&A' show the accuracy rates for earnings call sentences occuring the Management Discussion and Q&A sections of the calls, respectively. 'Mgmt' and 'Analyst' show the accuracy rates for sentences spoken by Managers and Analysts on the call.

Table 4.5 reports classification accuracy of earnings calls sentences, disaggregated firstly by call section (Management Discussion and Q&A), and secondly by the originator of the sentence (Managers and Analysts). The results offer further insights on the relative strengths and weaknesses of each classifier. The multimodal FinBERT DNN method continues to produce strong performance across all categories, and outperforms all other methods when classifying sentences originating from the Q&A section, as well as sentences delivered by manager participants on the call. This suggests that the inclusion of paralinguistic features within sentiment analysis models is beneficial in classifying messages that are spoken in the more conversational context often observed within the Q&A section of an earnings call. This is particularly important, given prior findings that sentiment originating from within the Q&A section of earnings calls has greater predictive power over market characteristics than the management discussion section (McKay Price et al. 2012; Borochin et al. 2017; Fu et al. 2019).

4.4.2. Additional Accuracy Measures

To further test the findings reported in Section 4.4.1, the performance of the multimodal, benchmark and robustness models are evaluated using Receiver Operating Characteristic curve and Area Under the Curve scores (ROC-AUC).[214] Given that ROC-AUC measures require a binary classification problem, and this chapter uses three classification categories (positive, neutral and negative), the methods used to understand each model's ability to classify must be slightly adjusted. Specifically, a 'One versus Rest' (OvR) method, which evaluates each class against all others, was used. The OvR method begins by evaluating the ability of each classifier to correctly assign positive sentiment by converting a correct classification to the positive class to a value of one, and an incorrect classification to the neutral or negative classes to a value of zero. The OvR model then adopts the same process for the neutral and negative categories. The average accuracy score across all three categories is then recorded for each classifier. In adopting this process, the ROC curve identifies the sensitivity (true positive rate) and specificity (true negative rate) of each sentiment classifier.

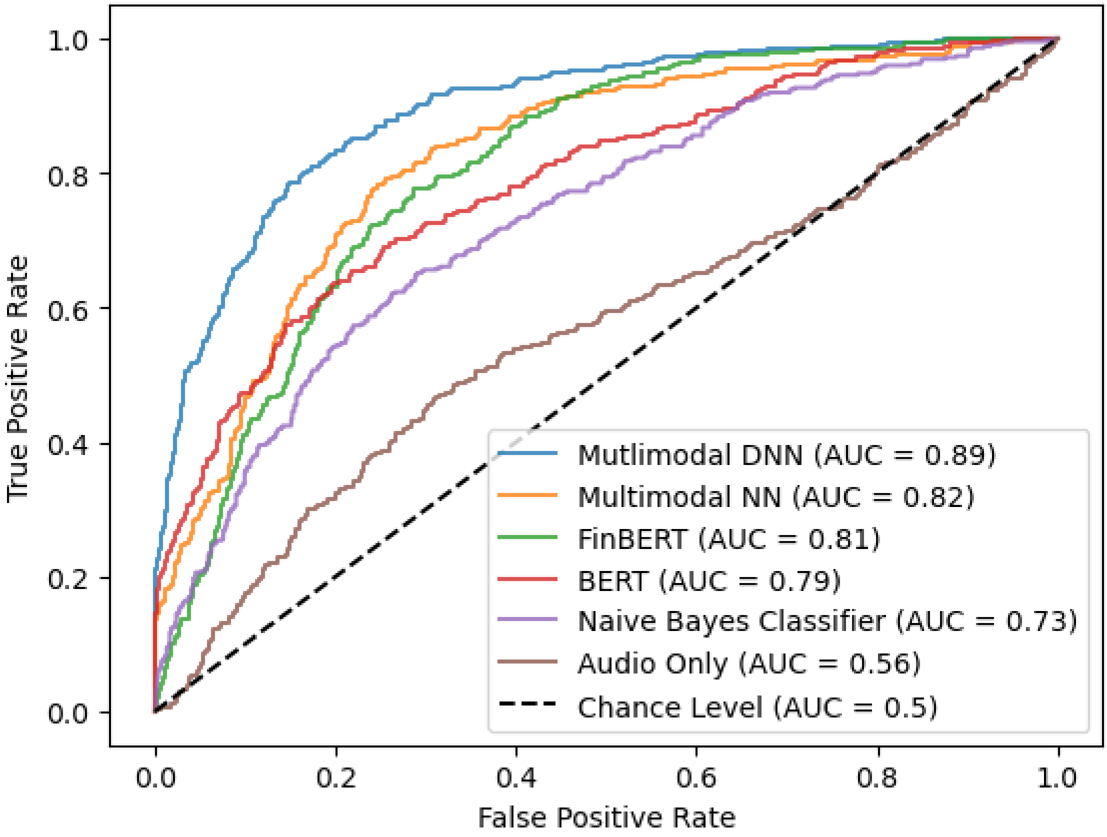
Mandrekar (2010) highlights that an ROC curve that intersects the coordinates (0,0) and (1,1) at a 45-degree angle represents pure chance, with any curve skewing towards the upper left corner of the plot representing a classification accuracy above that achieved by chance. Furthermore, the AUC score is an effective way to summarise the overall accuracy of a model, with an overall score of zero indicating

[214] Due to the configuration of Dictionary Methods, it is not possible to identify create a ROC curve. Hence, they have been omitted from Figure 4.3.

a perfectly inaccurate classifier, and a score of one reflecting perfectly correct classification. Results of the ROC-AUC test are shows in Figure 4.3.

The ROC curve offers additional validity to the classifier accuracies achieved by our models in earlier tests. Specifically, the ROC curves for each method skew toward the top left corner of the plot, and the AUC scores also gradually increase in the same fashion, as the classifier increases in (i) complexity and (ii) accuracy, as reported in Section 4.4.1. The Multimodal FinBERT DNN achieves the highest AUC score (0.89). This is followed by the Multimodal FinBERT NN (0.82), although this performance reflects only a marginal improvement on the single-modality FinBERT NN (0.81). Furthermore, the single-modality audio classifier is characterised by an ROC curve that lies very close to the pure chance line, and an AUC scores (0.56) only marginally higher than the chance level (0.50). Combined, the results suggest that audio characteristics in isolation only offer an incrementally increased ability to accurately predict sentiment.

Figure 4.1: Receiver Operating Characteristic (ROC) curves



Notes: This figure plots macro-averaged ROC curves for each of the models used within this study apart from the dictionary-based methods. It identifies the trade-off between True positive rates and False positive rates for each model. It also displays the AUC score for each model which is an effective summary of model accuracy. The closer an AUC score is to 1 the more robust the model is at making correct classifications.

4.4.3. Paralinguistic Feature Importance

The results reported in Sections 4.4.1 and 4.4.2 strongly suggest that incorporation of the audio modality to sentiment analysis models can slightly enhance accuracy. However, it may be the case that this enhanced performance is being driven by a small number of informative paralinguistic features, in which case, those audio cues that are barely informative may be removed from more streamlined dual-modality models in future. For this reason, permutation importance analysis is employed to assess the extent to which each of the specific paralinguistic features established in Table 4.2 inform the multimodal FinBERT DNN model, with results presented in Table 4.6.[215]

Table 4.6: Feature Importance Weights for Paralinguistic Features

Audio Feature	Weight
Fraction of Unvoiced	0.0811 ± 0.0168
Shimmer Local	0.0622 ± 0.0152
Mean Pitch	0.0443 ± 0.0113
Mean NHR	0.0442 ± 0.0126
Mean Autocorrelation	0.0406 ± 0.0118
Number of Periods	0.0368 ± 0.0158
Audio Length	0.0337 ± 0.0185
Number of Voice Breaks	0.0230 ± 0.0071
Jitter Local	0.0229 ± 0.0144
Mean Intensity	0.0126 ± 0.0052

Notes: This table represents the feature importance of each paralinguistic feature within the multimodal DNN model. The weight column represents the weight of each feature in relation to the other paralinguistic features in the dataset. The number to the left of the ± is the mean weight estimate with the number to the right of the symbol being the standard deviation of the estimate.

Permutation importance evaluates the importance of each feature in a classification model by measuring the impact of each feature on model accuracy, when specific features are randomly shuffled. If the model accuracy decreases with the shuffled feature data, the feature is considered to be important as the model no longer carries the same level of information. The results suggest that that all audio variables are, to differing degrees, somewhat important to the multimodal DNN model, which is somewhat expected given that the paralinguistic features were selected from a larger list of features, based on their prevalent use in prior studies outside of the finance domain.[216]

However, it is evident from the results that some features are more important than others. Namely, the fraction of unvoiced feature is found to be almost twice as informative as the third most informative feature (mean pitch). Shimmer local and the noise-to-harmonics ratio are also shown to be particularly important. The findings are generally consistent with prior literature suggesting that the fraction of unvoiced (Morrison, Wang and De Silva, 2007), shimmer (Li et al. 2007; Jacob, 2016) and pitch (Koolagudi and Rao, 2010; Koolagudi, and Krothapalli, 2012; Chebbi and Jebara, 2018) variables are

[215] We calculate permutation importance using Python's *eli5* library.

[216] The method used to select the final set of features is provided in Chapter 4.2.2.

important features for sentiment/emotion classification. Interestingly, mean intensity is found to be the least informative feature, though it does still hold a very small degree of influence on the predictions made by the multimodal FinBERT DNN classifier.

4.6. Conclusion

Textual analysis methods, namely sentiment analysis, have become increasingly popular within the academic finance literature in recent years. The techniques used to determine sentiment in each case vary considerably, with the most popular approaches – such as dictionaries and Naïve Bayesian classifiers – being comparatively more rudimentary than recent advancements, such as transformer architecture. This chapter offers two key contributions. Firstly, building on the comparison of disclosure sentiment methods by Frankel et al. (2002), this chapter creates a contemporary comparison of the most used sentiment analysis methods in academic finance to define financial sentiment, comparing against techniques currently employed within other domains. To do so, the analysis uses a dataset of 2,106 audio-text aligned sentences extracted from corporate earnings calls relating to twenty constituents of the S&P 100 index. The results strongly suggest that more computationally advanced classification models possess a greater ability to accurately classify corporate earnings call content. Secondly, the results show that the addition of a second modality, through the incorporation of vocal characteristics (paralinguistic features), allows for greater classification accuracy than existing text-based models.

The findings affirm results from extant literature which highlight that computationally advanced approaches appear to be more robust at capturing financial sentiment than commonly used dictionary methods (Kearney and Liu, 2014; Guo, Shi and Tu, 2016; Renault, 2017; Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Alamoudi and Alghamdi, 2021). Furthermore, the findings accentuate the conclusions made by Mayew and Venkatalcham (2012), and the wider social psychology literature, by finding that non-verbal information is incrementally informative in the communication process, albeit the effect is small. Particularly, in a financial setting this chapter’s results show that paralinguistic cues originating from earnings calls convey incremental information about the underlying sentiment of a message.

Although the multimodal sentiment classification model returns the highest classification accuracy for this dataset, there is still scope for further enhancement, mainly in regard to the creation of the classified sentiment data. For example, though the classifier is domain-specific to finance, it is not industry- or firm-specific. The companies used for the purposes of this study operate in a number of different industries and settings that differ, both in terms of the terminology used and the complexity of business models. Incorporation of these factors, perhaps through the use of industry experts to manually classify the training set, and validate results, could potentially improve the contextual understanding of the sentiment classifiers used.

Furthermore, classifying the sentiment of messages using both textual and audio data, as opposed to textual data only, would be more beneficial than only classifying using text in this study, as it would perhaps magnify the paralinguistic feature's ability to increase classification accuracies. Finally, a larger sample of classified messages would be beneficial to assess how much more of an understanding the multimodal model would achieve (deep learning models performer exponentially better on larger datasets). Despite the noted limitations, the multimodal model is still shown to produce greater classification accuracies over the most commonly used methods employed in extant financial literature, and thus has created a foundation for future research in this area.

5. An Event Study of Multimodal Earnings Call Sentiment and Abnormal Returns

5.1 Introduction

Behavioural finance is a specialised area of study that identifies the impact that psychological influences have on market outcomes. Lopez Cabarcos et al. (2020) highlight that a popular area for research within behavioural finance is the relationship between market sentiment and market reactions. This area of research has become popular due to its ability to identify the influence financial sentiment, which is shaped by cognitive biases and emotional responses, has on market behaviour. Subsequently, there has been a substantial interest in leveraging financial sentiment analysis methodologies to extract insights for asset pricing. Outside the financial domain, the field of Natural Language Processing (NLP) has seen substantial advancement in recent years, particularly due to new methods being developed to push state-of-the-art results forward. Two techniques that have pushed the capabilities of NLP are transformer architecture (Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Alamoudi and Alghamdi, 2021) and multimodal analysis (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021) both of which have been shown to increase sentiment classification accuracies in various domains.

Chapter 2 of this thesis and El Haj et al. (2018) however identify that the field of accounting and finance falls behind that of NLP studies, focussing on non-financial issues, in the classification of sentiment using state-of-the-art methods.[217] Particularly, models used to define financial sentiment in extant literature have mainly deployed dictionary-based approaches in their attempts to understand market behaviours. These approaches have been shown in the previous chapters to substantially underperform more advanced methods. In the comparative analysis conducted in Chapter 4 of this thesis, a multimodal sentiment classifier leveraging transformer architecture and paralinguistic cues is found to achieve the greatest accuracy in classifying earnings conference call sentiment. Building upon these results, the multimodal sentiment classifier established in Chapter 3 is now applied to 4,860

[217] El-Haj et al. (2018) highlights a potential reason for the lack of application of advanced NLP financial sentiment analysis models is due to the lack of substantial domain relevant datasets required for training and testing.

earnings conference calls relating to constituent firms of the Standard and Poor's (S&P) 100 index between 2006 and 2021.[218]

By applying this more advanced classifier, which leverages both textual and paralinguistic information from earnings calls, to an index-wide data set, the results of this analysis provide new insights into how behavioural factors influence market behaviours. This approach strengthens the behavioural finance perspective by demonstrating how sentiment and psychological cues impact investor reactions and market dynamics, further advancing the ongoing debate in asset pricing. Particularly, allowing us to identify whether market participants take into consideration the way in which information is framed[219] on these calls or whether the underlying fundamental information conveyed is only considered.

The analysis of this research generates key findings in relation to the theoretical field of asset pricing and the technical field of financial sentiment analysis. Firstly, with respect to the theoretical findings, the multimodal sentiment classifier reveals a highly significant and positive relationship with short-term Cumulative Abnormal Returns (CARs) whilst returning a highly significant negative relationship with longer period CARs. The relationship between sentiment and CARs in the short-term indicates that market participants quickly react to the information disseminated on these calls and react initially in a manner that follows the sentiment captured by the multimodal classifier. Subsequently, over the longer period, abnormal returns exhibit a negative relationship with sentiment indicating a reversal in returns. These findings suggest that investors initially overreact to the information conveyed on earnings calls, moving prices away from fundamental values, which compels market participants to reevaluate their initial positions and attempt to correct sentiment driven mispricing. This association between multimodal sentiment and CARs falls in line with a behavioural theory in that prices do not move to a fundamental value and reside there but instead initially overreact and continue to fluctuate.

Given the incorporation of paralinguistic features in this study and the extensive research exploring the distinct behavioural implications specific vocal characteristics have on individual decision-making, the analysis is extended to examine the impact earnings calls exhibiting significant differences in vocal characteristics among managerial and analyst participants have on market behaviour. This Chapter continues its investigation in this direction as the results arising from the main analysis indicate that market participants do indeed take into consideration the way in which information on earnings calls is expressed. Furthermore, the psychology literature indicates that greater emphasis on specific paralinguistic traits increases perceptions of speaker confidence and subsequently

[218] All of which contain a full repository of sentence level paralinguistic features.

[219] Framing is defined as how information is portrayed. Framing bias arises from framing and refers to the tendency of individuals to come to different conclusions about the same information due to the way such information is presented. Traditional theory indicates that rational agents make decisions under uncertainty using Expected Utility Theory (EUT). EUT, among other things, assumes descriptive invariance which implies that no matter how information is presented the same choice problem should lead to the same decision (Kircheler, Maciejovsky and Weber, 2004). Therefore, if framing bias is identified it would contradict EUT and give credence towards a behavioural theory which assumes market participants are sub-rational.

enhances persuasiveness (Mehrabian and Williams, 1969; London, Meldman and Lanckton, 1970; Erickson et al. 1978; Edinger and Patterson, 1983). Therefore, further exploring market reactions to calls which exhibit significant deviations form pallid paralinguistic communication creates a stronger more persuasive framing of company performance and in turn creates heightened market reactions in CARs.

As a result of this supplementary analysis, it is found that divergence in paralinguistic traits between the two sets of call participants (managers and analysts) evokes heightened market reactions to the information conveyed on earnings conference calls. Specifically, the findings indicate that calls displaying significantly higher levels of analyst intensity in comparison to managerial intensity, and calls showing significantly higher managerial jitter compared to analyst jitter elicit greater market reactions in short-term CARs. This provides further evidence, alongside the main results, that market participants are more liable to make sub-rational decisions based on the way information is portrayed around the time of a call rather than based purely on the fundamental information provided.

The technical contribution this study makes in relation to financial sentiment analysis is due to the insight provided surrounding the capabilities of a multimodal sentiment classifier applied to the financial domain. The inclusion of paralinguistic data allows this model to take into consideration the psychological influences that occur during communication that have been studied previously in psychology literature, and that have been shown to impact upon decision making (Erickson, Lind, Johnson and O'Barr, 1978; Conley, Lind and O'Barr, 1978; Apple et al. 1979; Wallbott, 1982; Brooke and Ng, 1986; Bradac, Mulac and House, 1988; Chattopadhyay et al. 2003; Feinberg et al. 2005; Klofstad, Anderson and Peters, 2012; Martín-Santana et al. 2015; Wang et al. 2018; Chua et al. 2020; Song et al. 2020), alongside the commonly used textual information to capture a more comprehensive sentiment signal. In doing so this chapter enriches our understanding of market dynamics through a deeper evaluation of the interplay between sentiment, market behaviour and market outcomes within the context of earnings calls. Furthermore, this investigation identifies that the multimodal sentiment classifier provides superior forecasting capabilities in comparison to singular modality models used in similar studies (Doran et al. 2012; Mayew and Venkatachalam, 2012; McKay Price et al. 2012; Brockman, Li and McKay Price, 2015) shown through higher R-square coefficients. Using the multimodal approach considered within this chapter, the results return an R^2 value of 0.7207 and an adjusted R^2 of 0.7187, which indicates considerably higher model explanatory power of abnormal returns than the aforementioned studies.

The remainder of this chapter is as follows: the next section develops the specific hypotheses that will be tested within this chapter, informed by relevant literature; Section 5.3 discusses the dataset being used to test these hypothesises; Section 5.4 provides details surrounding the content analysis methods and empirical methods used within this research; Section 5.5 provides the descriptive, main and additional analyses conducted. Finally, Section 5.6 concludes.

5.2 Hypothesis Development

To create a clear direction for this research and construct focused hypothesises that provide precise testable statements, this chapter relies on literature that has previously analysed the relationship between financial sentiment and CARs alongside theory to generate robust research questions. A multitude of research has been conducted on the associations between abnormal returns and various sources of financial information, including financial disclosures (Henry, 2006:2008; Li, 2010; Loughran and McDonald, 2011; Davis and Tama-Sweet, 2012; Jegadeesh and Wu, 2012; Twedt and Rees, 2012; Jiang et al. 2019) news (Tetlock, Saar-Tsechansky and Macskassy, 2008; Grob-Klubmann and Hautsch, 2010; Garcia, 2013; Ferguson et al. 2015; Sun, Najand and Shen, 2016; Audrino and Tetereva, 2019) and social media (Antweiler and Frank, 2004; Bollen, Mao and Zeng, 2011; Mao and Bollen, 2011; Sprenger et al. 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Siganos, Vagenas-Nanos and Verwijmeren, 2017; Gu and Kurov, 2020). This analysis builds on earlier work of McKay Price et al. (2012) by assessing the relationship between earnings call sentiment and two differing measures of CARs. The first measure of returns relates to short-term CARs, which is a summation of abnormal returns spanning one day prior (t-1) to one day post (t+1) the earnings call event (t). The second longer period measure looks at CARs over a longer time window, spanning from two days post (t+2) to sixty days post (t+60) the earnings call event. A broad consensus exists among related literature that abnormal returns move in the same direction as sentiment. Specifically, higher levels of positive sentiment drive higher short-term CARs, with higher levels of negative sentiment having a negative influence on short-term CARs.

For example, Tetlock (2007), Tetlock, Saar-Tsechanksy and MacKassy (2008), Loughran and McDonald (2011), and Garcia (2012) identify that higher levels of pessimism[220] across various financial sources incite a significant reduction in short-term abnormal returns.[221] Similarly, a number of studies identify that higher levels of positive sentiment across various financial information sources are positively associated with an increase in short-term abnormal returns (Antweiler and Frank, 2004; Henry, 2008; Twedt and Rees, 2012; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Azar and Lo, 2016; Jiang et al. 2019). Ferguson et al. (2015) and Bannier et al. (2017) both test the relationship between sentiment and short-term abnormal returns in both directions (i.e., how abnormal returns react to both positive and negative sentiment). They show, in line with the previous literature, that an increase in positive (negative) sentiment translates into an increase (decrease) in short-term abnormal returns.

Literature focussing on sentiment conveyed within earnings conference calls is generally in agreement with the aforementioned studies, in relation to short-term CARs. For example, McKay Price

[220] Pessimism in each of these studies is considered a higher negative word count in each respective information source.

[221] Tetlock (2007) uses Wall Street Journals (WSJ) “Abreast of the Market” opinion piece, Tetlock, Saar-Tsechanksy and MacKassy (2008) leverage news stories pertaining to S&P500 constituents contained within the WSJ and Dow Jones News Stories (DJNS), Loughran and McDonald (2011) use 10-K reports and Garcia (2013) uses media articles from the New York Times’.

et al. (2012) find that earnings calls in the highest (lowest) quintile of sentiment category have a positive (negative) impact on short-term returns. Doran et al. (2012) find that this result holds when examining earnings calls affiliated with Real Estate Investment Trusts (REITs); as earnings call sentiment becomes more positive (negative), the initial abnormal return response is higher (lower). Mayew and Venkatachalam (2012) show that a positive (negative) measure of qualitive information (defined using vocal characteristics) incites a positive (negative) reaction in initial CARs.[222] Brockman, Li and McKay Price (2015) identify that high (low) levels of sentiment on earnings calls produce a positive (negative) market reaction in abnormal returns. We therefore summarise the following hypothesis for short-term cumulative abnormal returns accordingly:

H1: Multimodal sentiment has a significant positive relationship with short-term CARs.

Previous literature suggests that an examination of the relationship between financial sentiment and abnormal returns over longer time periods yields stronger associations than short-term time periods. Engelberg (2008); Demers and Vega (2008) and McKay Price et al. (2012) all highlight the increased explanatory power of qualitative information contained in earnings calls when forecasting longer period abnormal returns. In contrast to this finding, Doran et al. (2012) and Antweiler and Frank (2004) both find that their respective sentiment measures are insignificant in predicting CARs for longer period returns. The majority of prior literature however agrees that the relationship between financial sentiment and longer period returns is inverse in nature (Lemmon and Portniaguina, 2006; Tetlock, 2007; Schmeling, 2009; Ho and Hung, 2012; Bathia and Bredin, 2013; Corredor, Ferrer and Santamaria, 2013; Gao and Yang, 2017).

To the author's knowledge, a majority of studies that analyse the relationship between earnings call sentiment and abnormal returns do so using return data over short time horizons, and seldom assess abnormal return's reaction to sentiment over longer time periods. A lack of consensus exists amongst the limited studies that do employ longer time periods. Jiang et al. (2019) show that a high level of manager sentiment is related to low excess aggregate returns in the next month. In line with this result, Siganos, Vagenas-Nanos and Verwijmeren (2014) find that Facebook sentiment's relationship with longer period returns is negative, although the results are not found to be statistically significant. Conversely, McKay Price et al. (2012) and Mayew and Venkatachalam (2012) find sentiment and abnormal returns move in the same direction for longer period abnormal returns. However, in line with the bulk of prior sentiment literature, including those studies using different information sources to earnings conference calls, an inverse relationship between earnings call sentiment and longer period abnormal returns is expected. Therefore, the following hypothesis for longer period abnormal returns has been established.

H2: Multimodal sentiment has a significant negative relationship with longer period CARs.

[222] See Chapter 2.3 for a more in-depth commentary on this paper.

As a next stage the relationship between sentiment and share pricing is further investigated, where the sample is limited to those earnings calls featuring the highest divergence of paralinguistic traits between managers and analysts. There has been a substantial amount of psychology research exploring the impact of specific language traits on listener perceptions and decision making (Nisbett and Ross, 1980; Aune and Kikuchi, 1993; Andersen and Blackburn, 2004; Craig and Blankenship, 2011; Clementson, Pascual-Ferra and Beatty, 2016),[223] and on the influence of varying paralinguistic (vocal) traits in the decision making process (Erickson et al. 1978; Conley, O'Barr and Lind, 1978; Brooke and Ng, 1986; Bradac, Mulac and House, 1988; Chattopadhyay et al. 2003; Feinberg et al. 2005; Klofstad, Anderson and Peters, 2012; Martín-Santana et al. 2015; Wang et al. 2018; Chau et al. 2020; Song et al. 2020). Amongst the relevant literature, it is generally found that distinctive language and vocal traits (such as pitch, intensity, jitter and shimmer) from speakers induce stronger responses from listeners in relation to agreeing with, and being persuaded towards, speaker opinions.

The four paralinguistic traits used to identify significant differences in vocal attributes between managers and analysts are (i) pitch, (ii) intensity, (iii) jitter and (iv) shimmer, primarily due to the extensive research conducted on each trait’s implications.[224] For instance, higher levels of vocal pitch are considered to create perceptions of immaturity, nervousness, lower credibility and lower persuasiveness (Chattopadhyay et al. 2003; Feinberg et al. 2005; Klofstad, Anderson and Peters, 2012; Martín-Santana et al. 2015; Wang et al. 2018; Chau et al. 2020; Song et al. 2020). Impressions of credibility and trustworthiness are thought to be related to higher levels of vocal intensity due to its relationship with efficiency in articulating information (Mehrabian and Williams, 1969; London, Meldman and Lanckton, 1970; Miller et al. 1976; Erickson et al. 1978; Apple et al. 1979; Edinger and Patterson, 1983). Vocal jitter and shimmer are shown to diminish in stressful situations which in turn potentially creates perceptions of doubt and uncertainty (Mendoza and Carballo, 1998; Park et al. 2011; Giddens et al. 2013). In a financial setting Bochkay et al. (2020) find that abnormal returns are more strongly correlated to extreme language in comparison to moderate language. They further identify that analyst revisions are strongly associated with extreme language, particularly positive language. In a similar light, Mayew and Venkatachalam (2012) show that their results in relation to unexpected earnings are stronger and more significant in high-scrutiny scenarios.

Hence, building upon the psychological implications of extreme language and vocal characteristics, it is expected that calls displaying the highest levels of paralinguistic divergence will emphasise the information being conveyed on such calls. Consequently, leading market participants to lend greater credence to this information in the financial decision-making process, and hence create a

[223] See Chapter 1.4 and Chapter 4.2.2 for further insight into the impact specific language traits have on listener perceptions and decision making.

[224] For a detailed discussion on how particular language traits, for instance language intensity, extremity and vividness aid persuasion attempts, see Chapter 1.4 of this thesis. Furthermore, for an insight into how paralinguistic traits impact perceptions of speaker confidence and consequently listener persuasion, see the following section 5.5.2.

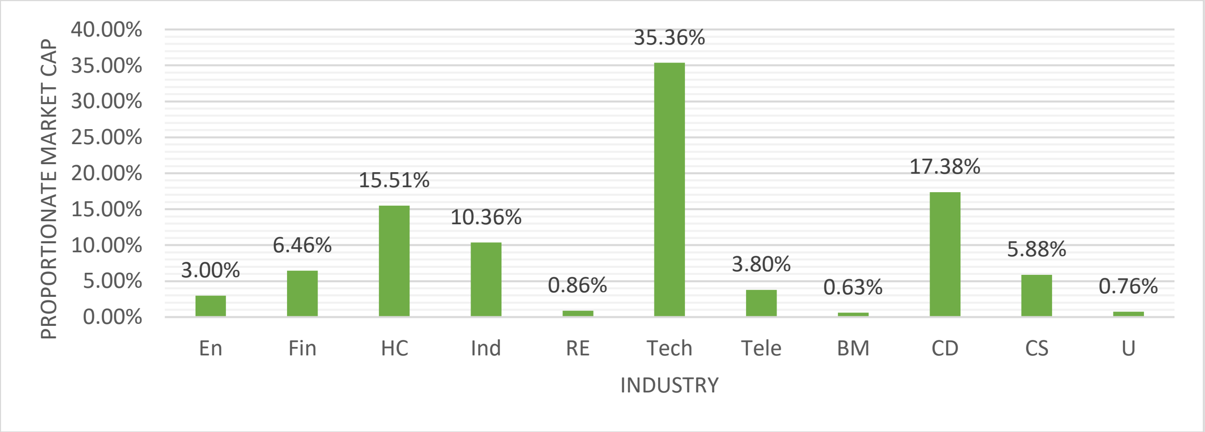
greater initial market reaction and a greater reversion of returns over a longer period. Therefore, the following hypothesises for the additional analysis in this research have been created:

H3: Multimodal sentiment from earnings calls defined by high levels of paralinguistic divergence has a significant positive (negative) relationship with short-term (longer period) CARs.

5.3 Data

For the purposes of this study, we use earnings call and share pricing data for 95 of the largest US-listed firms between 2005 and 2021. These 95 firms engaged in 4,860 earnings calls containing sentences, all of which contain a full repository of paralinguistic features.[225] Each firm was selected based upon a snapshot of S&P 100 constituents in 2021, and data for each firm was gathered through FinnHub[226] and Refinitiv.[227,228] Model training is based on a subset of earnings call sentences relating to twenty firms selected from the full dataset used, in accordance with the method used in Chapter 4. A full list of the companies used, including details of the market capitalisation, industry and sector of each constituent, the location of corporate headquarters, and the number of calls associated with each firm over the period, can be seen in Appendix 3.1. A further breakdown of this dataset by industry and market capitalisation is provided in Figure 5.1.

Figure 5.1: S&P 100 Firms' Market Capitalisation by Industry



Notes: This figure provides the market capitalisation of the overall sample set relating to the S&P100 discussed in the Methods Chapter (Chapter 3.2.3). The market capitalisation percentages in green represent each industries overall market share in the S&P100 index. Industry names: EN = Energy, Fin = Financials, HC = Health Care, Ind = Industrials, RE = Real Estate, Tech = Technology, Tele = Telecommunications, BM = Basic Materials, CD = Consumer Discretionary, CS = Consumer Staples, U = Utilities.

In comparison to extant literature examining earnings calls impact on financial markets, our final sample of earnings calls is of a roughly similar size. However, when compared to the small number of studies that focus specifically on paralinguistic characteristics of earnings conference calls, the current

[225] For a description of how paralinguistic features were calculated see Chapter 3.5
[226] FinnHub was used to download earnings conference call transcripts and corresponding audio.
[227] Refinitiv was used to download all data used surrounding company performance.
[228] For a full insight into the process of downloading and cleaning data see Chapter 3.

study analyses a substantially larger dataset of 4,860 calls: Mayew and Venkatachalam (2012) complete an analysis on 466 calls, while Chen, Han and Zhou (2023) use 848 calls and Li et al. (2020) use 3,443 calls.[229] As such, our more comprehensive dataset has the potential to offer additional insights into the effects of textual and paralinguistic cues, conveyed during earnings calls, on share pricing. A breakdown of the 4,860 calls across each year of the 16-year period can be seen in Table 5.1.

Table 5.1: Summary of Calls Per Year

Year	N. Sentences	MD	Q&A	Manager	Analyst	Sentiment	% change	Audio Length
2005	136	100%	0%	100%	0%	1.11	0%	35.74
2006	128	16%	84%	74%	26%	1.09	-2%	31.94
2007	125	34%	66%	83%	17%	1.33	22%	31.83
2008	125	23%	77%	78%	22%	1.26	-6%	34.23
2009	126	7%	93%	68%	32%	1.18	-6%	33.84
2010	123	10%	90%	66%	34%	1.21	2%	31.86
2011	124	2%	98%	58%	42%	1.26	4%	29.73
2012	121	0%	100%	50%	50%	1.3	3%	27.73
2013	127	1%	99%	55%	45%	1.31	0%	32.1
2014	137	5%	95%	57%	43%	1.24	-5%	38.65
2015	139	2%	98%	53%	47%	1.27	3%	38.38
2016	129	1%	99%	49%	51%	1.3	3%	33.12
2017	125	1%	99%	47%	53%	1.33	2%	31.13
2018	134	1%	99%	48%	52%	1.28	-4%	35.94
2019	144	0%	100%	43%	57%	1.24	-3%	43.8
2020	146	1%	99%	40%	60%	1.24	0%	47.04
2021	148	2%	98%	38%	62%	1.08	-13%	47.33

Notes: This table reports the average number of sentences associated to each section of the call, each participant set on the call, the average level of sentiment, the average cumulative abnormal returns over both the initial and longer periods being evaluated and the average audio length (in minutes) associated with the 4,860 calls across each specific year in this sample.

The first column of Table 5.1 indicates that the average number of sentences across all years remains reasonably consistent, with all calls on average containing between 120 to 150 sentences. Similarly, the average audio length of sentences used from each call within this analysis stays consistent across the years, suggesting that the length of earnings calls has not grown to any considerable degree in recent years.[230]

Evaluating the average number of sentences and average call audio lengths across the management discussion (MD) section and the question-and-answer (Q&A) section, the conversational Q&A section produces substantially more discussion (in terms of number of sentences) in comparison

[229] Li et al. (2020) take into consideration the initial management discussion section of earnings calls alone and do not use any paralinguistic information from the Q&A portion of these calls. Therefore, even though they have the closest number of calls to this study the number of sentences containing a full repository of paralinguistic features is significantly smaller – 394,277 sentences in comparison to the 637,220 used here.

[230] The average audio length used from each call indicates the length of audio associated with each sentence that returned a full repository of paralinguistic features. Not all sentences from all calls were returned successfully – see Chapter 3 section 3.5 for an insight in the full text audio alignment process and its accuracy rate.

to the MD section. Indeed, most years have upwards of 90% of sentences spoken on calls stemming from the Q&A section. Despite most sentences in this sample originating from the Q&A section, the split of sentences between managers and analysts is roughly equal with a slight skew towards managers. This roughly equal split provides this sample a substantial number of sentences from both sets of participants on earnings calls. Hence, giving this analysis the ability to examine the information produced by both sets of participants and how significant differences in participant communication can impact firm-level returns. The early years of the sample (pre-2011) contain on average double the number of manager sentences in comparison to analyst sentences. However, this divide moves towards a more equal split as the sample matures.

The sentiment columns of Table 5.1 shows that the average call sentiment over the course of the time frame remains constant. The percentage change column gives a clearer indication of the sentiment fluctuations over time. Further evaluating the sentiment variable, it identifies one year with a major increase in conference call sentiment followed by two substantial consecutive negative downturns. A 22% rise in the levels of sentiment is observed in 2007, with 2008 and 2009 both dropping 6% in their values of sentiment. These dates coincide with the Global Financial Crisis (GFC), where markets continued to rise over the course of 2007, before major indexes lost 20% of their value in 2008 and dropped further in value (54%) in 2009.

Table 5.2: Test of Differences of Means by Multimodal Sentiment Quartiles

Panel A - Differences of Means			
		CAR(-1,1)	CAR(2,60)
1 (High)	Mean	0.0118	-0.0072
	Std	0.0577	0.1286
	N. Observations	1211	1211
2	Mean	0.0082	-0.0063
	Std	0.0534	0.1316
	N. Observations	1212	1212
3	Mean	-0.0002	-0.0062
	Std	0.6068	0.1499
	N. Observations	1218	1218
4 (Low)	Mean	-0.0047	0.0101
	Std	0.0585	0.1854
	N. Observations	1216	1216
Panel B - T-tests			
Mean Q4 - Q1	T-statistic	-6.998	2.6778
	P-value	0.0000***	0.0075***

Notes: This table shows the differences in CARs when sorted into sentiment quartile portfolios using multimodal sentiment. Panel A provides the mean and standard deviation of each sentiment quartile in relation to both CAR measures. CAR (-1,1) is the short-term cumulative abnormal return across three days where the earnings call event is denoted as day 0 - abnormal returns are defined using the market model. CAR (2,60) is then the longer period *summation of cumulative abnormal returns from 2 days after an earnings call to 60 days after an earnings call. Panel B shows the differences in the highest and lowest sentiment groups and provides the test statistic and p-values for this comparison. Significance level indicators: * at 10%, ** at 5%, *** at 1%.*

In Panel A of Table 5.2, the 4,928 earnings calls sample is divided into quartiles based on the average sentence sentiment conveyed in the earnings call. Mean CARs and standard deviations are presented for each quartile, offering insights into potential associations between levels of sentiment expressed during earnings calls and subsequent market reactions. Mean levels of CAR(-1,1) begin negative in quartile 4 with a mean value of -0.0047, representing underperformance in the companies that reside within this quartile, and are generally found to increase through the higher sentiment categories until reaching an average positive response in CAR(-1,1) of 0.0118 in the most positive sentiment quartile. Therefore, implying that the market responds more positively, in terms of short-term abnormal returns, to earnings calls that exhibit higher levels of sentiment. The standard deviations relating to initial abnormal returns for each sentiment quartile remain consistent with no significant differences in the spread across the CAR(-1,1). These results indicate that the initial market response in abnormal returns has a positive relationship with earnings call sentiment i.e., as earnings call sentiment increases as too does the market reaction in abnormal returns.

The trend in mean levels of extended CAR(2,60) for each sentiment quartile return inverse results to that that found for short-term CARs. Specifically, the mean level of extended CARs is positive, 0.0101, for the lowest quartile of sentiment and becomes increasingly negative for the following three quartiles finally residing at -0.0072. The standard deviation of extended period returns also decreases as sentiment increases. Therefore, it can be said that as earnings conference calls increase in sentiment, the market responds with increased negativity over a longer horizon in abnormal returns. Furthermore, as earnings calls increase in sentiment the negative reaction in returns over the longer horizon is less varied as evidenced by a decreasing standard deviation. This potentially implies that positive calls evoke a larger initial market reaction in abnormal returns, but that market participants reach greater consensus surrounding the return reversal over the longer term.

5.4 Content Analysis and Empirical Method

5.4.1 Empirical Method

To examine the relationship between earnings call sentiment and abnormal returns this chapter builds upon previous literature that investigates similar associations. Following Tetlock (2007), Davis et al. (2008), Tetlock et al. (2008), Engelberg (2008), Frankel et al. (2010) and Mckay Price et al. (2012), who control for both the disclosure of additional information and other factors that are known to affect returns, the firm level effect of earnings call sentiment on security pricing is examined using cross-sectional regressions in the following form:

$$CAR_j = \propto_0 + \propto_1 SENTIMENT_{i,j} + \propto_2 SURP_{i,j} + CONTROLS_{i,j} + \in_{i,j} \text{ [5.01]}$$

In the above equation 5.01, CAR$_j$ refers to cumulative abnormal returns for a conference call j. CARs are defined using the following method:

$$CAR(-1,1) = \sum_{-1}^{1} AR_{i,j}$$

$$CAR(2,60) = \sum_{2}^{60} AR_{i,j}$$

To calculate abnormal returns, this analysis leverages the market model in line with previous research (Bowden, 2019; Doran et al. 2012).[231] These abnormal returns, as shown above, are then cumulated over two time periods to isolate the initial reaction and reaction over a longer horizon to each earnings call. The first CAR measure is a cumulation of abnormal returns over the three days (t-1 to t+1), beginning one day before the earnings call date (t-1) to one day after the call (t+1). This is consistent with prior studies using the same three-day window to investigate initial abnormal return reactions (Tetlock, 2007; Davis et al. 2008; Tetlock et al. 2008; Engelberg, 2008; Henry, 2008; Frankel et al. 2010; Mckay Price et al. 2012). Evaluating the lasting implications of sentiment, the second measure of CARs (t+2 to t+60) cumulates abnormal returns from two days after an earnings call (t+2) to sixty days after an earnings call (t+60), allowing for this investigation to capture any corrections in returns due to any potential over- or under-reaction to initial information.[232]

SENTIMENT$_{\mathrm{i,j}}$ in equation 5.01 represents the multimodal sentiment variable calculated for firm *j* at time *i*. The following section 5.4.2 provides further details on how the various sentiment variables used in this analysis have been calculated for each call. SURP$_{\mathrm{i,j}}$ expresses the unexpected earnings (also known as earnings surprise) for firm *j* at time *i*. The unexpected earnings variable is calculated, following the works of (Henry, 2008; Sadique, 2008; Akbas et al. 2013) in previous research, using a seasonal random walk approach as follows:

$$SURP_{i,j} = \frac{EPS_{i,j} - EPS_{i-4,j}}{PRICE_{i-4,j}}$$

Where:

$SURP_{i,j}$= The unexpected earnings for firm *j* at time *i*.
$EPS_{i,j}$= The earnings per share for firm *j* in quarter *i*.
$EPS_{i-4,j}$= The earnings per share in the same quarter of the previous year for firm *j*.
$PRICE_{i-4,j}$= The closing stock price of firm *j* in the same quarter of the previous year.

Ayers, Li and Yeung (2011) highlight that a number of prior studies use a seasonal random-walk approach to calculate unexpected earnings, however recent studies have used an alternative approach: analyst-based earnings surprises. As most prior research opts for the seasonal random-walk approach, this analysis adopts said method in calculating unexpected earnings. Further, Sadique (2008) uses both measures and finds no significant difference between the two.

[231] For a deeper explanation surrounding the reasoning and the method of calculating abnormal returns see Chapter 3.7.1.

[232] In the proceeding analysis the logarithm of CARs is used to create a more symmetrical comparison of short- and longer-term abnormal returns in the results.

The last variable to be discussed in equation 5.01 is in relation to the CONTROLS$_j$ variable. We use seven variables to control for external factors known to impact abnormal returns. Thus, the inclusion of these variables allows this analysis to focus directly on the impact multimodal sentiment has on CARs. The seven variables include measures of call length, firm size, book-to-market equity, profitability, leverage, trading volume and returns volatility. Call length is an indicator of the number of sentences within each earnings call. This variable is split into two columns which relate to the number of sentences associated with both portions of the earnings call, namely MD and Q&A sections. Firm size is calculated as the logarithm of the market capitalisation of a specific firm at the end of the quarter prior to the earnings call. Book-to-market equity is defined as the reciprocal of the market-to-book equity value directly downloaded from Refinitiv. Profitability (also known as return on assets, ROA) is calculated as the ratio between a given firms net income against total assets, expressed in percentage terms. Leverage is calculated as the ratio of total liabilities to total assets scaled by one hundred. Trading volume represents the logarithm of the volume of trades on the day of a given earnings call. Finally, returns volatility is calculated as the standard deviation of a given firms daily returns across the 90-day period beginning 100 days before a given call to 10 days before the call date.

This analysis applied the multimodal model to 4,860 conference calls discussed in the previous section. The sentiment model's classifications of the full sample of sentences have been aggregated into call level sentiment indicators using four differing methods identified in the following subsection 5.4.2.

5.4.2 Sentiment Variables

To create a robust analysis of the multimodal model's ability to capture financial sentiment and multimodal sentiment's relationship with abnormal returns, this chapter uses four differing methods to aggregate the sentence level classifications derived from the multimodal models earnings call sentence level classifications creating four call level indicators. As evidenced by Antweiler and Frank (2004), there are multiple ways in which to aggregate coded messages.[233] The authors assess the empirical performance of their naïve bayes classifier by employing three differing ways to aggregate messages. Identifying that the empirical performance of each measure they utilise is *"generally quite similar"* but not identical. Therefore, as this study looks to create a robust analysis of the multimodal models ability to classify financial sentiment, the following analysis uses four differing formulas to aggregate sentence level sentiment into a call level sentiment indicator.

The four formulas used to aggregate messages include the three measures used by Antweiler and Frank (2004) and a basic summation of sentiment. The first measure, the basic sentiment (BS) measure, can be calculated as follows:

$$BS = \frac{S^{Positive} - S^{Negative}}{N.\ of\ Sentences} \quad [5.01]$$

Where:

[233] Implying for this study, where each sentence within a call is classified as positive (1), negative (-1), or neutral (0), that each call can have a different level of sentiment depending upon how the sentence level sentiment is aggregated.

$S^{Positive}$= The number of positive sentences on a call.
$S^{Negative}$= The number of negative sentences on a call.
N. of Sentences = The total number of sentences on a call i.e., all positive, negative and neutral sentences.

This measure sums together all positive and negative sentences used within a call and divides by the total number of sentences in a call (positive + negative + neutral). The BS measure provides a call level sentiment figure that is bound between -1 and 1 i.e., a call with a sentiment of 1 contains only positive sentiment with a call sentiment value of -1 indicating that the call only contains negative sentiment. The following measures used to define sentiment in equations 5.02, 5.03 and 5.04 all originate from Antweiler and Frank (2004). Specifically, the first measure, defined in the following tables as AF1, is calculated in a similar fashion to the above basic sentiment measure and is also bound between -1 and 1, with a key difference in that the denominator only sums together the positive and negative messages.

$$AF1 = \frac{S^{Positive} - S^{Negative}}{S^{Positive} + S^{Negative}} \text{ [5.02]}$$

Antweiler and Frank (2004) highlight that this measure can be used to obtain all key results from their paper, however express a preference for the following measure:[234]

$$AF2 = \log\left(\frac{1 + S^{Positive}}{1 + S^{Negative}}\right) \text{ [5.03]}$$

The authors note that the third and final measure created is similar to AF2, and that both AF2 and AF3 differ from the initial AF1 measure as both measures increase in magnitude: (i) as the number of sentences being aggregated increase, and (ii) as the ratio of positive to negative messages increase. However, the first measure and the basic measure are homogenous with degree zero and hence independent of the number of messages considered.

$$AF3 = S^{Positive} - S^{Negative} \text{ [5.04]}$$

Each measure above has been used to calculate four different call sentiment figures in relation to the 4,860 earnings calls considered within this analysis.

5.5 Results

The descriptive analysis in the following subsections (5.5.1 and 5.5.2) is conducted at the sentence level. Sentence level analysis in this case refers to the inspection of all individual sentences that occur within each earnings call within this sample. These 4,860 calls translate into 637,220 sentences, and hence the sentence level analysis looks at the differences in these 637,220 sentences across the different sections and participants of earnings conference calls. Following the descriptive analysis, the main results are contained within subsection 5.5.3 and the results surrounding a

[234] Antweiler and Frank (2004) indicate their preference for this measure arises as it takes into consideration the number of traders expressing a particular sentiment. Furthermore, they note their preference for a homogenous measure of degree between zero and one.

supplementary analysis into the impact of calls exhibiting significant deviations in paralinguistic features, between both groups of call participants, is discussed in subsection 5.5.4.

5.5.1 Descriptive Statistics: Textual Data

Table 5.3 provides a comparison between sentences contained within the MD section of an earnings call versus the Q&A section (Panel A). A comparison is also provided between sentences spoken by managers versus sentences spoken by analysts (Panel B). A final comparison is conducted between the positive, neutral and negative sentiment categories (Panel C). Combined, these results allow for a greater general understanding of earnings call sentence characteristics. Panel A of Table 5.3 shows that most sentences in this sample stem from the Q&A portion (94%) of the call, with the remaining sentences stemming from the pre-scripted managerial introduction section (6%). As the Q&A section of the call requires participation from managers and analysts, this skew towards the Q&A section of the call does not lead to one set of participants dominating with a large majority of sentences. As expected, there are more questions asked during the Q&A section (71,306) in comparison to the MD section (4,130). The questions posed in the MD section arise due to rhetorical questions asked by managers in their initial presentation, questions asked by managers to other managers (often both the CEO and CFO are present on earnings calls), and questions asked to operators to amend call issues.[235]

Table 5.3: Descriptive Statistics of Classified Messages Textual Content

	Sentences		Length				Questions
	Total	%					
Section	N.	of	Min	Max	Mean	Median	N.
Panel A: Management Discussion vs Question and Answer							
MD	35,703	6%	1	254	19.566	16	4,130
Q&A	601,517	94%	1	273	19.172	16	71,036
Panel B: Manager vs Analyst							
Managers	353,131	55%	1	273	19.626	16	39,335
Analysts	284,089	45%	1	232	18.658	15	35,831
Panel C: Negative vs Neutral vs Positive							
Negative	117,439	18%	1	254	26.724	23	19,122
Neutral	329,599	52%	1	273	12.916	10	42,658
Positive	190,222	30%	1	232	25.424	22	13,386

Notes: This table provides a breakdown of the sentences contained within the different sections of earnings calls, across different participants and for the three different categories of sentiment. It identifies the total number (N.) of sentences, percentage (%) of sentences, minimum, maximum, mean and median sentence length and the number of questions posed.

Interestingly, even though there is a skew towards sentences originating from the Q&A section, Panel B identifies that there is a roughly equal split of sentences spoken by managers (353,131) and analysts (284,089) in our sample. The average length of sentence again is very similar for both

[235] An example of a question posed to a call operator is “And if we could have the answers on the screen?” and a question that is rhetorical in nature “So then the question is what do you do about it?”.

participants on the call, however the maximum sentence length for managers (273) is some forty words greater than that of analysts (232). This perhaps is due to managers need to effectively communicate financial performance in detail on these calls.

Panel C provides the number of sentences classified by the multimodal model as positive, negative or neutral across the full sample, giving an insight into the differences in these sentiment categories at the sentence level. The breakdown of positive (30%), negative (18%) and neutral (52%) sentences is broadly in line with prior literature surrounding sentiment distributions.[236] Positive and negative sentences on average are communicated with the same number of words, however neutral sentences are conveyed in roughly half as many words within earnings calls. Furthermore, as expected, most questions are posed on these calls in a neutral tone. Interestingly, questions are also asked with underlying positive and negative sentiment.[237]

5.5.2 Descriptive Statistics: Paralinguistic Data

The following descriptive statistics give insight into the differing average levels of paralinguistic attributes associated with both sets of participants (Panels A and B), both sections of the call (Panels C and D) and all three sentiment categories (Panels E, F and G). Table 5.4 shows that managers exert higher levels of pitch on average in comparison to analysts: the mean pitch of sentences spoken by managers on conference calls (167.613) is higher than that of analysts' (165.772). This result is also seen during the MD sections of these calls (174.078) in comparison to that of the Q&A sections (166.406). The manager subset being lower in comparison to the MD subset suggests that, during the Q&A section, managers lower their pitch when responding to analyst inquiries. These findings fall in line with Mayew, Parsons and Venkatachalam (2013), who find that managers lower their pitch in response to analysts rather than strategically lowering their pitch in the rehearsed MD section of the call.[238] Prior literature suggests that high levels of pitch create perceptions of immaturity, nervousness, lower credibility and persuasiveness (Chattopadhyay et al. 2003; Feinberg et al. 2005; Klofstad, Anderson and Peters, 2012; Martín-Santana et al. 2015; Wang et al. 2018; Chau et al. 2020; Song et al. 2020). Hence, a lower pitch in managerial replies may create perceptions of confidence and in turn could persuade market participants to align with a manager's opinion. This entails that managers', in response to analysts, are either confident in their replies to analysts or that they intentionally attempt to be perceived as confident.

[236] Sprenger et al. (2013) note that stock microblogs appear to be more balanced in terms of distribution of sentiment. The authors highlight that the majority of the messages they classified from stock microblogs were hold signals (49.6%) with buy signals (35.2%) being twice as likely as sell signals (15.2%). These findings are more in line with the breakdown of sentiment seen in this paper suggesting earnings calls have a more balanced distribution of sentiment that is similar to that of stock microblogs.

[237] An example of a question returning positive sentiment may be "How are you aiming to increase profitability over the next quarter?" and a negative sentiment "What causes the decline in share price in the previous quarter?"

[238] Mayew et al. (2013) note that it is still possible for managers to strategically manipulate their pitch in the MD section of the call.

Table 5.4: Paralinguistic Descriptive Statistics for each Call Section, Participant and Sentiment Category

Feature	Pitch	Intensity	N. of Periods	Fraction of Unvoiced	N. Voice Breaks	Jitter	Shimmer	Autocorrelation	NHR	Audio Length
Panel A: Management Discussion Section										
N.	35703	35703	35703	35703	35703	35703	35703	35703	35703	35703
Mean	174.078	28.962	835.070	52.146	38.249	2.816	17.436	0.752	0.391	17.192
SD	52.574	19.112	2971.168	24.331	117.839	1.099	2.710	0.074	0.146	44.352
Panel B: Question and Answer Section										
N.	601517	601517	601517	601517	601517	601517	601517	601517	601517	601517
Mean	166.406	28.947	802.910	50.848	36.683	2.595	16.497	0.765	0.368	16.133
SD	48.463	18.291	975.952	22.830	44.577	0.986	2.862	0.076	0.149	17.098
Panel C: Managers										
N.	368067	368067	368067	368067	368067	368067	368067	368067	368067	368067
Mean	167.613	29.341	809.783	50.712	36.780	2.628	16.685	0.764	6.280	16.151
SD	48.196	18.376	1309.167	22.989	55.940	0.991	2.838	0.075	2.339	21.280
Panel D: Analysts										
N.	269153	269153	269153	269153	269153	269153	269153	269153	269153	269153
Mean	165.772	28.410	797.776	51.206	36.760	2.579	16.365	0.764	6.314	16.250
SD	49.441	18.272	977.701	22.819	44.765	0.996	2.885	0.077	2.427	17.180
Panel E: Negative Sentiment										
N.	117439	117439	117439	117439	117439	117439	117439	117439	117439	117439
Mean	166.655	28.361	841.973	51.293	38.764	2.619	16.690	0.763	0.371	17.053
SD	48.208	17.955	1288.731	22.982	53.988	0.984	2.808	0.075	0.148	21.213
Panel F: Neutral Sentiment										
N.	329559	329559	329559	329559	329559	329559	329559	329559	329559	329559
Mean	166.558	29.377	777.916	50.573	35.555	2.602	16.492	0.765	0.368	15.642
SD	48.426	18.371	1210.623	22.748	53.128	0.993	2.883	0.075	0.147	19.996
Panel G: Positive Sentiment										
N.	190222	190222	190222	190222	190222	190222	190222	190222	190222	190222
Mean	167.429	28.568	828.131	51.292	37.648	2.611	16.564	0.763	0.372	16.616
SD	49.577	18.495	1049.674	23.164	46.850	1.001	2.855	0.077	0.151	17.948

Notes: This table provides basic descriptive statistics of paralinguistic features for each call section, call participant and each sentiment category. The descriptive metrics used are the number of sentences (N.) associated with each subsection, the arithmetic mean of each categories paralinguistic features and the standard deviation (SD) of each paralinguistic features distribution.

The results suggest that the Q&A section of the call, and particularly the analyst participants, are perceived as more reliable due to their lower average pitch on earnings calls, which in turn could heighten their ability to persuade listeners towards their opinions surrounding a firm's future direction. Furthermore, as managers speak in a higher pitch the market (audience) may be more risk averse to their suggestions. Chua et al. (2020), identify that during the communication process individual perceptions and judgements are impacted by speech cues. Particularly, the authors identify that low pitched speakers raise risk tolerance in listeners whilst high pitched speakers heightened risk aversion.

Assessing the average pitch associated with each of the three sentiment categories,[239] the positive (167.429), neutral (166.558) and negative (166.665) categories all display similar levels.

Regarding intonation,[240] the pre-scripted MD (52.574) section of the call has a higher standard deviation of pitch in comparison to that of the conversational Q&A (48.463) section. Managers however have a lower intonation (48.196) in comparison to analysts (49.441). The psychology literature relates lower (higher) levels of intonation with self-confidence and competence (lack of self-confidence). These results fall in line with the interpretations associated with the pitch variable, that managers are perceived with less confidence and competence in the initial scripted section of the call but are more confident and competent in conversational section of the call. Interestingly, the levels of intonation for neutral (48.426) and negative (48.208) categories are roughly similar with the positive category returning a higher average level of intonation (49.577). This linguistic analysis compliments extant financial literature which highlights that managers attempt to conceal bad news by using prepared scripts emulating positive language (Borochin et al. 2017). Furthermore, Mayew and Venkatachalam (2012) show that market participants react to both managerial positive and negative affective states but do so more prominently to negative states. These results potentially imply that positive comments are articulated with less confidence and create perceptions of dishonesty.

The MD section (28.962) and the manager subset (29.341) both exhibit similar levels of intensity in comparison to the Q&A section (28.947) and analyst subset (28.410). Intensity is found to be roughly equal for the positive (28.568) and negative (28.361) sentiment categories with the neutral (29.377) category carrying a higher average level of the measure. Higher levels of intensity suggest that a speaker has higher levels of efficiency in articulating arguments and subsequently creates impressions of credibility and trustworthiness (Erickson et al. 1978; Conley, Lind and O'Barr, 1978; Brooke and Ng, 1986; Bradac, Mulac and House, 1988).

Table 5.4 suggests that there isn’t much difference in the jitter or shimmer measures of each participant on the call, or each call subsection. However, a slight difference is noted in the jitter and shimmer measures across the MD and Q&A sections of the call. The MD section returns slightly higher levels of jitter and shimmer in comparison to the Q&A section. Prior literature suggests that measures of jitter and shimmer diminish with the implementation of experimentally induced stress (Mendoza and Carballo, 1998; Park et al. 2011; Giddens et al. 2013) indicating that the Q&A section of the earnings call contains speech that is spoken in a more stressful fashion. This aligns with Chen, Han and Zhou (2023) who highlight that the conversational section of an earnings call allows analysts to ask questions surrounding managers’ introductory statements, or firm performance, which can induce a high-stress environment.

[239] The sentiment categories are defined by splitting the 637,220 sentences into positive, negative and neutral subcategories based upon each sentence’s sentiment classification using the multimodal classifier.

[240] Gelinas-Chebat et al. (1996) state that intonation is the variation of pitch which reflects a voice’s melodic contour.

5.5.3 Main Results

Table 5.6 contains regression results for the multimodal model being evaluated in relation to short and longer period CARs. The results identify that each aggregation of multimodal sentiment returns significant results in relation to CARs, both in the initial period surrounding the call and the longer period. The results associated with short-term abnormal returns, all return positive coefficients falling in line with extant literature (Antweiler and Frank, 2004; Henry, 2008; Twedt and Rees, 2012; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Ferguson et al. 2015; Azar and Lo, 2016; Bannier et al. 2017; Jiang et al. 2019). Each of the aggregations of sentiment return highly significant results returning strong statistical significance at the 1% level. From these results hypothesis H1 can be accepted, confirming that multimodal sentiment has a significant positive relationship with short-term CARs. The coefficient of determination (R^2) and adjusted coefficient of determination (Adj R^2) for each aggregation of multimodal sentiment all surpass 0.71 for predicting short-term CARs, indicating the robust ability multimodal sentiment has in capturing initial market reactions. These findings imply that multimodal sentiment offers valuable insights into short-term market behaviours. This is observed across all four sentiment aggregation measures, suggesting a degree of robustness in the findings.

Comparing these results to previously conducted analyses of overall earnings call sentiment and short-term CAR, the results return similar statistical significance (Doran et al. 2012; Mayew and Venkatachalam, 2012; McKay Price et al. 2012). However, when considering the coefficient of determination, the results returned for the multimodal model far outperform that of previous research. Doran et al. (2012) consider both the Harvard-IV4 and Henry (2006) dictionary approaches to classify sentiment of earnings conference calls. From their results the Henry (2006) dictionary returns the highest R^2 for short-term CAR at 0.0069. Mayew and Venkatachalam (2012) return an adjusted R^2 of 0.0764 when calculating earnings call sentiment using a purely paralinguistic approach with managerial vocal cues. McKay Price et al. (2012) return an R^2 of 0.0016 when calculating overall call sentiment using the Henry (2006) specific dictionary approach. The results in the current study suggest an R^2 value of 0.7207 and an adjusted R^2 of 0.7187, which indicates considerably higher model explanatory power than the aforementioned studies.

Table 5.5: Estimation of the Association between Various Aggregations of Sentiment from each Classifier and CARs

Sentiment Measure	Basic Sentiment		AF Measure 1		AF Measure 2		AF Measure 3	
Cumulative Abnormal Returns	(-1,1)	(2,60)	(-1,1)	(2,60)	(-1,1)	(2,60)	(-1,1)	(2,60)
Constant	0.0059 (0.3587)	0.0577* (0.0641)	0.0058 (0.3645)	0.058* (0.0629)	0.0059 (0.3516)	0.0575* (0.0652)	0.0071 (0.2651)	0.0503 (0.1064)
Sentiment	***0.0109****** ***(0.0027)***	***-0.0642****** ***(0.0003)***	***0.0061****** ***(0.0012)***	***-0.0341****** ***(0.0002)***	***0.0022****** ***(0.0088)***	***-0.0142****** ***(0.0004)***	***0.0001****** ***(0.0016)***	***-0.0005****** ***(0.0002)***
Earnings Surprise	-0.0006 (0.9719)	-0.2173*** (0.0049)	0.0000 (0.9982)	-0.2207*** (0.0042)	-0.0003 (0.9868)	-0.2179*** (0.0048)	-0.0007 (0.9642)	-0.2165*** (0.0050)

Management Discussion	0.0000 (0.5591)	-0.0001 (0.6906)	0.0000 (0.5382)	-0.0001 (0.6580)	0.0000 (0.5220)	-0.0001 (0.6527)	0.0000 (0.8029)	0.0000 (0.9991)
Question and Answer	0.0000 (0.8251)	-0.0001 (0.4955)	0.0000 (0.8134)	-0.0001 (0.4812)	0.0000 (0.8007)	-0.0001 (0.4776)	0.0000 (0.8280)	0.0000 (0.8797)
Market Capitalisation	0.0000 (0.9714)	-0.0059** (0.0208)	0.0000 (0.9639)	-0.0059** (0.0208)	0.0000 (0.9841)	-0.0059** (0.0209)	0.0000 (0.9519)	-0.0058** (0.0224)
Book-to-Market	-0.0065*** (0.0000)	-0.0237*** (0.0012)	-0.0065*** (0.0000)	-0.0236*** (0.0013)	-0.0065*** (0.0000)	-0.024** (0.0011)	-0.0065*** (0.0000)	-0.0239** (0.0011)
Profitability	-0.0001 (0.3968)	-0.0003 (0.3026)	-0.0001 (0.4034)	-0.0003 (0.2973)	-0.0001 (0.4154)	-0.0003 (0.2802)	-0.0001 (0.3973)	-0.0003 (0.3021)
Leverage	-0.2741 (0.3176)	0.5927 (0.6580)	-0.2674 (0.3296)	0.567 (0.6720)	-0.2729 (0.3201)	0.5534 (0.6796)	-0.273 (0.3193)	0.5878 (0.6606)
Volume	0.0000 (0.1453)	0.0000 (0.1958)	0.0000 (0.1625)	0.0000 (0.1716)	0.0000 (0.1492)	0.0000 (0.1929)	0.0000 (0.1408)	0.0000 (0.2026)
Volatility	0.9378*** (0.0000)	0.0955** (0.0232)	0.9371*** (0.0000)	0.0985** (0.0194)	0.9379*** (0.0000)	0.097** (0.0213)	0.9377*** (0.0000)	0.0962** (0.0221)
N. Observations	4860	4860	4860	4860	4860	4860	4860	4860
R-sq	0.7206	0.0125	0.7207	0.0127	0.7205	0.0124	0.7206	0.0127
Adj R-sq	0.7186	0.0053	0.7187	0.0055	0.7184	0.0052	0.7186	0.0056
Industry Dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

*Notes: This table provides cross sectional regression results of CARs on measures of earnings call sentiment and control variable measures for calls that have been identified as having the largest average divergence in manager versus analyst paralinguistic features. CAR (-1,1) is the short-term cumulative abnormal return across three days where the earnings call event is denoted as day 0 - abnormal returns are defined using the market model. CAR (2,60) is then the longer period summation of cumulative abnormal returns from 2 days after an earnings call to 60 days after an earnings call. Each of the four sentiment indicators are aggregations of sentence level sentiment resulting from the application of the multimodal sentiment classifier, developed in Chapter 4, to the full sample of earnings call sentences. All four aggregation methods are discussed in Section 5.4.2. Industry and Time fixed effects are included to control for industry specific and year specific factors. N. Observations highlights the number of calls included within each regression. R^2 determines the proportion of variance in the dependent variable that can be explained by the independent variable. Adj R^2 is a corrected goodness of fit measure of R^2. Regression p-values are in parenthesis. Significance level indicators: * at 10%, ** at 5%, *** at 1%.*

In contrast to the initial period regressions, each aggregation of multimodal sentiment's relationship with longer period returns all return negative coefficients. These findings for longer period abnormal returns fall in line with previous literature who have evaluated the same question (Lemmon and Portniaguina, 2006; Tetlock, 2007; Schmeling, 2009; Ho and Hung, 2012; Gao and Yang, 2017; Bathia and Bredin, 2013; Corredor, Ferrer and Santamaria, 2013). Furthermore, each sentiment aggregation measure associated with multimodal sentiment returns a strong relationship with longer period CARs. Each of the measures return statistical significance at the 1% level. Combined, the results suggest that multimodal sentiment has an inverse relationship with longer period returns and confirms H2 can be accepted.

Although the relationship between multimodal sentiment and long-term CARs is highly significant, both the coefficient of determination (R^2) and adjusted coefficient of determination (Adj R^2) are consistently below 0.015 across each aggregation of sentiment. As the sentiment variable within this regression is shown to be highly significant but the coefficient of determination is low, it can be said that multimodal sentiment is strongly correlated to longer period CARs but holds limited ability in explaining variability of abnormal returns over a longer horizon. Comparing the forecasting capabilities of the multimodal model for abnormal returns at extended horizons, it is found that the model is slightly more robust when comparing to Doran et al. (2012) who return an R^2 of 0.0068. When considering the coefficient of determination returned within this analysis against that of McKay Price et al. (2012), who's study is directly compared to this analysis, it is found that again the multimodal model returns a greater R^2 value, even though both are relatively weak.[241]

The results shown in Table 5.5, when compared to previously completed studies in the same area, identify that the multimodal model captures a more robust sentiment variable than both singular modality models, agreeing with previous literature that highlights multimodal models outperform singular modality models due to the additional behavioural cues considered (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021).[242] Furthermore, these findings concur with previous financial literature that identify a positive relationship with short-term CARs and sentiment (Antweiler and Frank, 2004; Henry, 2008; Twedt and Rees, 2012; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Azar and Lo, 2016; Jiang et al. 2019; Seok , Cho and Ryu, 2019) and a negative association with long run CARs (Lemmon and Portniaguina, 2006; Tetlock, 2007; Schmeling, 2009; Ho and Hung, 2012; Bathia and Bredin, 2013; Corredor, Ferrer and Santamaria, 2013; Gao and Yang, 2017).[243] Previous literature explains these reactions stating that high sentiment predicts high short-term returns because of overpricing but the initial mispricing over the longer term reverts back toward a fundamental price, hence low future returns (De Long et al. 1990; Seok et al. 2019). These results, taken together with previous related literature from the computer science and financial domains, highlight that the multimodal model provides enhanced accuracy in capturing earnings conference call sentiment.[244]

From the theoretical perspective, the findings in relation to multimodal sentiment's relationship with CARs provides compelling insights into investor behaviour and market dynamics. The analysis reveals a highly significant positive relationship between multimodal sentiment and short-term CARs indicating that market participants quickly react to the information disseminated on earnings calls. The initial reaction aligns with the sentiment conveyed, suggesting that positive sentiment leads to positive abnormal returns and vice versa. Over a longer horizon, there is a highly significant negative relationship between sentiment and CARs. The initial positive relationship between sentiment and short-term CARs indicates that investors tend to overreact to the information presented in earnings calls. Due to the inclusion of paralinguistic cues within the multimodal sentiment classifier, this chapter posits that the classifier more accurately understands the way in which information is communicated on these calls (Guyer, Fabrigar and Vaughan-Johnston, 2018). Specifically, it is believed that this more accurate

[241] McKay Price et al. (2012) return a R^2 of 0.002 and statistical significance at the 5% level.

[242] This falls in line with Poria et al. (2015) who concludes that any combination of dual modality model is more robust at classifying sentiment than singular modality models.

[243] Brown and Cliff (2005) also find the same result for market wide sentiment.

[244] The most robust method of aggregating sentiment calculated from the multimodal model's classifications is the AF1 measure. All following analysis continues using this measure.

classifier identifies a framing bias within earnings call information which causes market participants to overreact to the intrinsic information disseminated leading prices to deviate from their fundamental values. The subsequent negative relationship over the longer term suggests that market participants reassess their positions, leading to a correction of the initial mispricing. These findings support behavioural theory, which postulates that prices do not immediately reflect fundamental values but instead exhibit patterns of overreaction and correction. Thus, suggesting inefficiencies in the market where prices are initially driven by the way fundamental information is portrayed rather than purely the fundamental information itself.[245]

5.5.4 Divergence of Paralinguistic Features

This section leverages the paralinguistic features contained within the multimodal model to assess economic agents' psychological perceptions associated with participants on earnings calls. Particularly, this subset of analysis takes the average levels of pitch, intensity, jitter and shimmer for each set of participants on these calls – managers and analysts – and identifies which calls have the highest divergence of paralinguistic traits.[246] To identify these calls the mean levels of each paralinguistic trait in relation to both groups of participants was created and then the mean and standard deviation of each calls' average level of paralinguistic traits were calculated. The calls with the highest levels of paralinguistic traits were then calculated as the calls that fell outside the range of the mean plus one standard deviation. Once these high divergence calls have been identified subsets of the overall earnings call dataset were created and their relationship with abnormal returns were evaluated. As this analysis looks at the divergence of paralinguistic traits and their implications on abnormal returns, earnings calls used within this subsection of analysis must contain both manager and analyst speech. This lowers the number of calls used to 4,395 and the results pertaining to this sample of calls are contained within Table 5.6.[247]

[245] Alternatively, this result may reflect the strategic use of sentiment by managers, where confident or optimistic tone is not grounded in fundamentals, leading to short-lived mispricing. This aligns with Mayew and Venkatalchalm (2012) suggesting that sentiment, particularly when conveyed through vocal cues, can distort market perceptions temporarily but is ultimately discounted as new information arrives.

[246] The analysis focuses on these paralinguistic traits as they have been identified, to the author's knowledge, as the most studied and understood features in psychological literature. Multiple studies have assessed the psychological implications of each of these traits and confirmed their implications of speaker perceptions. The paralinguistic features used within the multimodal model but not within this further analysis have been identified as improving multimodal sentiment classification but have not yet been as extensively studied in terms of their psychological implications.

[247] The text-audio alignment process discussed in Chapter 3.7 of this thesis identifies that the process is not 100% accurate. Due to the below perfect accuracy some sentences from this earnings call sample are missing and, in some cases, these missing cases leave a call with only sentences spoken from one group of participants.

Table 5.6: Estimation of the Association between Paralinguistic Features and CARs for Multimodal Sentiment

Paralinguistic Features	Pitch		Intensity		Jitter		Shimmer	
	High Manager	High Analyst	High Manager	High Analyst	High Manager	High Analyst	High Manager	High Analyst
Panel A: Initial Cumulative Abnormal Returns - CAR(-1,1)								
Const	0.0034 (0.8346)	0.0035 (0.8243)	0.0038 (0.7798)	-0.0061 (0.6806)	-0.0251 (0.1586)	-0.026 (0.1082)	-0.0327** (0.0487)	-0.0144 (0.382)
Sentiment	***0.0034 (0.5339)***	***0.0119** (0.019)***	***0.0013 (0.7789)***	***0.0122*** (0.0082)***	***0.017*** (0.0038)***	***-0.0056 (0.3191)***	***0.0125** (0.0236)***	***0.0025 (0.6486)***
Earnings Surprise	-0.0441 (0.34)	-0.0392 (0.4579)	0.0143 (0.7312)	0.0569 (0.1842)	-0.0044 (0.9413)	-0.0835 (0.1737)	-0.0122 (0.8142)	-0.0463 (0.4084)
Management Discussion	0.0000 (0.8225)	0.0001 (0.6821)	-0.0002 (0.6461)	0.0000 (0.7799)	0.0003 (0.2011)	0.0004 (0.1364)	0.0001 (0.7948)	0.0000 (0.8721)
Question and Answer	0.0000 (0.7881)	0.0001* (0.0765)	0.0000 (0.5042)	0.0000 (0.7855)	0.0000 (0.6621)	0.0001* (0.0883)	0.0001 (0.3829)	0.0000 (0.8495)
Market Capitalisation	-0.0013 (0.3273)	-0.0017 (0.2089)	-0.0014 (0.2377)	0.0007 (0.5961)	0.0015 (0.3215)	0.0015 (0.3066)	0.0016 (0.2371)	0.0011 (0.4478)
Book-to-Market	0.0041 (0.4215)	-0.0044 (0.3832)	0.0049 (0.2369)	-0.0091*** (0.0000)	0.0008 (0.8805)	-0.0067** (0.0167)	0.0008 (0.8721)	0.0098* (0.0624)
Profitability	0.0002 (0.4476)	-0.0002 (0.4357)	0.0000 (0.8396)	0.0000 (0.883)	-0.0002 (0.3364)	-0.0001 (0.5331)	-0.0001 (0.4909)	-0.0001 (0.598)
Leverage	0.9071 (0.1925)	-0.0937 (0.8862)	0.77 (0.213)	-0.3439 (0.561)	-0.1435 (0.8474)	0.1603 (0.8235)	0.4684 (0.4943)	0.3117 (0.6582)
Volume	0.0000 (0.7156)	0.0000*** (0.0007)	0.0000 (0.2573)	0.0000*** (0.0071)	0.0000 (0.4349)	0.0000 (0.9752)	0.0000 (0.155)	0.0000 (0.8326)
Volatility	0.9397*** (0.0000)	0.9311*** (0.0000)	0.9643*** (0.0000)	0.8974*** (0.0000)	0.869*** (0.0000)	0.9373*** (0.0000)	0.9117*** (0.0000)	0.9975*** (0.0000)
N. Observations	552	557	506	550	583	615	609	758
Panel B: Longer Period Cumulative Abnormal Returns - CAR(2,60)								
Const	0.0908 (0.259)	0.0672 (0.4493)	-0.1075 (0.1216)	-0.0858 (0.2351)	0.0909 (0.259)	0.0438 (0.6036)	0.1629** (0.0368)	-0.1453* (0.0577)
Sentiment	***0.0245 (0.3659)***	***0.0011 (0.9703)***	***-0.0137 (0.5572)***	***-0.0098 (0.6649)***	***-0.008 (0.7621)***	***-0.0589** (0.0441)***	***-0.0399 (0.1232)***	***-0.0056 (0.8263)***
Earnings Surprise	-0.3696 (0.1084)	-0.3284 (0.2763)	-0.1192 (0.5718)	-0.4384** (0.0369)	0.057 (0.8322)	-0.2578 (0.4213)	-0.337 (0.1689)	0.0735 (0.7767)
Management Discussion	-0.0008 (0.4076)	-0.004*** (0.0059)	-0.001 (0.675)	-0.0009 (0.2889)	-0.0005 (0.6126)	-0.0017 (0.2498)	0.0016 (0.1242)	-0.0014* (0.0871)
Question and Answer	0.0000 (0.8803)	-0.0004 (0.264)	0.0000 (0.9567)	0.0000 (0.9253)	0.0000 (0.8936)	0.0000 (0.9339)	-0.0003 (0.3891)	0.0002 (0.4123)
Market Capitalisation	-0.0084 (0.1949)	-0.0039 (0.6061)	0.0078 (0.1889)	0.0068 (0.2646)	-0.0084 (0.2295)	0.0013 (0.8657)	-0.009 (0.1593)	0.0113* (0.0798)
Book-to-Market	-0.0759*** (0.0028)	0.035 (0.2224)	0.0307 (0.1455)	-0.021** (0.0422)	-0.0302 (0.2151)	-0.0468*** (0.0015)	-0.0346 (0.1174)	-0.035 (0.1513)
Profitability	-0.0017* (0.0942)	0.0014 (0.208)	-0.0002 (0.792)	-0.0003 (0.7051)	-0.0013 (0.1826)	-0.001 (0.3551)	-0.0012 (0.2045)	-0.0009 (0.4104)
Leverage	5.4695 (0.1147)	0.7132 (0.8485)	1.392 (0.6556)	1.7984 (0.5345)	4.1904 (0.2151)	-3.4036 (0.3648)	-0.8615 (0.7892)	2.2949 (0.482)
Volume	0.0000*** (0.0335)	0.0000 (0.6793)	0.0000 (0.9852)	0.0000*** (0.0018)	0.0000 (0.254)	0.0000 (0.6387)	0.0000* (0.0943)	0.0000 (0.8847)
Volatility	0.1724 (0.2)	0.0351 (0.8173)	0.1917* (0.0722)	0.006 (0.9618)	0.0977 (0.4793)	0.1946 (0.1547)	0.1803 (0.1264)	0.2047* (0.0792)
N. Observations	552	557	506	550	583	615	609	758

Notes: This table provides cross sectional regression results of CARs on measures of earnings call sentiment and control variable measures for calls that have been identified as having the largest average divergence in manager versus analyst paralinguistic features. CAR (-1,1) is the short-term cumulative abnormal return across three days where the earnings call event is denoted as day 0 - abnormal returns are defined using the market model. CAR (2,60) is then the longer period summation of cumulative abnormal returns from 2 days after an earnings call to 60 days after an earnings call. Each of the *four sentiment indicators are aggregations of sentence level sentiment resulting from the application of the multimodal sentiment classifier, developed in Chapter 4, to the full sample of earnings call sentences. All four aggregation methods are discussed in Section 5.4.2. N. Observations highlights the number of calls included within each regression. Regression p-values are in parenthesis. Significance level indicators: * at 10%, ** at 5%, *** at 1%.*

In this divergence of paralinguistic analysis, the subset of calls which produce the most significant relationship with initial CARs is the set associated with high manager jitter in comparison to low analyst jitter. Wang et al. (2021) find that appeals of persuasion are more successful when persuaders vocal characteristics are less stressed. In this subset of data, the managers have on average higher levels of jitter and therefore are perceived as less stressed. The coefficient for this subset of data is statistically significant at the 1% level with a positive regression coefficient of 0.0170. In comparison to the results pertaining to the full sample of calls for short-term CARs, this result returns a higher positive coefficient.[248] This may suggest that managers are more confident in discussing past results and future performance of their firms in this specific subset of data and subsequently market participants lend greater credence to managerial sentiment on these calls in making financial decisions in the short term. This result magnifies the literatures findings discussed in the main results section that sentiment has a positive relationship with short-term CARs.

To further evaluate this conclusion, the average level of sentiment for managers and analysts on this subset of calls was calculated. Both participants return a similar level of sentiment,[249] with analysts returning a sentiment level of 0.252 and managers 0.253. Chen, Nagar and Schoenfeld (2018) find that on average the sentiment of earnings calls begins optimistic, due to managerial introductions, and moves towards a level of sentiment that fits with a firm's performance in the previous quarter. They further note that the level of sentiment produced by analysts begins closer to this *correct* level of sentiment. Hence, as managers are being perceived as calm and confident on these calls, coupled with the level of sentiment of managers and analyst being very similar, potentially implies that managers are not being overly positive about past results and are speaking at a level of sentiment which accurately reflects past performance. Looking at the longer period CARs regression for this set of calls it is found there is no significant result which would imply no reversal in returns. This agrees with the conclusion made previously which hypothesises that managers are speaking with a level of sentiment that accurately reflects firm performance and consequently market prices move to a fundamental level quickly with no further fluctuations.

Looking at the relationship with longer period returns for high divergence of jitter calls, the only significant result is in relation to the subset of calls that have high levels of analyst jitter and therefore low managerial jitter. Hence, managers are perhaps perceived as more stressful in this subset of calls in comparison to the analysts, who are perceived as being calmer in temperament. The regression

[248] The multimodal model using AF1 sentiment measure has a coefficient of 0.0061 and is significant at the 1% level.

[249] Average sentiment here and in the following discussion is calculated using the AF1 measure of sentiment on the multimodal model's sentiment classifications.

coefficient is negative, suggesting that abnormal returns decrease as sentiment increases for this subset of calls long-term. Again, this subset provides a stronger negative coefficient suggesting more of an inverse reaction to adjust security prices back to fundamental values.[250] This result is broadly in keeping with prior literature suggesting that long-run returns are negatively associated with earnings call sentiment. Furthermore, it suggests that there is an initial overreaction to call sentiment that is more aggressively reversed when managers are perceived as stressed on these calls. Managers in a heightened state of stress, as evidenced through low vocal jitter, could suggest that firm financial performance is lower than expected and hence the more prominent reversal to an initial positive reaction. The average levels of sentiment for this subset of calls shows that analysts (0.264) are less positive than managers (0.273). This could suggest an initial overreaction to managerial sentiment but a correction back toward analyst sentiment which has been defined as being more in line with fundamental performance.

The next significant result seen in Table 5.6 relates to calls that have high levels of analyst intensity and low levels of manager intensity. Reiterating the literature discussed in section 4.2.2 and extant psychological literature it is known that high levels of speaker intensity is related to credibility and trustworthiness. Priester and Petty (2003) find that spokesperson trustworthiness is influential for persuasion.[251] The coefficient in relation to the high analyst intensity regression is positive and significant at the 1% significance level suggesting that as sentiment becomes more positive for calls where analyst intensity is high, initial abnormal returns increase. This increase in initial abnormal returns is stronger for these calls which see high (low) levels of analyst (manager) intensity as evidenced through a stronger regression coefficient of 0.0122 in comparison to the overall set's coefficient associated with the multimodal model using AF1 of 0.0061. A potential reason for this stronger initial market reaction, following the implications laid out by Priester and Petty (2003), could be that markets perceive analysts as an expert source and trust their opinions surrounding the future directions of firm financial performance.[252] Therefore, if sentiment is a mechanism that portrays analysts' opinions about firm performance, and analysts are speaking in a way that is perceived as credible and trustworthy, market participants may unthinkingly accept the conclusion as valid (i.e. they accept their level of sentiment as a robust indicator of future performance). Average sentiment produced by analysts on high intensity calls (0.252) in comparison to the average sentiment produced by managers on high intensity calls (0.212) is more positive. Analysts speaking at a more positive level than managers, in a more credible and trustworthy fashion with no evidence of return reversals over the longer period CARs

[250] This result is significant at the 5% level which is not as strong as the 1% level seen in the full sample regression.

[251] There is a greater need to think about messages from expert sources who are dishonest over sources who are trustworthy. This is because if a message recipient can trust that an expert source will be providing accurate information, they do not need to complete the effortful task of scrutinizing and fact checking information. Hence, they can unthinkingly accept the information/ conclusion as valid.

[252]Chen, Nagar and Schoenfeld (2018) find that in comparison to management, analysts speak in a more neutral fashion and that their levels of sentiment lie closer to a level which fits with the firm's performance of the previous quarter. This potentially gives evidence that analysts sentiment is an expert source.

regression could suggest that the market in these instances take the analysts level of sentiment as a robust indicator of future performance.

The last subset of calls that contain notable levels of paralinguistic divergence is in relation to the shimmer characteristic. The high manager shimmer subset, implying low analyst shimmer, is statistically significant at the 5% level with a positive coefficient (0.0125) for the initial period regression. Hence, this finding implies that as sentiment increases on calls with high manager shimmer in comparison to low analyst shimmer, short-term abnormal returns increase. Low levels of shimmer are associated with the same psychological underpinnings as jitter – as stress increases shimmer diminishes. Therefore, this result confirms the conclusions drawn from the high manager jitter analysis, that managers are calmer and perceived as confident on these subsets of calls when discussing firm performance. Consequently, market participants lend greater credence to managerial sentiment on these calls in making financial decisions in the short term. Neither of these subsets return statistical significance for long-run abnormal returns, which in line with prior literature would be a reversal of returns, suggesting that managers on these calls are not attempting to be perceived as calm but in fact are calm and that they are discussing positive results.

To further explore the interaction between multimodal sentiment and CARs on calls that produce significant differences in paralinguistic traits, the conditional regression analysis conducted in Table 5.6 has been replicated using a dummy variable regression analysis shown below in Table 5.7.[253] This additional analysis produces similar results as those found in Table 5.6 for short-term CARs, but no results are found to be in agreement for the longer-term analysis. It is found that calls which produce higher levels of manager intensity and jitter return significant results at the 5% level for the short-term period. Furthermore, the results indicate that higher levels of analyst intensity, jitter and shimmer all return significant results with short-term CARs. Comparing the short-term results of Table 5.7 with Table 5.6 it is found that two subsets of calls show significant results in both tests, and both return positive coefficients. Calls exhibiting high levels of analyst intensity and calls showing high levels of manager jitter both have a positive impact on short-term abnormal returns.[254] Confirming the effect of high levels of analyst intensity through this additional analysis gives backing to the insights laid out above leaning on the psychological underpinnings of speaker intensity in that analysts are considered experts on these calls and their level of sentiment is used by market participants as a robust indicator of future firm performance. Moreover, the subset of calls with high manager jitter again returning significant results gives further evidence towards market participants reacting more positively to calls where managers speak in a calm and unstressed fashion.

[253] The conditional regression only applies the regression equation to calls that fall within the paralinguistic parameters defined. However, the dummy variable regression uses all calls and includes a dummy variable column to identify calls that show significant differences in paralinguistic traits.

[254] This is shown through the coefficient in front of the sentiment variable in Table 5.7 and through the coefficient of the Dummy X Sentiment variable in Table 5.8.

This analysis pertains to subsets of calls that produce evident differences between the two sets of participants within this sample. For each significant short-term result in Table 5.6 and all but two significant results in Table 5.7, we identify stronger coefficients and similar significance in comparison to the regressions that use the full sample of calls. This suggests that the multimodal model not only outperforms singular modality models but is also effective in picking up subtle differences in call participant paralinguistic traits shown through the inflated results found. Although each significant result found in Tables 5.6 and 5.7 provide insights into market participant behaviour resulting from earnings calls, the two short-term results found in both regression methods (high analyst intensity and high manager jitter) give evidence towards these paralinguistic features, defined through participant vocal characteristics, as having strong implications for short-term CARs. This analysis confirms, in a financial setting, the impact vocal cues have on speaker perceptions and judgement that have been found in general settings (Burgoon, Birk and Pfau, 1990; Friestand and Wright, 1994; Gaertig and Simmons, 2018; Guyer, Fabrigar and Vaughan-Johnston, 2018; Van Zant and Berger, 2020) and further identifies that the way in which information is framed has a significant impact on the decision-making process of financial market agents. These results lend support to the framing bias found in the main results of this chapter, and further highlight the influence market participant psychology has on market reactions in line with behavioural theory.

Table 5.7: Estimation of the Association between Paralinguistic Features and CARs for Multimodal Sentiment

Panel A: Initial Cumulative Abnormal Returns - CAR(-1,1)

Paralinguistic Features	Pitch		Intensity		Jitter		Shimmer	
	High Manager	High Analyst	High Manager	High Analyst	High Manager	High Analyst	High Manager	High Analyst
const	-0.0033 (0.5424)	-0.0029 (0.5952)	-0.0041 (0.4474)	-0.0028 (0.6025)	-0.0026 (0.6303)	-0.0037 (0.4954)	-0.0026 (0.6308)	-0.0035 (0.5119)
Sentiment	0.0073*** (0.0001)	0.0063*** (0.0008)	0.0083*** (0.0000)	0.0051*** (0.0075)	0.0056*** (0.0029)	0.0077*** (0.0000)	0.0059*** (0.0017)	0.0078*** (0.0000)
Dummy	0.0032 (0.1489)	-0.0023 (0.2784)	0.004** (0.0225)	-0.0044** (0.0109)	-0.0048** (0.0213)	0.0036* (0.0792)	-0.0027 (0.1612)	0.0046** (0.0183)
Interaction	-0.0088 (0.1387)	0.0029 (0.6085)	-0.0116** (0.0135)	0.0096** (0.0424)	0.0099* (0.0825*)	-0.0122** (0.0274)	0.0052 (0.3140)	-0.0112** (0.0290)
Earnings Surprise	0.0123 (0.4363)	0.0112 (0.4788)	0.0134 (0.3958)	0.0111 (0.4849)	0.0114 (0.4718)	0.0128 (0.4200)	0.0114 (0.4729)	0.0127 (0.4241)
Management Discussion	0.0000 (0.9556)	0.0000 (0.9238)	0.0000 (0.9170)	0.0000 (0.9775)	0.0000 (0.9498)	0.0000 (0.9406)	0.0000 (0.9605)	0.0000 (0.8700)
Question and Answer	0.0000 (0.6522)	0.0000 (0.6518)	0.0000 (0.6030)	0.0000 (0.6008)	0.0000 (0.6395)	0.0000 (0.5818)	0.0000 (0.6901)	0.0000 (0.6322)
Market Capitalisatio n	0.0002 (0.6311)	0.0002 (0.6127)	0.0002 (0.5950)	0.0002 (0.6209)	0.0002 (0.5995)	0.0002 (0.6084)	0.0002 (0.6125)	0.0002 (0.6379)
Book-to-Market	-0.006*** (0.0000)	-0.006*** (0.0000)	-0.006*** (0.0000)	-0.0058*** (0.0000)	-0.0061*** (0.0000)	-0.006*** (0.0000)	-0.006*** (0.0000)	-0.006*** (0.0000)
Profitability	-0.0001 (0.1395)	-0.0001 (0.1372)	-0.0001 (0.1427)	-0.0001 (0.1613)	-0.0001 (0.1289)	-0.0001 (0.1244)	-0.0001 (0.1304)	-0.0001 (0.1404)
Leverage	-0.0051 (0.9826)	-0.0129 (0.9558)	0.0003 (0.9990)	0.0056 (0.9807)	-0.0315 (0.8928)	-0.013 (0.9555)	-0.0148 (0.9493)	-0.0039 (0.9868)
Volume	0.0000 (0.7668)	0.0000 (0.7674)	0.0000 (0.7613)	0.0000 (0.7815)	0.0000 (0.8037)	0.0000 (0.8114)	0.0000 (0.7911)	0.0000 (0.8173)

Volatility	0.9374*** (0.0000)	0.9374*** (0.0000)	0.9375*** (0.0000)	0.938*** (0.0000)	0.9374*** (0.0000)	0.9381*** (0.0000)	0.9374*** (0.0000)	0.9383*** (0.0000)
N. Observations	4395	4395	4395	4395	4395	4395	4395	4395
Panel B: Longer Period Cumulative Abnormal Returns - CAR(2,60)								
const	0.0358 (0.1713)	0.0353 (0.1777)	0.0366 (0.1635)	0.0359 (0.1708)	0.0354 (0.1766)	0.035 (0.1821)	0.034 (0.1957)	0.0345 (0.1886)
Sentiment	-0.0323*** (0.0004)	-0.0277*** (0.0025)	-0.0278*** (0.0031)	-0.0304*** (0.0012)	-0.0289*** (0.0017)	-0.0252*** (0.0059)	-0.026*** (0.0051)	-0.0323*** (0.0005)
Dummy	-0.0209* (0.0573)	-0.0035 (0.7337)	-0.0096 (0.2585)	-0.0138 (0.0990)	-0.0036 (0.7245)	-0.0063 (0.5359)	0.0044 (0.6385)	-0.007 (0.4651)
Interaction	0.0621** (0.0332)	0.0064 (0.8182)	0.0055 (0.8105)	0.0226 (0.3268)	0.0183 (0.5122)	-0.0141 (0.6003)	-0.009 (0.7205)	0.0422* (0.0923)
Earnings Surprise	-0.1958** (0.0114)	-0.1941** (0.0122)	-0.195** (0.0118)	-0.196** (0.0114)	-0.194** (0.0122)	-0.193** (0.0127)	-0.1922** (0.0131)	-0.1953** (0.0116)
Management Discussion	-0.0005 (0.2775)	-0.0005 (0.3094)	-0.0005 (0.2871)	-0.0004 (0.3490)	-0.0005 (0.2953)	-0.0005 (0.3042)	-0.0005 (0.3029)	-0.0004 (0.3304)
Question and Answer	0.0000 (0.6379)	-0.0001 (0.6106)	0.0000 (0.6568)	0.0000 (0.6440)	0.0000 (0.6340)	-0.0001 (0.6237)	0.0000 (0.6342)	0.0000 (0.6377)
Market Capitalisation	-0.0013 (0.5660)	-0.0013 (0.5463)	-0.0014 (0.5239)	-0.0014 (0.5375)	-0.0013 (0.5399)	-0.0013 (0.5529)	-0.0013 (0.5461)	-0.0012 (0.5713)
Book-to-Market	-0.0207*** (0.0010)	-0.0207*** (0.0010)	-0.0206*** (0.0011)	-0.0203*** (0.0013)	-0.0208*** (0.0010)	-0.0206*** (0.0011)	-0.0206*** (0.0011)	-0.0204*** (0.0012)
Profitability	-0.0008** (0.0181)	-0.0008** (0.0197)	-0.0008** (0.0188)	-0.0008** (0.0230)	-0.0008** (0.0188)	-0.0008** (0.0196)	-0.0008** (0.0198)	-0.0008** (0.02)
Leverage	0.6364 (0.5768)	0.6807 (0.5507)	0.6537 (0.5666)	0.7588 (0.5064)	0.655 (0.5661)	0.7003 (0.5392)	0.6865 (0.5473)	0.6515 (0.5678)
Volume	0.0000* (0.0839)	0.0000* (0.0837)	0.0000* (0.0805)	0.0000* (0.0843)	0.0000* (0.0834)	0.0000* (0.0726)	0.0000* (0.0812)	0.0000* (0.0854)
Volatility	0.094** (0.0324)	0.0936** (0.0332)	0.0941** (0.0322)	0.0947** (0.0312)	0.0939** (0.0327)	0.0933** (0.0339)	0.0937** (0.0330)	0.092** (0.0363)
N. Observations	4395	4395	4395	4395	4395	4395	4395	4395

*Notes: This table provides cross sectional regression (containing a dummy and interaction variable) results of CARs on measures of earnings call sentiment and control variable measures for calls that have been identified as having the largest average divergence in manager versus analyst paralinguistic features. CAR (-1,1) is the short-term cumulative abnormal return across three days where the earnings call event is denoted as day 0 - abnormal returns are defined using the market model. CAR (2,60) is then the longer period summation of cumulative abnormal returns from 2 days after an earnings call to 60 days after an earnings call. Each of the four sentiment indicators are aggregations of sentence level sentiment resulting from the application of the multimodal sentiment classifier, developed in Chapter 4, to the full sample of earnings call sentences. All four aggregation methods are discussed in Section 6.4.2. N. Observations highlights the number of calls included within each regression. Regression p-values are in parenthesis. Significance level indicators: * at 10%, ** at 5%, *** at 1%.*

5.6 Conclusion

The bulk of work completed using qualitative information to forecast financial markets has been done so using textual information sources (see Chapter 2). Henry (2006) was one of the first papers to identify that the inclusion of qualitative information above the traditionally used quantitative data provides substantial gains in financial market forecasting accuracy. Mishew et al. (2020) highlight that sentiment analysis has become an important tool for analysing qualitative information in the financial domain due to its ability to summarise large amounts of financial text. This is particularly beneficial for finance as it allows a previously unused modality of information,[255] to be understood and used at scale. Hence, if a robust sentiment measure that accurately depicts qualitative information and the way in

[255] The efficient markets hypothesis states that all relevant information is reflected in share prices.

which such information is disseminated is created, then an advantage can be gained. Building on accounting and finance research (Das and Chen, 2007; Loughran and McDonald, 2011; Kearney and Liu, 2014; Guo, Shi and Tu, 2016; Renault, 2017; Hiew et al. 2019; McGurk, Nowak and Hall, 2020) that highlights a more accurate measure of financial sentiment returns stronger relationships with financial market metrics and in turn provides superior forecasting ability for users, and various studies (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021) that identify the use of paralinguistic information, alongside extensively used textual information, is beneficial in defining a more accurate measure of sentiment, this chapter leverages state-of-the-art natural language processing techniques to create a robust multimodal sentiment classifier which classifies financial sentiment from earnings conference calls by considering not only the textual information produced in these financial disclosures but also additionally utilises behavioural cues contained within the paralinguistic element of the calls.

This research represents a comprehensive investigation into the ability of a multimodal sentiment classifier in capturing financial sentiment by assessing its relationship with firm level abnormal returns. In doing so the main analysis of this research generates key theoretical and computational findings. Firstly, multimodal sentiment generates results in line with previous literature. Specifically, it is found that multimodal sentiment returns a positive relationship with short-term CARs and a negative relationship with longer-term CAR, both results returning statistical significance. This direction of relationship adds to the consensus found in previous research (Antweiler and Frank, 2004; Lemmon and Portniaguina, 2006; Tetlock, 2007; Henry, 2008; Tetlock, Saar-Tsechanksy and MacKassy, 2008; Schmeling, 2009; Loughran and McDonald, 2011; Doran et al. 2012; Ho and Hung, 2012; Mayew and Venkatachalam, 2012; McKay Price et al. 2012; Twedt and Rees, 2012; Bathia and Bredin, 2013; Corredor, Ferrer and Santamaria, 2013; Garcia, 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Brockman, Li and McKay Price, 2015; Ferguson et al. 2015; Azar and Lo, 2016; Bannier et al. 2017; Gao and Yang, 2017; Jiang et al. 2019). Furthermore, the results indicate that investors tend to initially overreact to the information presented during earnings calls, causing prices to deviate from fundamental values in the short-term. This initial overreaction perhaps is due to the way in which information is framed on these calls. As the sentiment classifier used in this analysis takes into consideration behaviour cues known to heighten persuasion and impact decision making and is shown in the previous chapter to have a better understanding of financial sentiment, it may be the case that this model is able to pick up subtleties in the way in which information is portrayed which identifies market participants use of paralinguistic information in their decision-making process. These initial decisions, then prompts market participants to reassess their initial sub-rational decisions and work to correct mispricing due to such framing bias, captured in the results with a reversal in CARs over the longer horizon. The relationship between multimodal sentiment and CARs underscores the complexity of market dynamics and the influence of investor psychology on decision-making, emphasising the need for a nuanced approach to interpreting sentiment and its impact on market returns.

Due to the inclusion of paralinguistic features in this analysis, and the extensive research done on the behavioural implication's distinct levels of specific paralinguistic features have on individuals, the analysis extended into looking at specific calls where managerial and analyst vocal characteristics had the greatest variance. As a result of this additional analysis, this chapter identifies that a divergence in certain paralinguistic traits, between managerial and analyst participants, creates increased market reactions to the information being portrayed on earnings conference calls. Specifically, the results show that calls that produce on average higher levels of divergence evoke a greater market reaction in short-term CAR in comparison to the main results. Multiple subsets of calls have been identified, through significant differences in participant paralinguistic features, as having a significant impact on abnormal returns. However, the strongest results (as found in both regressions displayed in Table 5.6 and 5.7) show that when sentiment increases on calls where analyst intensity is significantly higher than managerial intensity, and managerial jitter is significantly greater than analyst jitter the market responds in a heightened manner shown through greater reactions in short-term CARs. An interpretation of these results further suggests that these subsets of calls provide evidence that financial market participants are persuaded into making decisions in line with the way in which information is portrayed rather than the underlying meaning of the information produced, principally for initial decisions around the call date. Particularly, on calls when managers are perceived as calm and on calls when analysts are perceived as confident, there is a greater initial overreaction in CARs around the time of the earnings conference call. Providing evidence in line with behavioural theory by again identifying a potential framing bias having a significant impact on the decision-making process of market participants.

From the computational perspective, the results identify that the multimodal sentiment classifier is substantially more adept at predicting short- and longer-term CARs than that of singular modality classifiers that have been used in similar studies in the past (Doran et al. 2012; Mayew and Venkatachalam, 2012; McKay Price et al. 2012). The multimodal classifier returns a higher coefficient of determination (R^2) implying that the combination of textual and paralinguistic data captures a sentiment variable that is more adept at understanding fluctuations in market behaviour over and above models which only consider single modalities of information. Particularly, this chapter identifies that leveraging both modalities on earnings calls substantially increases the ability to forecast short-term CARs, returning R^2 values of more than 0.6 points above the previously conducted studies answering the same question.[256]

This research strongly supports existing literature, underscoring the superiority of computationally advanced methods over traditional dictionary-based approaches in capturing financial sentiment (Kearney and Liu, 2014; Guo, Shi and Tu, 2016; Renault, 2017; Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Alamoudi and Alghamdi, 2021). Moreover, our findings reinforce the

[256] Increases in R^2 for longer horizon CARs are also found but this increase is marginal and not as pronounced as that of the short-term analysis.

conclusions drawn by Mayew and Venkatalcham (2012) and broader social psychology research, indicating that non-verbal cues play a modest yet significant role in communication. Specifically, within financial contexts, this study demonstrates that paralinguistic cues from earnings calls offer additional insights into the decision-making process of market participants and identifies the use of such cues in sub-rational decision making in reaction to the information conveyed on earnings calls.

6. An Event Study of Multimodal Earnings Call Sentiment and Abnormal Trading Volume

6.1 Introduction

Behavioural finance is a branch of finance that explores how psychological biases and cognitive errors influence investor behaviour and market outcomes (Shefrin, 2001). Unlike traditional finance theories, which assume that investors are rational and always in in their best interest, behavioural finance recognises that human decision-making is often influenced by emotions, heuristics and social factors. It studies phenomena such as overconfidence, loss aversion, framing and anchoring, which can lead to sub-rational investment decisions and market inefficiencies. By understanding these behavioural biases, researchers and practitioners in behavioural finance aim to improve the understanding of financial markets, develop more accurate models of investor behaviour, and ultimately enhance the investment decision-making processes.

Whilst there are many areas within behavioural finance, one that has gained traction in recent years is the investigation into the relationship between investor sentiment and stock market characteristics (Lopez Cabarcos et al. 2020). A particular market characteristic which has received attention surrounding its relationship with sentiment is trading volume. Models of investor sentiment, such as De long et al. (1998) theorise that low levels of sentiment will generate downward price pressure while extreme high or low levels of sentiment will induce high levels of trading volume. This chapter aims at making a theoretical addition to the body of work that has evaluated the relationship between sentiment and trading volume by providing further evidence about the sentiment and trading volumes relationship. In asset pricing literature there are two theories which offer different perspectives to the sentiment and trading volume relationship – informational and sentiment theory. Traditional theory suggests an informational theory which theorises that sentiment acts as a proxy for fundamental information and therefore positive (negative) sentiment generates a positive (negative) reaction in returns and consequently creates an increase in trading volume in both cases as investors respond to the news. Over the longer horizon, the information theory posits that the news provided within earnings calls is incorporated into share prices efficiently in the short -term and thus following the initial increase in trading volume, no subsequent trading occurs.

Traditional theory and behavioural theory both agree on the way in which trading volume reacts in the short term but have two differing views on what creates these reactions. Behavioural theory aligns with sentiment theory which indicates that short-horizon returns will be reversed in the long-term and sentiment will initially create increased trading volume as market participants overestimate the information provided and consequently lead to a surge in buying or selling activity. Over the longer horizon, the sentiment theory speculates that trading volume is sustained as the overreactions to initial information move prices away from fundamental values and hence the market reacts with continued trading in attempts to correct mispricing. The difference between the explanations provided by the traditional and behavioural theories for short-term volume is that the traditional theory implies that trading volume is increased due to the release of new fundamental information to the market. This initial uptick in trading volume is thought to continue due to rational market participants trading on new information and consequently moving prices to fundamental values quickly with no subsequent movements. Alternatively, behavioural theory emphasises the psychological responses to information and the subsequent biases introduced within market participants' decision-making processes. Thus, there will be an initial uptick in trading volume when information is released and a continued increased level of trading volume as market participants attempt to correct initial mispricing.

Expanding on this relationship, trading volume serves as a key measure capturing the intensity and duration of market reaction to information. From an informational perspective, trading volume spikes immediately following the release of new fundamental information, as investors adjust their expectations. This view predicts that trading volume will quickly move to fundamental levels as prices fully incorporate the news, resulting in a short-term volume effect closely tied to sentiment. Over the longer horizon, the informational perspective indicates that volume will revert to baseline levels as the new information has been acted upon and calculated into the updated fundamental price. Sentiment theory proposes a more nuanced interpretation to the sentiment trading volume dynamic. It suggests that investor reactions to sentiment are influenced by cognitive biases and emotional responses triggered not only by the content but also by the delivery of information — factors especially salient in multimodal contexts where vocal tone and expression carry additional meaning. Under this framework, trading volume is not limited to the short-term but can persist over longer horizons as market participants engage in continued trading to resolve mispricing caused by initial over- or under-reactions to sentiment.

This chapter attempts to understand the behaviour of financial market participants who leverage earnings conference calls as a source of information for financial decision-making purposes. In doing so it aims to contribute to the literature previously undertaken surrounding sentiment and trading volume by adopting a state-of-the-art multimodal sentiment classifier for financial decision making introduced in Chapter 4. To the authors' knowledge, previous research evaluating the relationship between sentiment and trading volume has used dictionary or traditional machine-learning approaches to determine financial sentiment (see Chapter 2). The inclusion of a more robust model, which takes into consideration a previously under used modality of information, for financial sentiment classification therefore has the potential to further contribute to our understanding of market dynamics.

As this research takes into consideration a multimodal sentiment classifier, leveraging behavioural cues that are shown to impact decision-making, the application of such classifier will provide a more robust investigation into whether market behaviour follows an information or a sentiment theory. Particularly, this analysis aims to identify if psychological biases occur in financial market agents' decision-making process, leading them to initially overvalue the information provided on earnings calls and subsequently move prices away from fundamental values. Therefore, it contributes to prior literature that has looked at the relationship between sentiment and trading volume (Antweiler and Frank, 2004; Tetlock, 2007; McKay Price et al. 2012; Garcia, 2013; Sprenger et al. 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014:2017), and explores the average levels of earnings conference call multimodal sentiments impact on Cumulative Abnormal Trading Volume (CAV).

The relationship between the absolute values of sentiment and trading volume have been well documented (Baker and Wurgler, 2006). However, Siganos et al. (2017) highlight that an absolute level of sentiment residing at zero on a given day is the same as half of the population having a positive sentiment and half having a negative sentiment. Therefore, as a natural continuation of the investigation into average levels of sentiment conducted in the main analysis of this chapter, an additional analysis exploring the relationship between earnings calls that exhibit the highest levels of participant disagreement, through four Divergence of Sentiment (DoS) measures, and the same CAV variables is included following prior research (Karpoff 1986; Harris and Raviv, 1993; Hong and Stein, 2007; Banerjee and Kremer, 2010; Atmaz and Basak, 2016).

This paper adds to the finance literature in two ways. Firstly, it evaluates the relationship between average levels of earnings call sentiment and trading volume. Leveraging the multimodal sentiment analysis classifier, which has been shown to outperform the most used sentiment classifiers in previous research, the main empirical analysis shows that the multimodal financial sentiment classifier returns a positive but insignificant relationship with short-term CAVs whilst returning a negative and highly significant relationship with extended CAVs. More succinctly, as the sentiment on earnings calls increases, initial trading volumes also increase, before decreasing over subsequent days. The initial positive reaction in trading volume to sentiment reacts in line with both the sentiment and informational theories however this result is insignificant. Over the longer horizon, the negative and statistically significant result of CAV in relation to multimodal sentiment implies a reduction in trading volume back to base line levels which would follow the informational theory. In general, neither traditional or behavioural theories fully explain the results found within the main analysis which looks at the relationship between absolute values of multimodal earnings conference call sentiment and CAV.

Secondly, it builds upon the first paper to use a divergence of sentiment measure as a proxy for investor disagreement (Siganos et al. 2017). In doing so this chapter evaluates a more focused dataset of earnings conference calls, assessing the relationship between calls with high levels of participant disagreement and abnormal trading volume. This is the first study, to the author's knowledge, that identifies disagreement using linguistic and paralinguistic content for earnings conference calls. In the

DoS analysis it is found that calls typified by high (low) levels of disagreement evoke a larger (smaller) short-term market reaction when compared to the main results evaluating absolute levels of sentiment. The results indicate that DoS, which identify calls with significant differences in managerial and analyst sentiment, are robust predictors of short-term cumulative abnormal volume (CAV). While both types of call sentiment (managerial vs. analyst optimism) predict short-term CAV, calls with greater managerial optimism elicit a heightened market reaction in CAV. This pattern holds both for the overall call and the Q&A section.[257] In summary, when there is a sentiment disagreement, particularly with managers being more optimistic, the market responds with increased abnormal trading volume in the short term identified through stronger statistical significance and larger positive coefficients than the main results. Multimodal sentiment becoming statistically significant in the DoS analysis falls in line with theoretical models of disagreement which indicate disagreement between market participants induces heightened trading.

Models relating to market behaviours in correspondence to investor disagreement do not make predictions about market movements over the longer term. Therefore, within this analysis I provided information on the relationship between disagreement on earnings calls and its impact on longer horizon CAV. The findings indicate that, calls which display disagreement in sentiment between manager and analyst participants reduce CAV in the long term. Although Tables 6.4 and 6.5 differ on which participants' optimistic sentiment affects long-term trading volume, both indicate that increased sentiment in calls with disagreement reduces trading volume over the long term. This may suggest that disagreement between managers and analyst on earnings calls reveals new fundamental information, leading to better initial decision-making and stabilizing prices at fundamental values, thereby dampening trading volume over time.

The remainder of the paper is structured as follows. Section 6.2 identifies the relevant literature pertaining to the relationship between sentiment and trading volume to develop appropriate hypotheses. Following on Section 6.3 discusses the dataset being used to test these questions. Section 6.4 identifies the empirical methods being applied to the dataset in order to robustly answer the questions posed. The results associated with this analysis are then contained within Section 6.5 with Section 6.6 concluding the paper.

6.2 Hypothesis Development

As evidenced in the Chapters 3 and 5 of this thesis, a substantial amount of research has evaluated the relationship between financial sentiment and abnormal returns, of which a number also consider the relationship between sentiment and trading volume (Antweiler and Frank, 2004; Tetlock, 2007; McKay Price et al. 2012; Garcia, 2013; Sprenger et al. 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014:2017). Hong and Stein (2007) highlight that many interesting phenomena between pricing and

[257] However, these results become insignificant and return a negative coefficient when the relationship is evaluated using a dummy variable regression in Table 6.5.

returns are closely linked with trading volume,[258] noting that any comprehensive asset pricing model will give a prominent role to trading volume. Hence, evaluating the relationship between multimodal sentiment and trading volume will allow us to more directly measure the impact that sentiment has on firm level trading volume, alongside the previously studied abnormal returns. As the multimodal sentiment classifier identified a framing bias in Chapter 5 of this thesis in relation to abnormal returns, this empirical chapter continues to investigate whether the framing of earnings call information has any significant impact on the decision-making process of financial market participants that can be identified through CAV. Continuing in the same vein as Chapter 5, the inclusion of paralinguistic variables known to impact perceptions of speakers and decision making will allow this study to identify whether the way in which information is framed on earnings calls has any significant impact on trading volume.

The consensus in the literature surrounding the relationship between financial sentiment and trading volume denotes that high levels of positive or negative sentiment will increase trading volume (De Long et al. 1990). If market participants are exposed to high levels of positive sentiment surrounding a specific firm, this positive signal about future firm performance leads to increased trading volume, as investors attempt to capitalise on the perceived future profits. Alternatively, if there is a negative consensus around a given firm, investors change their future belief around future share return prospects to a negative one and look to sell out of their position to avoid potential losses. This has been evidenced using a number of different financial sentiment indicators calculated using information sources other than earnings conference calls (Mao and Bollen, 2011; Garcia, 2013; Sprenger et al. 2013; Bochkay et al. 2020; Gu and Kurov, 2020).

Antweiler and Frank (2004) show that levels of posting activity on internet message boards predict increased levels of contemporaneous trading volume. Similarly, Sprenger et al. (2013) identify that a 1% increase in twitter message volume translates into a 10% increase in trading volume. In both cases increases in message volume, either on message boards or social media, are thought to indicate the arrival of new information to the market. Furthermore, many messages posted on these platforms stem from *noise traders* and represent buy signals (Dewally, 2003). Hence, the assumption that increased message volume represents agreement amongst market participants using these platforms, and in other words positive market sentiment, can be made. Looking directly at the relationship between calculations of sentiment and trading volume, Tetlock (2007) find that unusually high or low levels of pessimism, calculated using the Wall Street Journal, predict increased trading volume at the market level. Similarly, Siganos, Vagenas-Nanos and Verwijmeren (2014) show that negative sentiment calculated from Facebook posting activity has a significant negative contemporaneous relationship with

[258] The authors give examples of high-priced glamour stocks tending to exhibit higher trading volume in comparison to low-priced value stocks, holding all else equal. Hong and Stein (2007) also note a stocks future returns tend to be lower when it has higher trading volume, controlling for a stock’s ratio of price to fundamentals. In other words, trading volume seems to be an indicator of sentiment.

trading volume across international markets. In other words, negative sentiment is associated with an increase in contemporaneous trading volume.[259]

Focussing specifically on earnings call sentiment and trading volume, Frankel et al. (1999) find that earnings conference calls add new information to the market and consequently increase trading volume.[260] McKay Price et al. (2012) finds a positive relationship between sentiment and trading volume for both short and longer period analyses. The positive and significant sentiment variables over the longer horizon indicate that the market more slowly incorporates information conveyed by positive call sentiment. In accordance with the extant literature discussed surrounding sentiment and trading volume, this chapter creates the following hypothesis:

H1: High absolute values of multimodal sentiment have a significant positive relationship with short-term CAV.

McKay Price et al. (2012) contribute the only study to evaluate the impact earnings call sentiment has on abnormal trading volume over long-term trading horizons. The authors find that after the initial increase in trading volume around the call date, trading volume remains present over the longer horizon as evidence through a positive regression coefficient although this coefficient is small. They explain this continued trading volume post call using an information theory suggesting that the market slowly incorporates the fundamental information produced on these calls. Informational theory suggests that market movements are induced by new information being released to the market. This information is then used by market participants to adjust their future expectations surrounding a given firm and for the subsequent purchase or sale of said firm shares, hence increasing trading volume initially. McKay Price et al. (2012) findings imply that market participants continue to trade on the information produced on earnings calls over longer horizons, implying that the market slowly incorporates the fundamental information from these calls. However, from a purely theoretical perspective, the information theory indicates that over the longer-term, trading volume diminishes as prices move quickly to fundamental values and hence no new trading ensues.

The alternative to the informational theory in explaining market movements is the sentiment theory (Tetlock, 2007). A sentiment theory indicates that returns deviate from fundamentals when new information comes to light due to the sub rational decision making of market participants. In contrast to the informational theory which sees returns quickly move to fundamental values, the sentiment theory implies that market participants overestimate the news in any new information which consequently leads to a reversal back towards fundamentals. For instance, during periods of positive (negative) sentiment market participants may become overly optimistic (pessimistic) about future returns leading to overvaluation (undervaluation) which subsequently reverses. Tetlock (2007) concisely explains these

[259] This result falls in line with Erbert and Tesser (1993) findings that negative sentiment causes investors to trade more, as they look to overcome their negative sentiment with a positive outcome from an alternative activity.

[260] The authors note that increased trading volume is seen during these calls and conclude that large investors trade in real time during these calls based on the information conveyed.

two theories in relation to longer horizons returns and volume by noting that the sentiment theory predicts initial overreaction in both returns and volume, followed by a reversion in returns over the longer horizon which is facilitated by persistent volume as market participants attempt to capitalise on mispricing. Alternatively, the information theory predicts that returns will persist indefinitely implying trading volume will revert back to baseline after the initial reaction.

McKay Price et al. (2012) use traditional classifiers[261] to define earnings conference call sentiment in comparison to the multimodal model used within this chapter. Multimodal sentiment classifiers have been shown to have greater capabilities in understanding natural language in alternative domains (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021) and the specific classifier used within this analysis has been shown in Chapter 4 to produce stronger classification accuracy on an earnings conference call dataset. Furthermore, the incorporation of paralinguistic traits, that have been shown in previous psychology literature to impact the decision-making process are included within this model (Mehrabian and Williams, 1969; London, Meldman and Lanckton, 1970; Erickson et al. 1978; Edinger and Patterson, 1983). In Chapter 5 of this thesis this multimodal classifier showed increased capabilities in predicting abnormal returns and identified a framing bias present in market participants decision making process. Due to the multimodal model having been shown to (i) have a greater understanding of natural language over text based classifiers in a financial context, (ii) includes paralinguistic information that is shown to impact decision-making and (iii) its ability to detect framing bias in Chapter 5 of this thesis, we posit that the behavioural insights associated with this classifier will identify that financial markets reaction in trading volume will follow that of the sentiment theory.

A relationship between longer-period CAV and sentiment is expected based on both behavioural finance theory and the dynamics of information processing in financial markets. Sentiment — particularly when captured multimodally — conveys not only the semantic content of communication but also affective signals such as tone, emphasis, and emotional intensity. These non-verbal cues can have a prolonged influence on investor psychology, shaping perceptions of risk, confidence, and uncertainty over time. In cases where information is ambiguous, the delivery and tone of communication can heavily influence how investors interpret its implications, potentially leading to extended trading activity as market participants continue to assess and reassess their positions. Particularly, if framing effects are evident in financial decision-making processes, which the results of Chapter 5 suggest, the heuristic may also anchor investors to certain narratives that influence trading beyond the initial disclosure window. As such, multimodal sentiment may not only trigger immediate reactions but also exert a persistent effect on trading volumes as market participants digest the

[261] McKay Price et al. (2012) compare the performance of a general dictionary (Harvard IV-4 Psychosocial) and a specific dictionary (Henry (2008) Dictionary).

communicated information in light of new developments or evolving market conditions. Therefore, the following hypothesis is proposed:

H2: Multimodal sentiment has a significant positive relationship with longer period CAV.

The literature cited in the development of H2 demonstrates the relationship financial sentiment has on trading volume. If the market has high absolute values of sentiment, we can assume that there is consensus among market participants. In other words, a high positive sentiment entails market participants agree on a buy signal with high negative sentiment implying market participants have negative future expectations and look to sell.[262] In both cases, the market agrees surrounding the future expectations of a given firm and hence increased trading ensues. A substantial amount of literature has also focused on the impact disagreement among investors has on trading volume (Karpoff, 1986; Harris and Raviv, 1993; Antweiler and Frank, 2004; Banerjee and Kremer, 2010; Atmaz and Basak, 2016). Disagreement in financial markets is defined as the difference in opinions of market participants surrounding the current and future value of a given asset.

Karpoff (1986) examines the relationship between trading volume and price changes, finding that the presence of disagreement among investors is a key driver of trading volume. Higher disagreement correlates with increased trading volume, especially during periods of significant market events or announcements. Harris and Raviv (1993) explore the role of information asymmetry in financial markets. Differences in information or interpretation can lead to disagreement among investors. Disagreement is linked to increased trading volume as investors seek to exploit differences in valuation, leading to active trading to correct perceived mispricing’s. Hong and Stein (2007) discuss how investor sentiment and attention affect market dynamics. When investors disagree, it often leads to differing trading strategies and behaviours. Disagreement results in higher trading volume, particularly when sentiment shifts or new information emerges, as investors attempt to align their positions with their beliefs about the market. Banerjee and Kremer (2010) analyse how social interactions and information dissemination among investors can lead to differing opinions. Disagreement can arise from different interpretations of available information or varying levels of information access. Increased disagreement tends to drive up trading volume, as investors engage in speculative trading to capitalize on perceived mispricing’s based on their unique views. Atmaz and Basak (2016) focuses on the dynamics of investor sentiment and how it influences trading behaviour. High levels of disagreement among investors can lead to increased trading volume as individuals react differently to new information, resulting in heightened market activity. Greater disagreement often correlates with higher volatility and trading volume, as investors actively adjust their positions based on differing expectations. In summary, the findings from this area of research imply that disagreement between market participants on the

[262] This is consistent with Cao et al. (2020) who suggest that individuals are more likely to trade when they know others have received the same signal as them. As this agreement takes time to be revealed, they note that higher levels of agreement will result in more trades the following day.

prospects and risks of a given firm leads to disagreement on the value of a share. Consequently, disagreement is thought to be positively related to trading volume.

Hong and Stein (2007) highlight three underlying mechanisms, gradual information flow, limited attention, heterogenous priors, that are thought to create differences in investors prior beliefs and hence generate disagreement amongst investors. Gradual information flow highlights that certain pieces of value-relevant information will reach some investors before others. If the information is positive, the investors who receive it first will revise their valuations of the stock upward, while those who have not yet received it will maintain their current valuations. This will increase the disagreement between the two groups of investors, leading those who received the information first to buy from those who have not yet seen it. Several recent papers, including Hirshleifer and Teoh (2003) and Peng and Xiong (2006), emphasise the concept of limited attention, where cognitively overloaded investors focus on only a subset of publicly available information. For practical purposes, this concept is similar to gradual information flow, though it places less emphasis on the dynamics of information diffusion. Like gradual information flow, limited attention alone does not produce significant patterns in prices or volume. Instead, it needs to be combined with the assumption that investors are unsophisticated in another distinct way: they do not account for the fact that their valuations are based only on a portion of the relevant information when trading with others. The underlying mechanism, heterogenous priors, that creates disagreement implies that even when a piece of news is made publicly available to all investors at the same time, and they all pay attention to it, the news can still increase their disagreement about the fundamental value of the stock. As discussed by Harris and Raviv (1993) and Kandel and Pearson (1995), this occurs if investors have different economic models that cause them to interpret the news differently. Each of the underlying mechanisms discussed to produce disagreement among investors all have different distinct features but all share a common feature of identifying disagreement among investors.

Although each of the disagreement mechanisms above can be tested within the market, this chapter focuses directly on earnings conference calls. Therefore, this analysis looks at disagreement between manager and analyst content on these calls. Previous literature evaluating market responses to earnings conference call content have identified that investors in general react differently to the sentiment of managers and analysts (Brockman, Li and McKay Price, 2015; Chen, Nagar and Schoenfeld, 2018). Brockman, Li and McKay Price (2015) evaluate the relationship that both managerial and analyst sentiment has with abnormal returns. They find that both sets of call participants sentiment levels are significantly associated with abnormal returns. However, results indicate that the market reacts to analyst sentiment in a more pronounced manner in comparison to managerial sentiment. Chen, Nagar and Schoenfeld (2018) also analyse manager-analyst conversations on earnings calls finding similar results. The authors show that intraday share prices significantly respond to analyst sentiment with evidence suggesting that the effect strengthens when analyst sentiment is relatively negative. Therefore, as there is evidence indicating that the market responds differently to manager and analyst sentiment produced on earnings calls, we posit that differences in managerial and analyst sentiment may be an underlying mechanism that creates differences in investors' beliefs and hence produces disagreement between the wider market.

Particularly, we build upon the initial work conducted by Siganos et al. (2017) who develop a DoS measure that acts as a proxy for investor disagreement, measured as the distance between positive and negative sentiment of Facebook status updates, and evaluate the contemporaneous reaction in trading volume at the country level. In their paper, the authors use their DoS measure as a proxy for investor disagreement, noting that a higher divergence of sentiment implies diverging views on financial prospects and risks. Their findings fall in line with the previously mentioned disagreement literature, showing that divergence of sentiment is positively related to trading volume. They also highlight the limitations surrounding their research, such as: (i) DoS is hard to quantify, (ii) it is difficult to control for the arrival of news, and (iii) they are constrained to relatively general tests due to their dataset of Facebook status updates.[263] Building on this research, this chapter attempts to adopt a more accurate DoS measure by applying the previously mentioned multimodal sentiment classifier to a more focused financial information dataset: corporate earnings calls. Our earnings call dataset allows us to control for other factors known to impact market characteristics around call dates, and therefore focus directly on the market reaction in trading volume to differences in opinions on earnings calls.

This chapter builds upon Siganos et al. (2017) by adopting a divergence of sentiment measure to quantify differences of opinions on earnings conference calls. In this chapter, disagreement is defined as the difference in sentiment between participants on these calls. We assume that this divergence in sentiment acts as a proxy for a difference of opinions between call participants and hence adds noise to the fundamental information produced on earnings calls. Furthermore, in line with Brockman, Li and McKay Price (2015) and Chen, Nagar and Schoenfeld (2018), it is expected that market participants view the sentiment of manager and analyst differently. Differences in the weighting provided to each participant groups sentiment may imply a disagreement in the information provided and hence in line with the theoretical models of investor disagreement (Karpoff, 1986; Harris and Raviv, 1993; Antweiler and Frank, 2004; Banerjee and Kremer, 2010; Atmaz and Basak, 2016) heighten abnormal trading volume in the short-term. Models of investor disagreement do not speak to the relationship between sentiment and longer period trading volume. Miller (1977) reports that disagreement among investors, when there is the existence of short-selling constraints within the market, leads to overpricing. Similar to that of sentiment theory, as an initial mispricing occurs, we posit that trading will continue to persist over the longer horizon as the market attempts to move prices back to fundamental values. Therefore, the following hypothesis has been established.

[263] The authors use general Facebook status updates with no specific focus on financial information or posts that are related to financial decision making.

H3: Multimodal sentiment from earnings calls defined by high levels of divergence has a significant positive relationship with short-term and longer period CAV.

6.3 Data

The dataset employed in this study comprises financial data and earnings call information for 95 distinct firms spanning a 15-year duration from 2006 to 2021. Throughout this period, these firms collectively participated in 4,928 earnings calls, generating a corpus of 637,220 sentences, all of which encapsulate a comprehensive array of paralinguistic features.[264] Selection of these 95 firms was predicated upon their inclusion in the S&P 100 index as of 2021, with relevant data for each firm sourced from FinnHub[265] and Refinitiv.[266,267] Detailed information regarding the companies included, encompassing each firm's market capitalization, industry and sector classification, headquarters location, and the volume of calls attributed to each firm across the timeframe under examination, is presented in Appendix 3.1. Additionally, an illustrative breakdown of this dataset categorized by industry and market capitalization is depicted in Figure 5.1 in Chapter 5.

6.4 Empirical Method and Content Analysis

6.4.1 Empirical Method

In Chapter 4, a comparison is presented between the most frequently used methods to determine financial sentiment and our trained multimodal sentiment classifier. The results from the previous chapters indicate that the multimodal classifier is more adept at classifying financial sentiment of earnings calls and produces the most robust measure of sentiment for predicting abnormal returns. Hence, the analysis completed within this chapter focuses on further evaluating the relationship between multimodal sentiment and CAV. Following on from associated literature (Tetlock, 2007; Tetlock et al. 2008; Engelberg, 2008; Frankel et al. 2010; McKay Price et al. 2012) this analysis adopts the same equation used to assess Cumulative Abnormal Returns (CARs), however changing the dependent variable to CAV and removing the volume control variable from the analysis. McKay Price et al. (2012) also adopt this approach, noting that the control variables used within equation 6.01 are appropriate for confirming the effects of conference call sentiment on trading volume.

$$CAV_j = \propto_0 + \propto_1 SENTIMENT_{i,j} + \propto_2 SURP_{i,j} + CONTROLS_{i,j} + \in_{i,j} \; [6.01]$$

In the above equation 6.01, CAV_j represents the cumulative abnormal trading volume for conference call j. Adopting the same approach as Barber and Odean (2008) and McKay Price et al. (2012), we calculate daily abnormal volume for firm *j* at time *t* ($\text{AV}_{\text{j,t}}$) as such:

$$AV_{j,t} = \frac{V_{j,t}}{\overline{V_{J,t}}}$$

[264] For a list of these paralinguistic features and a description of how the features were calculated see Chapter 3.5.

[265] FinnHub was used to download earnings conference call transcripts and corresponding audio.

[266] Refinitiv was used to download all data used surrounding company performance.

[267] For a full insight into the process of downloading and cleaning data see Chapter 3.

Where $V_{j,t}$ is the volume for a specific firm *j* on day *t*, and $\overline{V}_{j,t}$ is the average trading volume for firm *j* across the period t = -252 to t = -2, which is calculated as follows:

$$\overline{V_{J,t}} = \sum_{t-252}^{t-2} \frac{V_{j,t}}{252}$$

$AV_{j,t}$ is then cumulated over two event periods. The first period CAV(-1,1) captures the initial period, summing together abnormal trading volume for the three days surrounding an earnings call from a day before (t-1) to a day after (t+1). The CAV(-1,1) allows us to assess the initial market reaction in abnormal trading volume in relation to earnings call sentiment. The second period looks at a longer period cumulation of abnormal volume, summing together daily abnormal values across 2 days post (t+2) to 60 days post (t+60) earnings call, CAV (2,60). This longer period measure allows for an understanding of the lasting effects earnings call sentiment has on abnormal trading volume over a longer horizon.

$SENTIMENT_{i,j}$ in equation 6.01 represents the sentiment variable calculated for firm *j* at time *i*. Following Antweiler and Frank (2004) and the previous chapter, we run each regression with various aggregations of the sentence level sentiment classifications from the multimodal model. Further insight into the different methods of aggregating earnings call sentiment is provided in section 6.4.2. The next variable included within equation 6.01, $SURP_{i,j}$, represents unexpected earnings (also known as earnings surprise) for firm *j* at time *i* and is calculated using a seasonal random walk approach in line with (Henry, 2008; Sadique, 2008; Akbas et al. 2013) as follows:

$$SURP_{i,j} = \frac{EPS_{i,j} - EPS_{i-4,j}}{PRICE_{i-4,j}}$$

Where:

$SURP_{i,j}$= The unexpected earnings for firm *j* at time *i*.
$EPS_{i,j}$= The earnings per share for firm *j* in quarter *i*.
$EPS_{i-4,j}$= The earnings per share in the same quarter of the previous year for firm *j*.
$PRICE_{i-4,j}$= The closing stock price of firm *j* in the same quarter of the previous year.

Most prior literature use a seasonal random-walk approach to calculate unexpected earnings, however recent studies have used an alternative approach – analyst-based earnings surprises (Ayers, Li and Yeung, 2011). Following the majority of prior literature, this analysis adopts the seasonal random-walk approach in calculating unexpected earnings.[268]

The last variable in equation 6.01, $CONTROLS_{i,j}$, represents six variables used to control for external factors known to impact abnormal returns. Thus, the inclusion of these variables allows this analysis to better isolate the impact that sentiment has on CAVs. The six variables include measures of call length, firm size, book-to-market equity, profitability, leverage and returns volatility. Call length is an indicator of the number of sentences within each earnings call. This variable is split into two columns which relate to the number of sentences associated with both portions of the earnings call, namely

[268] Sadique (2008) uses both measures and conclude that there is not any significant difference between them.

Management Discussion (MD) and Question & Answer (Q&A) sections. Firm size is calculated as the logarithm of the market capitalisation of a specific firm at the end of the quarter prior to the earnings call. Book-to-market equity was defined as the reciprocal of the market-to-book equity value directly download from Refinitiv. Profitability (also known as return on assets, ROA) is calculated as the ratio between a given firm's net income against total assets, multiplied by one hundred to create a percentage. Leverage is calculated as the ratio of total liabilities to total assets scaled by one hundred. Finally, returns volatility is calculated as the standard deviation of a given firm's daily returns across the period 90-day period beginning 100 days before a given call to 10 days before the call date.

This analysis focuses on the relationship between earnings call sentiment classifications of a multimodal sentiment classifier developed for financial decision making. In doing so the multimodal sentiment classifier is applied to 4,928 conference calls that form the basis of the earlier Chapter 5. The multimodal sentiment model's classifications of the full sample of sentences have been aggregated into call level sentiment indicators using four differing methods identified in the following subsection 6.4.2.

6.4.2 Sentiment Analysis

To create sentiment measures for each earnings call, this chapter applies the multimodal sentiment classifier introduced in Chapter 4 to each individual sentence on all 4,928 conference calls within the sample set.[269] Each sentence is then classified as positive (1), negative (-1) or neutral (0). To generate a sentiment value for each call we adopt four varying methods to aggregate the sentence level classifications made by the multimodal classifier. In line with Antweiler and Frank (2004), this analysis adopts four methods to aggregate sentence level classification into a call level sentiment indicator.

The four formulas used to aggregate messages include the three measures used by Antweiler and Frank (2004) and a basic summation of sentiment. The first measure, the basic sentiment (BS) measure, can be calculated as follows:

$$BS = \frac{S^{Positive} - S^{Negative}}{N.\ of\ Sentences} \quad [6.02]$$

Where:

$S^{Positive}$= The number of positive sentences on a call.
$S^{Negative}$= The number of negative sentences on a call.
N. of Sentences = The total number of sentences in a call i.e., all positive, negative and neutral sentences.

This measure sums together all positive and negative sentences used within a call and divides by the total number of sentences in a call – positive, negative and neutral messages. The basic sentiment measure provides a call level sentiment figure that is bound between -1 and 1 i.e., a call with a sentiment of 1 contains only positive messages with a call sentiment value of -1 indicating that the call only contains negative messages. The following measures used to define sentiment in equations 6.03, 6.04 and 6.05 all stem from Antweiler and Frank's (2004) paper. The first measure, defined in the following

[269] The multimodal sentiment classifier for financial decision-making is trained at the sentence level and hence must be applied to earnings calls sentences for the best results.

tables as AF1, is the measure that Antweiler and Frank (2004) use extensively in early drafts of their influential paper. The measure is calculated in a similar fashion to the above BS measure and is also bound between -1 and 1, the only key difference is that the denominator only sums together the positive and negative messages.

$$AF1 = \frac{S^{Positive} - S^{Negative}}{S^{Positive} + S^{Negative}} \text{ [6.03]}$$

Antweiler and Frank (2004) highlight that this measure can be used to obtain all key results from their paper however they prefer the following measure:

$$AF2 = \log\left(\frac{1 + S^{Positive}}{1 + S^{Negative}}\right) \text{ [6.04]}$$

The authors note that the third and final measure created is similar to AF2 and that both AF2 and AF3 differ from the initial AF1 measure as both measures increase in magnitude as the number of messages being aggregated increase and as the ratio of positive to negative messages increase. However, AF1 and BS are homogenous with degree zero and hence independent of the number of messages considered.

$$AF3 = S^{Positive} - S^{Negative} \text{ [6.05]}$$

Each measure above has been used to calculate four different call sentiment figures for the multimodal classifiers sentence level classifications in relation to the 4,928 earnings calls considered within this analysis.

6.5 Results

6.5.1 Descriptive Statistics

Descriptive statistics along with correlations for the four sentiment aggregations of multimodal sentiment and the two periods of CAV are shown in Table 6.1. Looking directly at the descriptive statistics for both CAV variables, the longer period cumulation of abnormal trading volume is substantially greater than the short-term variable. This is expected due to the longer period CAV cumulating abnormal trading volume over a longer time frame. The short-term CAV measure ranges from 1.31 to 43.44 with a mean of 5.11. However, when looking at the interquartile range we can see most observations are less dramatic and fall within 3.68 at the 25th percentile and 5.92 at the 75th percentile. Ranging from 20.36 to 421.01 with an interquartile range of 47.25 at the 25th percentile to 64.81 at the 75th percentile, the longer period CAV measure demonstrates a similar, less pronounced distribution, to the short-term variable. This indicates that abnormal trading volume does not react in a linear manner to each call within this sample at either the short or long term, with certain calls receiving significantly dampened or excessive trading reactions. Each of the aggregations of multimodal sentiment return consistent relationships with both CAV measures. Multimodal sentiment returns a positive relationship with short-term CAV and an inverse relationship with long-term CAV across all aggregations of sentiment in line with expectations.[270]

Table 6.1: Descriptive Statistics and Correlations for the Multimodal Sentiment Classifier and CAVs

	BS	AF1	AF2	AF3	CAV(-1,1)	CAV(2,60)
Panel A: Descriptive Statistics						
count	4928	4928	4928	4928	4928	4928
mean	0.11	0.24	0.50	15.03	5.11	58.70
std	0.14	0.28	0.63	18.65	2.34	20.28
min	-1.00	-1.00	-4.96	-142.00	1.31	20.36
25%	0.03	0.06	0.11	3.00	3.68	47.25
50%	0.12	0.25	0.50	15.00	4.59	54.46
75%	0.20	0.43	0.89	27.00	5.92	64.81
max	0.60	1.00	3.04	86.00	43.44	421.01
Panel B: Correlation						
BS	1.00					
AF1	0.95	1.00				
AF2	0.96	0.98	1.00			
AF3	0.97	0.92	0.94	1.00		
CAV(-1,1)	0.05	0.05	0.05	0.05	1.00	
CAV(2,60)	-0.09	-0.09	-0.08	-0.08	0.41	1.00

Notes: This table presents descriptive statistics (Panel A) and correlations (Panel B) pertaining to the S&P 100 sample encompassing 4,928 earnings conference calls across the period 2005-2021 used within this analysis. CAV(-1,1) denotes the short-term 3-day cumulative abnormal trading volume, with day 0 representing the conference call date. Abnormal trading volumes are estimated utilizing the approach which is elaborated upon in Section 6.4.1. Similarly, CAV(2, 60) is computed following the same methodology as CAV(-1, 1), albeit cumulated over the period spanning days 2 through 60. Each of the four sentiment indicators are aggregations of sentence level sentiment resulting from the application of the multimodal sentiment classifier, developed in Chapter 4, to the full sample of earnings call sentences. All four aggregation methods are discussed in Section 6.4.2.

Each aggregation of multimodal sentiment produces different distributions of call level sentiment. The BS and AF1 methods of aggregation produce a similar distribution of call sentiment for the 4,928 calls. This is due to both measures being homogenous with degree zero. In contrast, the second and third methods of aggregation introduced by Antweiler and Frank (2004) are not bound and tend to increase in magnitude as the number of sentences being evaluate increases. Interestingly, AF1 and AF2 return the strongest correlation with one another. These methods of aggregation also return the strongest relationship with CARs in the previous chapter.

Table 6.2: Test of Differences of Means by Multimodal Sentiment Quartiles

Panel A - Differences of Means			
		CAV(-1,1)	CAV(2,60)
1 (High)	mean	5.42	57.8
	std	2.37	17.61
	N. Observations	1232	1232

[270] See Chapter 6.2 for a discussion of previous literature surrounding sentiment and trading volume.

2	mean	4.93	56.88
	std	1.88	17.72
	N. Observations	1231	1231
3	mean	5.05	59.11
	std	2.48	21.92
	N. Observations	1232	1232
4 (Low)	mean	5.03	61.03
	std	2.57	23.08
	N. Observations	1233	1233
Panel B - T-tests			
Mean Q4 - Q1	T-statistic	-3.93	3.91
	P-value	0.00008***	0.00009***

*Notes: This table shows the differences in CAVs when sorted into sentiment quartile portfolios using the AF2 measure of aggregating sentiment. Panel A provides the mean and standard deviation of each sentiment quartile in relation to both CAV measures. CAV(-1,1) denotes the short-term 3-day cumulative abnormal trading volume, with day 0 representing the conference call date. Abnormal trading volumes are estimated utilizing the approach which is elaborated upon in Section 6.4.1. Similarly, CAV(2, 60) is computed following the same methodology as CAV(-1, 1), albeit cumulated over the period spanning days 2 through 60. Panel B shows the differences in the highest and lowest sentiment groups and provides the test statistic and p-values for this comparison. Significance level indicators: * at 10%, ** at 5%, *** at 1%.*

Panel A of Table 6.2 splits the 4,928 earnings calls sample into quartiles based on their sentiment scores and shows the mean and standard deviations of CAV for each quartile. From a theoretical standpoint (De Long et al. 1990), a U-shaped relationship between sentiment and abnormal trading volume is expected, where both highly negative and highly positive calls prompt stronger trading responses. Under this logic, earnings calls with moderate or neutral sentiment would generate less trading activity, as they may be perceived as less informative or ambiguous. However, the results in Table 6.2 do not support this expectation. Instead of a U-shape, the relationship between sentiment and abnormal volume appears asymmetric and directional. The mean levels of short-term abnormal volume increase through each quartile, with the exception of quartile three which, even though produces the second highest average level of sentiment, generates the lowest average reaction in abnormal trading volume. The standard deviations relating to reactions in short-term CAV decrease from the lowest quartile of sentiment to the highest quartile of sentiment. However, quartile three again goes against this trend returning a lower dispersion in short-term CAV in comparison to the highest sentiment quartile. Minus the results found in quartile three for the short-term CAV, the results indicate that there is a smaller reaction in abnormal trading volume to earnings calls that produce negative sentiment in comparison to the calls which produce positive sentiment. However, there is a more diverse reaction to calls producing negative sentiment than positive. Implying that market participants may more easily agree on trading signals when those signals are positive in nature.

For the longer-term CAV across each sentiment quartile, it is seen that as sentiment increases (from the lowest quartile to the highest quartile) the mean reaction in abnormal trading volume decreases. Similarly, the standard deviation in longer-term CAV also diminishes as the sentiment on these calls increases. Moreover, it is found that the market produces a larger response in CAV over the extended period to negative calls in comparison to positive calls. However, negative calls evoke a more varied market response in CAV which again could imply that market participants are on average in greater consensus to the information produced on positive calls than on calls exhibiting negative information. Furthermore, Panel B shows the results of the independent two-tailed t-test completed for the CAVs associated with the lowest and highest quartiles of call sentiment. The results imply that there is in fact a significant difference between the level of CAV associated with the most positive and most negative calls within this sample.

6.5.2 Main Results

The analysis contained within this section evaluates the explanatory power the multimodal sentiment classifier has on abnormal trading volume. Table 6.3 contains firm level regressions, analysing the impact that various aggregations of multimodal sentiment have on both short-term and longer period CAV. The results show that multimodal earnings call sentiment has a positive but non-significant relationship with initial period CAV. The highest positive multimodal sentiment coefficient relating to short-term CAV stems from the BS measure at 0.0234. However, the high p-value associated with this result (0.5548) suggests that the positive uptick in CAV is insignificant. From these results, it can be said that short-term CAV increases around the time of earnings conference calls however multimodal sentiment does not have any significant relationship with CAV. Therefore, H1 can be rejected confirming no association between multimodal sentiment and short-term CAV. In contrast to McKay Price et al. (2012) who find that the length of the Question-and-Answer (Q&A) section of earnings calls has a significant positive relationship with initial period CAV, our findings do not arrive at the same conclusion. Hence, from this analysis it can be confirmed that call sentiment canno explain fluctuations in short-term trading volume, and the length of call has no significant impact on initial firm level trading volume.

Comparing the results to the theoretical models of traditional and behavioural finance that explain trading volume market behaviours the results documented in this analysis are not in line with either theory. Traditional theory, which aligns with an informational story, implies that sentiment is a proxy for fundamental information. It identifies that trading volume increases surrounding the earnings call date as new fundamental information is brought to the market which induces market participants to begin trading and move prices to their new fundamental values. Following the initial spike in abnormal trading volume there is no subsequent trading based upon the information released in earnings calls as it is already incorporated into prices. Behavioural theory implies that market participants overweight their initial decisions based on the information provided in earnings calls which initially increases trading volume but consequently moves share prices away from fundamental values. Specifically for earnings calls, a potential bias that may cause overweighting of information on these calls is a framing bias where financial market participants overweight the information on these calls based on the way the information is disseminated, creating mispricing within the market.

The results from the main analysis, looking at absolute values of multimodal sentiment's relationship with CAV do not fit exactly with either of the theories discussed. Although there is an initial uptick in CAV, multimodal sentiment does not return a significant relationship with CAV. From a traditional theoretical perspective, where sentiment is a robust piece of fundamental information, the multimodal model doesn't contain any association with short-term CAV and such is not a robust predictor of fundamental information. Similarly, behavioural theory indicates, where psychological biases, particularly framing bias is included in financial market participant decision making, the behavioural cues contained within the multimodal model do not indicate evidence of any framing bias which would cause an uptick in short-term trading volume.

Table 6.3: Estimation of the Association between Various Calculations of Sentiment Models and Log(CAVs)

Sentiment Measure	Basic Sentiment		AF Measure 1		AF Measure 2		AF Measure 3	
Cumulative Abnormal Volume	(-1,1)	(2,60)	(-1,1)	(2,60)	(-1,1)	(2,60)	(-1,1)	(2,60)
Constant	2.1765*** (0.0000)	4.0429*** (0.0000)	2.1765*** (0.0000)	4.0423*** (0.0000)	2.176*** (0.0000)	4.0419*** (0.0000)	2.1795*** (0.0000)	4.0348*** (0.0000)
Sentiment	***0.0234*** ***(0.5548)***	***-0.0792****** ***(0.0087)***	***0.0131*** ***(0.5204)***	***-0.0373****** ***(0.0163)***	***0.0085*** ***(0.3424)***	***-0.0149***** ***(0.0291)***	***0.0002*** ***(0.4647)***	***-0.0005***** ***(0.0241)***
Earnings Surprise	-0.2599 (0.1318)	-0.0755 (0.5656)	-0.2588 (0.1333)	-0.0807 (0.5390)	-0.2621 (0.1285)	-0.078 (0.5527)	-0.261 (0.1302)	-0.077 (0.5581)
Management Discussion	0.0001 (0.6905)	0.0000 (0.9476)	0.0001 (0.6859)	0.0000 (0.9842)	0.0001 (0.6959)	0.0000 (0.9908)	0.0001 (0.7548)	0.0001 (0.7858)
Question and Answer	0.0002 (0.5032)	-0.0002 (0.3642)	0.0002 (0.5013)	-0.0002 (0.3510)	0.0002 (0.5066)	-0.0002 (0.3479)	0.0001 (0.5782)	-0.0001 (0.5482)
Market Capitalisation	-0.0452*** (0.0000)	-0.0065 (0.1216)	-0.0452*** (0.0000)	-0.0065 (0.1220)	-0.0453*** (0.0000)	-0.0065 (0.1188)	-0.0452*** (0.0000)	-0.0065 (0.12)
Book-to-Market	-0.0285* (0.0824)	0.0117 (0.3489)	-0.0285* (0.0823)	0.0121 (0.3307)	-0.0278* (0.0906)	0.0119 (0.3412)	-0.0283* (0.0847)	0.0121 (0.3323)
Profitability	-0.001 (0.1485)	-0.0005 (0.3881)	-0.001 (0.1493)	-0.0005 (0.3858)	-0.001 (0.1542)	-0.0005 (0.3741)	-0.001 (0.1490)	-0.0005 (0.3917)
Leverage	-11.3377*** (0.0001)	4.9896** (0.0267)	-11.3178*** (0.0001)	4.9816** (0.0271)	-11.2468*** (0.0001)	4.9783** (0.0272)	-11.3184*** (0.0001)	5.0378** (0.0253)
Volatility	-0.1597* (0.0887)	-0.2276*** (0.0015)	-0.1612* (0.0864)	-0.2266*** (0.0016)	-0.1645* (0.08)	-0.2291*** (0.0014)	-0.1611* (0.0859)	-0.2305 (0.0013)
N. Observations	4928	4928	4928	4928	4928	4928	4928	4928
R-sq	0.1844	0.1354	0.1844	0.1352	0.1845	0.1350	0.1844	0.1351
Adj R-sq	0.1787	0.1294	0.1787	0.1292	0.1788	0.1290	0.1787	0.1291
Industry Dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table provides cross sectional regression results of log CAV on measures of earnings call sentiment and control variable measures for the multimodal sentiment classifier. CAV (1,1) denotes the short-term 3-day cumulative abnormal trading volume, with day 0 representing the conference call date. Abnormal trading volumes are estimated utilizing the approach which is elaborated upon in Section 6.4.1. Similarly, CAV (2, 60) is computed following the same methodology as CAV (1, 1), albeit cumulated over the period spanning days 2 through 60. Each of the four sentiment indicators are aggregations of sentence level sentiment resulting from the application of the multimodal sentiment classifier, developed in Chapter 4, to the full sample of earnings call sentences. All four aggregation methods are discussed in Section 6.4.2. Industry and Time fixed effects are included to control for industry specific and year specific factors. N. Observations highlights the *number of calls included within each regression. R^2 determines the proportion of variance in the dependent variable that can be explained by the independent variable. Adj R^2 is a corrected goodness of fit measure of R^2. Regression p-values are in parenthesis. Significance level indicators: * at 10%, ** at 5%, *** at 1%.*

The relationship between multimodal sentiment and longer period CAV returns a highly significant coefficient at the 1% level, uncovering an inverse association. These results remain consistent for each aggregation of earnings call sentiment – all returning significance at the 1% or 5% level with negative coefficients. This implies that there is indeed a highly significant relationship between multimodal earnings call sentiment and longer period abnormal trading volume.[271] However, as the relationship is negative H2 is again rejected. The association between multimodal sentiment and longer period CAV aligns more closely with traditional theory, suggesting that over the long term, trading volume decreases. Informational theory indicates that this reduction in trading volume is generated as share prices in the short term quickly incorporate the fundamental information shared during earnings calls therefore no further opportunities are available to profit from a correctly priced share and thus no motivation to trade.

Although as there is no indication of an initial significant relationship between sentiment and CAV, the results here do not align fully with traditional or behavioural theory. To explain the relationship between earnings call sentiment and abnormal trading volume take for instance two calls, one positive and one negative. A positive call, based on our findings, creates an increased initial reaction in abnormal trading volume. This potentially suggests that the market agrees on the positive outlook of a given firm based upon the positive sentiment information disclosed to the market through its earnings call.[272] However, multimodal sentiment does not have any ability to explain this initial uptick in CAV. In this instance it is assumed that the market begins to buy the shares of said firm and trading volume increases. This increased trading volume is then followed by a drop in trading volume over the longer period analysis. This result suggests that the information portrayed to the market through the initial positive earnings call was accurate hence moving prices to their fundamental values with no future large movements.

A negative call does not generate a positive reaction in initial trading volume. As there is a positive relationship between sentiment and initial period CAV, the lower the sentiment on earnings call the smaller the initial market reaction. However, as there is an inverse relationship with longer-period CAV, the results imply that an earnings call which portrays negative sentiment information, will experience increased trading volume over the longer horizon. This could potentially be explained due to the short-selling constraints associated with noise traders within the market. If there is a negative sentiment, which would imply negative future expectations of returns, only sophisticated market agents

[271] This result returns greater statistical significance in comparison to McKay Price et al. (2012) who return significant results at the 5% level when using a dictionary approach to define earnings call sentiment.

[272] This falls in line with the descriptive results contained with Chapter 5.3 Table 5.3 which indicates that there is a greater consensus on the future of firm when calls are more optimistic compared to more pessimistic calls. This is shown through a smaller standard deviation in short- and longer-term CAR for calls in the top two quartiles of sentiment compared to the bottom two.

have the ability to short sell. If the trading volume in the short-term is driven by short sellers, it is expected that the share price to drop over the initial period. This lower share price may then attract investors to begin buying said share again and explain the increased trading volume over longer horizons for calls with negative information.

6.5.3 Divergence of Sentiment and Cumulative Abnormal Volume

The previous section’s results and McKay Price et al. (2012) identify that earnings conference calls contain value relevant information that significantly impacts both CAR and CAV. Particularly, they show that classifying the sentiment of these calls create a robust indicator of initial and longer period market characteristics. Future research stemming from their paper highlights the importance of studying earnings call dialogue at the participant level. Evaluating the nuances in communication across both sets of participants on these calls will allow for a deeper understanding of the implications of sentiment in the manager-investor communication process. This chapter builds upon the analysis completed by McKay Price et al. (2012) and disagreement literature (Karpoff, 1986; Harris and Raviv, 1993; Antweiler and Frank, 2004; Banerjee and Kremer, 2010; Atmaz and Basak, 2016), by looking particularly at various differences in sentiment produced by participants on these calls and how these disagreements in sentiment are digested by the wider market.

Table 6.4 provides conditional regression results for each of the four divergence of sentiment measures leveraged within this study and both CAV periods. In this analysis the four DoS measures relate to the difference in manager versus analyst sentiment over the full call (Overall DoS), the difference in manager versus analyst sentiment in the Question-and-Answer section only (Q&A DoS), the difference between MD and Q&A sentiment (MD Vs Q&A DoS) and the difference between individual analyst sentiment (Analyst DoS). These four measures of DoS allow us to identify differences in opinion of participants on these calls and understand how the various measures of disagreement impact firm level abnormal trading volume.

Table 6.4: Estimation of the Association between DOS and Log CAVs for Multimodal Sentiment

Panel A: Initial Cumulative Abnormal Volume

Divergence of Sentiment	Overall Sentiment		Q&A Sentiment		Intro Vs Q&A		Analyst	
	High Manager	High Analyst	High Manager	High Analyst	High Intro	High Q&A	High Divergence	Low Divergence
Const	2.3546*** (0.0000)	2.1802*** (0.0000)	2.3716*** (0.0000)	2.1558*** (0.0000)	1.9841** (0.0354)	1.8937*** (0.0036)	1.855*** (0.0000)	2.343*** (0.0000)
Sentiment	0.1191*** (0.0000)	0.0664** (0.0278)	0.1136*** (0.0000)	0.0597** (0.0411)	-0.0524 (0.751)	0.0412 (0.6255)	0.0643** (0.0163)	-0.0106 (0.5611)
Earnings Surprise	-0.5456 (0.2254)	0.3796 (0.4283)	-0.7254* (0.0968)	0.2982 (0.5382)	-5.52 (0.1419)	-2.3839 (0.1429)	-0.1229 (0.7965)	0.2575 (0.6051)
Management Discussion	-0.0023 (0.3204)	0.0077* (0.0764)	-0.0027 (0.2321)	0.0061** (0.0413)	0.0007 (0.9395)	-0.0084* (0.0942)	-0.0014 (0.5677)	0.003 (0.2236)
Question and Answer	0.0007 (0.2616)	0.0000 (0.9699)	0.0008 (0.2013)	0.0003 (0.67999)	-0.0033 (0.3065)	-0.0013 (0.5146)	0.0004 (0.5518)	0.0008 (0.1006)
Market Capitalisation	-0.0797*** (0.0000)	-0.0534*** (0.0003)	-0.0825*** (0.0000)	-0.0543*** (0.0002)	-0.008 (0.9068)	-0.0037 (0.9435)	-0.0224* (0.0883)	-0.0731*** (0.0000)

Book-to-Market	-0.0783** (0.0114)	-0.1302*** (0.0072)	-0.0837*** (0.0064)	-0.1269*** (0.0075)	-0.0336 (0.4397)	-0.3652 (0.107)	-0.1433*** (0.002)	-0.0548* (0.0588)
Profitability	0.0011 (0.632)	-0.0009 (0.6441)	0.0017 (0.4572)	-0.0006 (0.7471)	-0.0104 (0.3251)	0.0113 (0.182)	-0.0019 (0.4020)	0.0003 (0.8808)
Leverage	-5.2365 (0.4521)	-4.6007 (0.5182)	-4.2463 (0.5346)	-5.6711 (0.4208)	20.8811 (0.5502)	-6.2114 (0.8122)	-7.6741 (0.2878)	-8.2467 (0.2160)
Volatility	-0.2269 (0.3602)	-0.5884* (0.0514)	-0.1759 (0.4692)	-0.7073** (0.0203)	-0.5532 (0.7072)	0.7543 (0.3871)	-0.5846** (0.0268)	0.1138 (0.6653)
N. Observations	657	613	667	624	51	65	762	736
R-sq	0.0956	0.0540	0.1020	0.0582	0.1180	0.2046	0.0382	0.0626
Adj R-sq	0.0830	0.0399	0.0897	0.0444	-0.0756	0.0744	0.0267	0.0510

Panel B: Extended Cumulative Abnormal Volume

Divergence of Sentiment	Overall Sentiment		Q&A Sentiment		Intro Vs Q&A		Analyst	
	High Manager	High Analyst	High Manager	High Analyst	High Intro	High Q&A	High Divergence	Low Divergence
Const	3.969*** (0.0000)	4.0871*** (0.0000)	3.9689*** (0.0000)	4.0443*** (0.0000)	2.8529*** (0.0000)	3.9845*** (0.0000)	4.0385*** (0.0000)	4.0675*** (0.0000)
Sentiment	0.0078 (0.7108)	-0.0388* (0.099)	0.0082 (0.6912)	-0.039* (0.0799)	0.0086 (0.932)	0.0293 (0.6173)	-0.015 (0.4025)	-0.0306** (0.0425)
Earnings Surprise	-0.0088 (0.9794)	0.1302 (0.7275)	-0.0978 (0.7699)	0.0315 (0.9319)	-2.0093 (0.376)	-0.563 (0.6153)	0.0118 (0.9706)	-0.1853 (0.6522)
Management Discussion	-0.0011 (0.535)	0.0034 (0.3083)	-0.0011 (0.5329)	0.0007 (0.763)	0.0069 (0.1922)	-0.002 (0.5716)	-0.0003 (0.8718)	0.0031 (0.1257)
Question and Answer	0.0003 (0.5496)	0.0000 (0.9646)	0.0003 (0.5663)	0.0004 (0.385)	0.0011 (0.5691)	-0.0012 (0.4139)	0.0002 (0.7513)	0.0003 (0.4374)
Market Capitalisation	-0.0028 (0.7771)	0.0027 (0.8149)	-0.0037 (0.7074)	0.0009 (0.9385)	0.0795* (0.0604)	0.0365 (0.3204)	0.0008 (0.9264)	-0.0086 (0.4052)
Book-to-Market	0.0405* (0.085)	-0.0508 (0.1774)	0.0421* (0.0729)	-0.0439 (0.2239)	0.1274*** (0.0000)	-0.1364 (0.3818)	-0.0397 (0.2007)	-0.0114 (0.6341)
Profitability	-0.0003 (0.8741)	-0.0025 (0.1073)	0.0001 (0.9368)	-0.0023 (0.1116)	-0.0043 (0.4997)	-0.0133** (0.0265)	-0.0008 (0.5939)	-0.0017 (0.2457)
Leverage	3.6187 (0.4945)	-5.7318 (0.3022)	5.1348 (0.3275)	-5.6957 (0.2886)	12.2867 (0.5637)	-10.9379 (0.5474)	-1.7041 (0.7241)	8.3128 (0.1310)
Volatility	0.0783 (0.678)	-0.1907 (0.4177)	0.103 (0.5801)	-0.484** (0.0371)	-0.687 (0.4449)	1.7335*** (0.0056)	-0.4692*** (0.0079)	-0.3873* (0.0749)
N. Observations	657	613	667	624	51	65	762	736
R-sq	0.0077	0.0130	0.0086	0.0201	0.4566	0.2246	0.0138	0.0293
Adj R-sq	-0.0061	-0.0017	-0.0050	0.0057	0.3373	0.0977	0.0020	0.0173

*Notes: This table provides conditional regression results of log CAVs on earnings call divergence of sentiment measures and control variable measures for the multimodal model. CAV (-1,1) is the short-term cumulative abnormal volume across three days where the earnings call event is denoted as day 0 - abnormal volume is defined in section 6.4.1. CAV (2,60) is then the longer period summation of cumulative abnormal volume from 2 days after an earnings call to 60 days after an earnings call. In the first four columns the divergence of sentiment measures identifies calls that have the greatest difference in managerial versus analyst sentiment across the full call and within the question-and-answer section of earnings calls. The high manager column relates to calls that have a high divergence of sentiment where managers are significantly more optimistic than analysts, whilst the high analyst columns are calls where analysts are more optimistic than managers. The fifth and sixth columns identify calls where there is a significant difference between earnings call MD sentiment and Q&A sentiment. The high MD column relates to a subset of calls that produce a significantly more optimistic introductory sentiment compared to the Q&A section sentiment, whilst the high Q&A column identifies calls that produce a more optimistic Q&A sentiment compared to the MD sentiment. The last two columns relate to differences in analyst sentiment on earnings calls. The high divergence call captures a subset of calls that have a high degree of difference in average analyst sentiment on these calls, whilst the low divergence subset identifies calls where analyst sentiment agrees. N. Observations highlights the number of calls included within each subset's regression. R^2 determines the proportion of variance in the dependent variable that can be explained by the independent variable. Adj R^2 is a corrected goodness of fit measure of R^2. Regression p-values are in parenthesis. Significance level indicators: * at 10%, ** at 5%, *** at 1%.*

The first divergence of sentiment (DoS) measure analysed within this analysis looks at the contrast between manager and analyst sentiment concerning the overall call. Specifically, identifying calls with the highest differences in average managerial sentiment compared to average analyst sentiment. Panel A of Table 6.4 shows the explainability each of the DoS measures has for short-term CAV. The findings indicate that the subset of calls exhibiting significantly higher manager sentiment compared to analyst sentiment serves as a significant predictor of short-term CAV. Specifically, a one percent increase in earnings call multimodal sentiment on this particular subset of calls, CAV increases by roughly 0.12%. The obtained p-value of 0.0000 suggests a highly significant result, with a positive coefficient implying that an increase in sentiment on the calls characterised by high manager sentiment and low analyst sentiment leads to an initial heightened reaction in abnormal trading volume.

This relationship is also seen in the reciprocal set of calls where there is again a significant difference in overall call sentiment, however in this case where analysts express more positivity in comparison to managers. This result retains significance at the 5% level again with a positive coefficient (albeit smaller than the high manager DoS subset regression), suggesting that calls featuring substantial differences in average participant sentiment explain initial movements in abnormal volume. This result is intensified for calls eliciting substantially higher levels of managerial sentiment as they are shown to be followed by an immediate uptick in abnormal volume of greater magnitude in comparison to calls that display higher average levels of analyst sentiment. These results return larger coefficients in comparison to the most robust result (AF2) in relation to the main regressions for short-term CAV. However, the DoS high manager subset for the overall call is the only subset to produce the same level of statistical significance as the main results. Hence implying that calls where managers are significantly more optimistic than their call counterparts generate a greater market response in short-term CAV than calls in general.

The next DoS measure that was evaluated looked again at disagreement between managers and analysts but particularly within the Q&A section of earnings calls. Similar to the results for the overall call DoS measure, it is found that both sets of Q&A DoS (positive manager sentiment vs negative analyst sentiment and negative manager sentiment vs positive analyst sentiment) have positive and statistically significant relationships with short-term CAV. In these conditional regressions we again see that the DoS subset which relates to calls where managers speak on average with significantly greater optimism than their call counterparts generate a greater market reaction (higher coefficient) and stronger relationship (stronger p-value) than the reciprocal set. Comparing the DoS measure for the overall call and the DoS measure that focuses particularly on the Q&A set we see that there is not much difference in the results however the Q&A set returns a stronger R^2 value. Therefore, implying that when looking to identify disagreement on earnings calls focusing on differences in sentiment in the natural language portion of the call will allow for marginally stronger prediction abilities for following movements in short-term abnormal trading volume. Comparing the Q&A DoS results to the main results, we find a similar narrative as the overall DoS results, that is the DoS subset pertaining to high manager sentiment returns stronger statistical significance and a larger positive coefficient for short-term CAV.[273]

Looking at the relationship both overall and Q&A DoS measures have with longer period CAV, the subset of calls that produce more optimistic analyst sentiment with more pessimistic managerial sentiment are the only sets to return a significant relationship. Both the overall call DoS measure and the Q&A DoS measure that exhibit higher levels of analyst sentiment in comparison to managers, return a negative relationship with longer period CAV at the 10% level. Previous literature has identified general differences in the way in which managers and analysts speak on these calls (Brockman, Li and McKay Price, 2015; Chen, Nagar and Schoenfield, 2018). Managers are found to speak with significantly greater optimism, at greater length and use less complex language in comparison to their call counterparts. Interestingly, when analysts speak more optimistically on earnings calls in comparison managers the level of sentiment on these calls has greater forecasting power for abnormal trading volume over extended periods. These results return a stronger negative coefficient in comparison to the main results which would entail a stronger market reaction in CAV, however the results do not return as strong in terms of statistical significance.

The MD vs Q&A DoS measure evaluated within this analysis falls in line with the abnormal sentiment measure leveraged by Blau, DeLisle and McKay Price (2015). The authors develop a measure of inflated talk by taking the difference between MD sentiment and Q&A sentiment, to identify whether short-sellers incorporate soft information into their decision making process.[274] Chen, Nagar and Schoenfield (2018) show that earnings calls' typically begin overly optimistic due to the managerial discussion introduction and begins to move towards a sentiment which fits with the firm's performance of the previous quarter. This analysis looks at the divergence between average introductory statement sentiment versus average Q&A sentiment to assess whether calls which display a large difference in sentiment create disagreement within the market and consequently induce trading. From Table 6.4 it can be seen that there is no significant relationship between said DoS measure and short or longer period CAV.

The final divergence of sentiment indicator that was analysed was the difference between individual analyst sentiment produced on these calls. To calculate this divergence measure, the average individual level of sentiment produced by each analyst on each of our 4,928 earnings calls was identified and we selected calls which had the highest standard deviations of average analyst sentiment.[275] Two

[273] Brockman, Li and McKay Price (2015) and Chen, Nagar and Schoenfeld (2018) both find that the market responses significantly stronger in reaction to analyst sentiment compared to managerial sentiment. However, they identify this response using abnormal returns whether as this analysis evaluates market reactions in terms of abnormal trading volume.

[274] Kartik et al. (2007) identify that inflated talk should be considered bad news – hence the authors are evaluating whether short sellers are sophisticated enough to process inflated talk information in their forecasts as bad news.

[275] The subset of analyst disagreement calls was identified as the calls that fell outside the mean plus one standard deviation of the distribution of analysts' standard deviations on these calls. The analyst agreement subset was identified as calls that fell below the mean minus one standard deviation.

subsets were then created an evaluated in relation to short and longer period CAV – a high analyst divergence set, and a low analyst divergence set. These two sets act as proxies for analyst disagreement and analyst agreement, respectively. Table 6.4 indicates that calls which have high levels divergence amongst analyst participants are significantly and positively related to short-term CAV. In other words, as the average sentiment on these calls increase, where there is an evident disagreement in analyst tone, the market responds with increased abnormal trading volume. This relationship with short-term CAV is not found on calls where there are high levels of analyst agreement. However, when evaluating both the analyst disagreement and agreement measures, only calls that produce analyst agreement have a significant relationship with longer period CAV. Specifically, calls that produce high levels of analyst agreement have a significant negative relationship with longer period CAV. That is, as the average sentiment on calls that produce high levels of analyst agreement increases, abnormal trading volume decreases over a longer horizon.

Continuing the investigation into the interaction between earnings calls sentiment and CAV on calls that produce significant differences of opinion Table 6.5 evaluates the same subsets of calls as Table 6.4 but uses a different regression technique. Comparing both sets of analysis, results in each do not complement one another but neither do they contradict. The only result which returns significance in both regressions is in relation to the low analyst divergence subset of earnings calls for long-term CAV. Both results indicate that as sentiment increases on calls which have a high level of analyst agreement, long-term abnormal volume decreases. Furthermore, the subset of calls relating to analyst agreement also returns a highly significant positive result for short-term CAV i.e., as the sentiment on calls that produce high levels of analyst agreement increases, the reaction to abnormal volume post earnings call is increased (even more so than general calls).

As illustrated in Table 6.5 the only other significant results stemming from the dummy variable analysis are in relation to longer period CAV. These significant results come from the subsets of calls that are identified as having significantly higher manager sentiment compared to analyst sentiment on both overall calls and only the Q&A portion of the call. For the subset of calls where managers are significantly more optimistic than analysts across the full length of the call the findings indicate that as sentiment on these calls increase, longer-term CAV decreases. Alternatively, for the subset of calls concerned purely with managers that are more optimistic than analysts in the Q&A section of the call the reciprocal is true – as the sentiment on these calls increase CAV also increases over the longer term.

Table 6.5: Estimation of the Association between DOS and Log CAVs for Multimodal Sentiment

Panel A: Initial Cumulative Abnormal Volume								
	Overall Sentiment		Q&A Sentiment		Intro Vs Q&A		Analyst Divergence	
Divergence of Sentiment	High Manager	High Analyst	High Manager	High Analyst	High Intro	High Q&A	High Divergence	Low Divergence
	Coefficient	**Coefficient**	**Coefficient**	**Coefficient**	**Coefficient**	**Coefficient**	**Coefficient**	**Coefficient**
const	2.1704*** (0.0000)	2.1699*** (0.0000)	2.1696*** (0.0000)	2.1695*** (0.0000)	2.1669*** (0.0000)	2.1652*** (0.0000)	2.1595*** (0.0000)	2.1709*** (0.0000)

Sentiment	***0.0301****** ***(0.0011)***	***0.0342***.*** ***(0.0002)***	***0.0304****** ***(0.0010)***	***0.0345****** ***(0.0002)***	***0.0377****** ***(0.0000)***	***0.037****** ***(0.0000)***	***0.0327****** ***(0.0005)***	***0.0532****** ***(0.0000)***
Dummy Variable	-0.0276 (0.2018)	-0.0321 (0.1261)	-0.021 (0.3257)	-0.0304 (0.1440)	0.0255 (0.6234)	0.0739 (0.2183)	-0.0021 (0.9128)	0.0522*** (0.0026)
Interaction	0.0722** (0.0107)	0.0294 (0.3257)	0.0662 (0.0178)	0.0246 (0.4027)	-0.0837 (0.4861)	0.0026 (0.9714)	0.0355 (0.1648)	-0.0693*** (0.0004)
Earnings Surprise	-0.1002 (0.5817)	-0.0914 (0.6156)	-0.1005 (0.5809)	-0.0903 (0.6198)	-0.0946 (0.6033)	-0.0899 (0.6215)	-0.0942 (0.6050)	-0.0759 (0.6765)
Management Discussion	0.0003 (0.3162)	0.0003 (0.3795)	0.0003 (0.3083)	0.0003 (0.3772)	0.0003 (0.3469)	0.0003 (0.3172)	0.0003 (0.2699)	0.0003 (0.3520)
Question and Answer	0.0003 (0.1850)	0.0003 (0.2115)	0.0003 (0.1837)	0.0003 (0.21000)	0.0003 (0.2229)	0.0003 (0.1915)	0.0003 (0.1620)	0.0002 (0.2998)
Market Capitalisation	-0.053*** (0.0000)	-0.0525*** (0.0000)	-0.053*** (0.0000)	-0.0525*** (0.0000)	-0.0526*** (0.0000)	-0.0528*** (0.0000)	-0.0523*** (0.0000)	-0.0533*** (0.0000)
Book-to-Market	-0.1013*** (0.0000)	-0.1017*** (0.0000)	-0.1013*** (0.0000)	-0.1018*** (0.0000)	-0.1021*** (0.0000)	-0.1009*** (0.0000)	-0.1013*** (0.0000)	-0.1032*** (0.0000)
Profitability	-0.0002 (0.7878)	-0.0002 (0.7478)	-0.0002 (0.7888)	-0.0002 (0.7495)	-0.0002 (0.7546)	-0.0002 (0.7806)	-0.0003 (0.6987)	-0.0002 (0.7597)
Leverage	-7.2969*** (0.0054)	-7.2862*** (0.0055)	-7.3001 (0.0054)	-7.2821*** (0.0055)	-7.2876*** (0.0055)	-7.1326*** (0.0066)	-7.4258*** (0.0047)	-7.5198*** (0.0041)
Volatility	-0.126 (0.2123)	-0.1276 (0.2066)	-0.1249 (0.2161)	-0.1277 (0.2063)	-0.1248 (0.2168)	-0.1263 (0.2115)	-0.1221 (0.2270)	-0.125 (0.2154)
N. Observations	4928	4928	4928	4928	4928	4928	4928	4928
R-sq	0.0425	0.0416	0.0424	0.0416	0.0413	0.0417	0.0418	0.0439
Adj R-sq	0.0404	0.0395	0.0403	0.0394	0.0392	0.0396	0.0396	0.0418

Panel B: Extended Cumulative Abnormal Volume

	Overall Sentiment		Q&A Sentiment		Intro Vs Q&A		Analyst Divergence	
Divergence of Sentiment	High Manager	High Analyst	High Manager	High Analyst	High Intro	High Q&A	High Divergence	Low Divergence
	Coefficient	**Coefficient**	**Coefficient**	**Coefficient**	**Coefficient**	**Coefficient**	**Coefficient**	**Coefficient**
const	4.0628*** (0.0000)	4.0588*** (0.0000)	4.0625*** (0.0000)	4.0587*** (0.0000)	4.0595*** (0.0000)	4.0594*** (0.0000)	4.0602*** (0.0000)	4.0614*** (0.0000)
Sentiment	***-0.0303****** ***(0.0000)***	***-0.0275****** ***(0.0001)***	***-0.0304****** ***(0.0000)***	***-0.0271****** ***(0.0001)***	***-0.0281****** ***(0.0000)***	***-0.0285****** ***(0.0000)***	***-0.0298****** ***(0.0000)***	***-0.0237****** ***(0.0014)***
Dummy Variable	-0.0305* (0.0592)	-0.0034 (0.8301)	-0.0276* (0.0850)	-0.0033 (0.8342)	-0.0395 (0.3085)	-0.0109 (0.8088)	-0.0095 (0.5110)	0.0225* (0.0833)
Interaction	0.0288 (0.1738)	-0.0042 (0.8510)	0.0285 (0.1734)	-0.0086 (0.6965)	-0.028 (0.7552)	0.0595 (0.2744)	0.0172 (0.3673)	-0.0156 (0.2902)
Earnings Surprise	0.0541 (0.6913)	0.0533 (0.6958)	0.0535 (0.6946)	0.0526 (0.6997)	0.053 (0.6970)	0.0554 (0.6845)	0.0533 (0.6958)	0.0607 (0.6560)
Management Discussion	0.0008*** (0.0006)	0.0008*** (0.0004)	0.0008*** (0.0006)	0.0008*** (0.0004)	0.0008*** (0.0005)	0.0008*** (0.0004)	0.0008*** (0.0004)	0.0008*** (0.0004)
Question and Answer	-0.0002 (0.2170)	-0.0002 (0.2374)	-0.0002 (0.2205)	-0.0002 (0.2386)	-0.0002 (0.2110)	-0.0002 (0.2475)	-0.0002 (0.2473)	-0.0002 (0.1817)
Market Capitalisation	0.0003 (0.9380)	0.0002 (0.9538)	0.0003 (0.9439)	0.0002 (0.9542)	0.0003 (0.9321)	0.0001 (0.9853)	0.0002 (0.9635)	-0.0001 (0.9806)
Book-to-Market	-0.0221** (0.0488)	-0.0222** (0.0478)	-0.0221** (0.0485)	-0.0222** (0.0477)	-0.0219* (0.0508)	-0.0221** (0.0483)	-0.022** (0.0494)	-0.0223** (0.0463)
Profitability	-0.0014** (0.0141)	-0.0014** (0.0140)	-0.0014** (0.0142)	-0.0014** (0.0144)	-0.0014** (0.0142)	-0.0014** (0.0143)	-0.0014** (0.0130)	-0.0014** (0.0146)
Leverage	3.6797* (0.0608)	3.7938* (0.0535)	3.6827* (0.0607)	3.816* (0.0521)	3.7477* (0.0563)	3.8034* (0.0527)	3.7125* (0.0588)	3.6835* (0.0606)
Volatility	-0.2109*** (0.0053)	-0.2105*** (0.0054)	-0.2103*** (0.0054)	-0.2109*** (0.0053)	-0.209*** (0.0057)	-0.2089*** (0.0058)	-0.21*** (0.0055)	-0.2112*** (0.0052)
N. Observations	4928	4928	4928	4928	4928	4928	4928	4928
R-sq	0.0205	0.0198	0.0204	0.0199	0.0200	0.0201	0.0200	0.0204
Adj R-sq	0.0183	0.0177	0.0182	0.0177	0.0178	0.0179	0.0178	0.0182

*Notes: This table provides regression (containing a dummy and interaction variable) results of log CAVs on earnings call divergence of sentiment measures and control variable measures for the multimodal model. CAV (-1,1) is the short-term cumulative abnormal volume across three days where the earnings call event is denoted as day 0 - abnormal volume is defined in section 6.4.1. CAV (2,60) is then the longer period summation of cumulative abnormal volume from 2 days after an earnings call to 60 days after an earnings call. In the first four columns the divergence of sentiment measures identifies calls that have the greatest difference in managerial versus analyst sentiment across the full call and within the question-and-answer section of earnings calls. The high manager column relates to calls that have a high divergence of sentiment where managers are significantly more optimistic than analysts, whilst the high analyst columns are calls where analysts are more optimistic than managers. The fifth and sixth columns identify calls where there is a significant difference between earnings call MD sentiment and Q&A sentiment. The high MD column relates to a subset of calls that produce a significantly more optimistic introductory sentiment compared to the Q&A section sentiment, whilst the high Q&A column identifies calls that produce a more optimistic Q&A sentiment compared to the MD sentiment. The last two columns relate to differences in analyst sentiment on earnings calls. The high divergence call captures a subset of calls that have a high degree of difference in average analyst sentiment on these calls, whilst the low divergence subset identifies calls where analyst sentiment agrees. N. Observations highlights the number of calls included within each subset's regression. R^2 determines the proportion of variance in the dependent variable that can be explained by the independent variable. Adj R^2 is a corrected goodness of fit measure of R^2. Regression p-values are in parenthesis. Significance level indicators: * at 10%, ** at 5%, *** at 1%.*

6.6 Conclusion

A vast amount of research has identified that psychology plays a significant role in financial decision making and impacts asset pricing (Ackert, Church and Deaves, 2003). Traditional theories that explain movements in financial market characteristics assumed that market participants are purely rational and make decisions based on fundamental information using the laws of probability (Fama, 1970). The traditional theories, that based their models upon rational agents, leave no room for the emotional feelings of financial market participants even though the psychology literature has shown the significant effects that emotional states can have on decision making (Erickson et al. 1978; Conley, Lind and O'Barr, 1978; Apple et al. 1979; Wallbott, 1982; Brooke and Ng, 1986; Bradac, Mulac and House, 1988; Chattopadhyay et al. 2003; Feinberg et al. 2005; Klofstad, Anderson and Peters, 2012; Martín-Santana et al. 2015; Wang et al. 2018; Chua et al. 2020; Song et al. 2020). This chapter builds upon previous literature that evaluates the impact that financial sentiment has on financial market characteristics. In particular, it leverages a multimodal sentiment classifier that incorporates behavioural cues extracted from the audio aspect of earnings conference calls, along with the textual content produced on these calls, to further evaluate the impact that sentiment has on Cumulative Abnormal Trading Volume (CAV). The inclusion of these paralinguistic features is based on the psychology literature which identifies the way in which information is communicated conveys substantial information beyond the content of such communication (Guyer, Fabrigar and Vaughan-Johnston, 2018). Thus, this analysis evaluates the relationship between earnings call sentiment and CAV from a deeper behavioural standpoint.

The present study conducts an in-depth analysis into the impact absolute values of multimodal sentiment and measures of participant divergence of said sentiment on earnings conference calls has on both short- and longer-term CAV. Controlling for general movements in the overall market and specifically evaluating the fluctuations in abnormal trading volume due to the release of information within earnings calls using multimodal sentiment, this chapter has four key findings. Firstly, this paper looks at the relationship between multimodal sentiment and longer period CAV which is relatively understudied in the domain. The findings indicate that multimodal sentiment has a negative relationship with longer period CAV and produces stronger results than the only previous paper which similarly evaluates earnings call sentiment and longer period CAV (McKay Price et al. 2012). From the findings in relation to absolute values of sentiment and both measures of CAV, this chapter identifies that the inclusion of behavioural cues coupled with commonly used textual information creates a sentiment measure that more accurately captures investors' reaction in abnormal trading volume to information produced on earnings calls. These additional behavioural cues extract insights from information produced on earnings calls, that previously have been underutilised, and highlight their importance in forecasting longer-term trading volume. These results further show the significant role paralinguistic information plays in forging investors' expectations surrounding a firm's future performance.

The three following key findings emerging from this analysis are in relation to the work conducted using divergence of sentiment measures. The results indicate that the DoS measures that identify calls with significant differences in managerial and analyst sentiment over the full earnings call are able to explain movements in short-term CAV. Evaluating two subsets of calls, first where managers are significantly more positive than analysts and secondly the reciprocal of this set where analysts are notably more optimistic than managers. The evidence suggests that both are associated with short-term CAV, however the results indicate that calls where managers are significantly more positive than analysts, the market generates a heightened reaction in terms of CAV. Furthermore, focusing on the difference in sentiment across the Q&A section of the call, the are in line with the findings established for the overall call. Both sets of earnings calls which exhibit DoS in manager vs analyst sentiment produce a significant positive relationship with short-term CAV, however the subset where managers speak with more positive sentiment generates a greater market reaction in trading volume. Additionally, these subsets identifying managerial optimism return stronger statistical significance and larger positive coefficients than the main results. These results, taken together, imply that when there is a disagreement in sentiment between the two sets of participants on the call, particularly when managers are more optimistic, the market responds with increases levels of abnormal trading volume compared to the average reaction in the short-term.

For both DoS measures comparing managerial and analyst sentiment, looking at the full earnings call and the Q&A section, the results indicate that calls producing DoS with high analyst sentiment significantly predicts longer period CAVs when evaluating the conditional regression. This relationship is found to be negative for the overall call DoS measure and the Q&A section DoS measure. Hence, on calls where analysts are significantly more optimistic, there is a reduction in abnormal trading volume over the longer term. This could imply that the market credits analysts as an expert source and take their level of sentiment as a robust predictor of fundamental information, hence generating a consensus in trading signals which initially increase CAV but subsequently dampen over the longer term as prices quickly move to fundamental values. However, when running the results using a dummy variable regression these results become insignificant. Table 6.5 produces results which indicate that calls which contain a high level of optimism in managerial sentiment in comparison to analyst sentiment return a negative significant relationship with longer period CAV. The results from both of these regressions return differing results for extended period CAV. Although Table 6.4 and Table 6.5 produce different conclusion in relation to which set of participants optimistic sentiment generates a significant association with trading volume over a longer horizon, both indicate that as sentiment increases on calls which display disagreement in sentiment, trading reduces over the longer period. This may entail that disagreement among participants on these calls teases out new fundamental information which gives market participants as a whole a better understanding of future firm performance allowing for better decisions to be made initially, moving prices to fundamental values with a dampening of trading over the longer horizon as prices reside at a correct level.

Finally, looking directly at the subsets of calls which display divergence and agreement between analyst participants, calls which evidence analyst disagreement generate an increase in short-term CAV however calls that have analyst agreement do not. Alternatively, calls producing analyst agreement show increased CAV over the longer period whilst calls producing analyst disagreement do not have a significant relationship with longer period CAV. Moreover, the subset relating to analyst agreement significantly predicts longer term CAV. This falls in line with the previous findings that analyst sentiment is considered an expert source and if all analysts sentiment on the call is in agreement, then there is less trading volume over the longer period as market prices quickly move to fundamentals.

When considering these findings collectively it can be said that the addition of behavioural cues to a sentiment classification model increases forecasting abilities of trading volume. Additionally, the use of the multimodal classifier highlights the weight that sentiment of both sets of call participants carries in relation to short-term and longer-term CAV. Particularly, optimistic managerial sentiment is found to have a stronger relationship with short-term CAV whilst optimistic analyst sentiment has a more robust association with abnormal trading volume over the longer-term. Specifically, calls where there is DoS among participants evoke a larger market reaction in comparison to general reactions, evidenced through larger coefficients than the main findings, in short-term CAV.

7. Conclusion

7.1 Introduction

In recent years the application of sentiment analysis in the financial domain has seen a substantial amount of interest. Researchers have used sentiment analysis techniques to understand the complexities of qualitative information and its impact on financial markets across various forms of qualitative data (Soleymani et al. 2017). A key theme identified through the examination of prior related literature concerns the substantial lack of advanced methods used to define sentiment in academic finance. However, those studies that have been conducted across the finance, NLP and psychology domains demonstrate that the application of more advanced methods can result in more robust measures of sentiment. These improved measures potentially allow for a deeper understanding of the nuances in natural language communication to be developed. Specifically, there are two sentiment classification techniques that have shown demonstrably greater abilities in understanding natural language across a number of academic domains: transformer architecture (Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Alamoudi and Alghamdi, 2021) and multimodal analysis (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021).

This thesis addresses the lack of advanced methods that have been applied to financial disclosures by creating a multimodal sentiment classifier that leverages textual and audio information, in order to quantify a more robust measure of sentiment for the financial information disclosure earnings conference calls. In doing so, this research sheds light on the extent to which the subtleties of language disseminated by earnings call participants' impact upon wider investor decision making. Leveraging these methods offers the opportunity to make key technical contributions, outlined in Section 7.1.2, to the finance domain. However, the primary reason these techniques are leveraged is to create a model that more adeptly understands the nuances of financial communication, thus allowing for a deeper investigation into market behaviour surrounding earnings conference calls. Specifically, the incorporation of the audio modality of earnings calls offers the opportunity to assess whether the responses to paralinguistic cues that have been demonstrated in the psychology literature (Erickson et al. 1978; Conley, O'Barr and Lind, 1978; Brooke and Ng, 1986; Bradac, Mulac and House, 1988; Chattopadhyay et al. 2003; Feinberg et al. 2005; Klofstad, Anderson and Peters, 2012; Martín-Santana et al. 2015; Wang et al. 2018; Chau et al. 2020; Song et al. 2020) are evident in financial decision-making and consequently impacts upon financial markets.276 Thus, the results from this thesis are embedded in behavioural theory.

Building upon the investigation of previous literature in Chapter 2, which identifies a lack of advanced NLP methods being used in the financial sentiment analysis area, Chapter 3 of this thesis outlines the steps taken to create the multimodal sentiment classifier used throughout this thesis and the accompanying multimodal S&P 100 dataset the empirical investigations are conducted using. This thesis addresses a critical gap in finance by introducing state-of-the-art methods for defining sentiment. A key metric for evaluating whether these advanced techniques enhance comprehension of financial context is to compare the accuracy of such models against traditional models that are used in the prior literature. Chapter 2 highlights that most prior research on earnings call sentiment focuses on U.S.-listed companies. The S&P 100, being among the largest and most dynamic global markets, presents an ideal setting for this thesis as it allows for the results stemming from this research to be compared to similar research using more rudimentary sentiment classification methods. The empirical chapters (Chapters 4,

[276] The impact differing levels of paralinguistic cues have on speaker perceptions and subsequently the decision-making process surrounding the information communicated is discussed in Chapter 1.4 and Chapter 4.2.2.

5 and 6) then apply the multimodal sentiment classifier to the S&P 100 dataset to understand how market participants react and use the information portrayed on these calls.

Firstly, in Chapter 4 a comparative study is conducted to analyse the performance of the multimodal sentiment classifier against the most used sentiment classification techniques in previous literature. Specifically, Chapter 4 compares the multimodal sentiment analysis framework, defined in Chapter 3, that integrates FinBERT with paralinguistic features and a deep learning classifier, with benchmark models including general and specific dictionary-based approaches, traditional machine learning models, and recent BERT-based techniques. This allows for classification of the classifier which has the best understanding of the language used on these disclosures.

Chapter 5 then investigates the impact that sentiment has on market behaviour in terms of cumulative abnormal returns (CARs) (Antweiler and Frank, 2004; Henry, 2006:2008; Tetlock, Saar-Tsechansky and Macskassy, 2008; Li, 2010; Grob-Klubmann and Hautsch, 2010; Loughran and McDonald, 2011; Bollen, Mao and Zeng, 2011; Mao and Bollen, 2011; Davis and Tama-Sweet, 2012; Jegadeesh and Wu, 2012; Twedt and Rees, 2012; Garcia, 2013; Sprenger et al. 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Ferguson et al. 2015; Sun, Najand and Shen, 2016; Siganos, Vagenas-Nanos and Verwijmeren, 2017; Jiang et al. 2019; Audrino and Tetereva, 2019; Gu and Kurov, 2020). This is a natural continuation from the results found in Chapter 4 which indicate that the multimodal classifier has greater capabilities in classifying earnings call sentiment in comparison to the most used methods in prior literature. Although prior studies have evaluated the relationship between sentiment and returns, most studies use basic approaches to classify sentiment which do not have as robust of an understanding of financial context compared to the multimodal classifier introduced by this thesis. The analysis of Chapter 5 then extends to focus on market behaviours in accordance with the way in which information is portrayed on earnings calls. Particularly, due to this thesis taking into consideration the paralinguistic modality of earnings calls, Chapter 5 evaluates whether differences in the way in which information is communicated on these calls has any impact on the decision-making process of investors at large.

Finally, Chapter 6 conducts the same investigation but looks to evaluate market behaviour from the lens of cumulative abnormal trade volumes (CAVs) instead of CARs. While previous studies have explored the link between financial sentiment and trading volume (Antweiler and Frank, 2004; Tetlock, 2007; McKay Price et al. 2012; Garcia, 2013; Sprenger et al. 2013; Siganos et al. 2014, 2017), this relationship remains less examined than the well-studied connection between sentiment and returns. Similar to prior academic studies conducted on the relationship between sentiment and CARs, basic approaches are widely used to define sentiment for studies analysing the relationship between sentiment and CAV. With the results in Chapter 4 and 5 indicating the multimodal sentiment classifier used in this thesis has a deeper understanding of financial context, evaluating the sentiment CAV dynamic with this model is a logical progression to deepen the areas knowledge of this relationship. Building on the initial evaluation of average levels of multimodal sentiments impact on CAV, Chapter 6 extends its analysis by evaluating the impact disagreement on earnings calls, proxied through a divergence of multimodal sentiment measure like that used by Siganos et al. (2014, 2017), has on the behaviour of CAV.

The contributions to existing literature made throughout this thesis are outlined in the following section 7.2. Then the limitations associated with the research conducted are discussed in section 7.3 with 7.4 highlighting directions for future research.

7.2 Contributions to Existing Literature

This thesis offers significant contributions to the literature on the impact of earnings conference calls on market dynamics from two perspectives: technical and theoretical. First, it presents a technical advancement in the field of financial sentiment analysis by introducing a more robust sentiment classifier than those predominantly used in prior studies. Second, it provides a theoretical contribution by employing this more accurate multimodal classifier to deepen the understanding of financial market participants' behaviour in response to information disclosed during earnings conference calls. Each of these technical and theoretical contributions to the literature are comprehensively discussed within subsections 7.2.1 and 7.2.2, respectively.

7.2.1 Technical Contributions

This thesis makes four key technical contributions to the area of financial sentiment analysis: (i) the creation of the largest multimodal earnings conference call dataset used for the purposes of academic study; (ii) the creation of a multimodal sentiment classifier which outperforms, in terms of classification accuracy, the most commonly employed models in the previous literature; (iii) the identification that a multimodal sentiment classifier returns the strongest association with abnormal returns, and; (iv) evidence suggesting that the multimodal sentiment classifier has stronger forecasting abilities of cumulative abnormal trading volumes in comparison to single-modality sentiment measures used in prior related studies. Each of these contributions are expanded upon individually below.

El Haj et al. (2018) note that a potential reason for the lack of advanced textual analysis methods within the area of financial sentiment analysis is the unavailability of manually classified domain-relevant sentiment datasets on which these models can be trained and validated. The first primary contribution to the literature of this thesis is, to the authors knowledge, the utilisation of the largest financial multimodal dataset to date, containing a full repository of aligned paralinguistic features for 4,860 earnings calls, translating into 637,220 sentences. The creation of the multimodal dataset was a crucial element to allow for state-of-the-art methods used within this thesis to be trained and tested appropriately. Furthermore, the inclusion of this multimodal dataset stands as a highly beneficial addition to the behavioural finance domain due to its dual modality characteristics and ability to investigate insights provided on these calls beyond that of the written word. Across various domains outside of finance, multiple studies indicate that the analysis of natural language from a multimodal perspective allows for a more comprehensive measure of sentiment to be captured (Morency, Mihalcea and Doshi, 2011; Houjeij et al. 2012; Wollmer et al. 2013; Bhaskar, Sruthi and Nedungadi, 2014; Poria, Cambria and Gelbukh, 2015; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021). This increased comprehension is due to the additional behaviour characteristics that are conveyed through the audio and visual modalities outside that of the textual modality alone.

The process of aligning calls and generating paralinguistic features is a challenging endeavour, consuming considerable time and requiring significant technical precision. This thesis analyses a significantly larger dataset of 4,860 calls. In comparison, Mayew and Venkatachalam (2012) analysed a dataset of 466 earnings calls, while Chen, Hand and Zhou (2023) examined 848 calls, and Li et al. (2020) utilized a substantially larger dataset comprising 3,443 calls.[277] This thesis employs a unique dataset specifically consisting of earnings call data from S&P 100 companies, which is unprecedented in its scale, allowing for the adoption of state-of-the-art techniques from the NLP domain to be used in this analysis, and providing the basis for analysis of market behaviour in relation to the significantly understudied paralinguistic modality. The ability to apply state-of-the-art methods to financial disclosures, offered by this dataset, allows for an enhanced sentiment indicator to be captured and, in combination with the ability to analyse non-textual information's impact on market behaviours throughout Chapters 5 and 6, creates the platform for the theoretical contributions to the literature to be made.

Secondly, in the comparative study presented in Chapter 4, the multimodal sentiment classifier developed in Chapter 3 demonstrates superior performance in classifying the sentiment of earnings calls, outperforming widely used approaches used in previous research (Antweiler and Frank, 2004; Henry, 2006; Tetlock, Saar-Tsechansky and Macskassy, 2008; Grob-Klubmann and Hautsch, 2010; Li, 2010; Loughran and McDonald, 2011; Bollen, Mao and Zeng, 2011; Mao and Bollen, 2011; Davis and Tama-Sweet, 2012; Jegadeesh and Wu, 2012; Twedt and Rees, 2012; Garcia, 2013; Sprenger et al. 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014:2017; Ferguson et al. 2015; Sun, Najand and Shen, 2016; Audrino and Tetereva, 2019; Jiang et al. 2019; Gu and Kurov, 2020). This framework surpasses dictionary-based models, traditional machine learning techniques, and general-purpose transformer architectures by integrating state-of-the-art NLP models with paralinguistic features—an underexplored area in prior literature (see Chapter 2). The multimodal classifier achieves the highest testing accuracy of 74.88% in the comparative analysis, slightly exceeding the performance of the same model trained solely on textual data (74.64%) and significantly outperforming Loughran and McDonald's (2011) domain-specific dictionary-based method (47.39%).

Notably, the inclusion of paralinguistic features enhances performance in a multimodal context compared to single-modality models, underscoring the value of incorporating multiple modalities in

[277] Li et al. (2020) focusses solely on the initial management discussion section of earnings calls and exclude any paralinguistic information from the Q&A portion. As a result, despite having a call count similar to this study, the number of sentences that include a complete set of paralinguistic features is considerably smaller—394,277 sentences compared to the 637,220 used here.

sentiment analysis. Although the incorporation of paralinguistic information leads to increased classification accuracy, the increase is marginal. Even so, this underscores the conclusions made by Mayew and Venkatalcham (2012) and the wider psychology literature that non-verbal information is incrementally informative in communicating the true sentiment of a message. These findings indicate that the multimodal approach captures more nuanced aspects of communication, demonstrating the benefits of multimodal frameworks. Combined, our findings contribute to a growing body of literature across various domains, which consistently shows that adding modalities beyond singular modality classifiers improves performance (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021). This study extends this evidence by highlighting the efficacy of multimodal sentiment classifiers in a financial context. However, the gains in classification accuracy come with significant costs in terms of computation power, the time required to generate paralinguistic features, and a lack of output explainability when compared to traditional models. Therefore, there exists a trade-off between marginal improvement of sentiment classification with the computation power and time required to generate these paralinguistic features. Creating a more computationally and time efficient method as a result of this trade-off is highlighted in Chapter 7.4 as a future direction resulting from this thesis.

Furthermore, the substantial increase in classification accuracy over Loughran and McDonald’s (2011) dictionary approach (a difference of 32.49%) underscores the significant performance advantages of the proposed multimodal classifier. Applied to the full dataset of 637,220 sentences used in this thesis, this improvement in classification accuracy results in 207,033 additional sentences being correctly classified. This measurable advancement demonstrates the tangible benefits of the multimodal approach in financial sentiment analysis, and stands as a significant contribution given that the finance-specific dictionary introduced by Loughran and McDonald (2011) remains the most commonly used approach to define sentiment in previous related studies.[278] Although there have been various finance-specific dictionaries developed within the domain, such as Henry (2006), Loughran and McDonald (2011) stands as a seminal paper in the area of financial sentiment analysis and provides a robust reference point to frame the results found using the multimodal classifier introduced here. Furthermore, the significant increased accuracy of our model over the most commonly used approach to define sentiment in previous studies provides a foundation to further investigate this avenue of research and analyse questions previously answered using single-modality, traditional models.

Thirdly, Chapter Five provides evidence that the multimodal classifier developed in this thesis exhibits a strong and significant relationship with both short-term and long-term abnormal returns. Specifically, the analysis demonstrates a stronger association, as indicated by a marked increase in the coefficient of determination ($R^2 = 0.6$) for short-term Cumulative Abnormal Returns (CARs). This result surpasses the most notable findings from prior research (e.g., Doran et al. 2012; Mayew and

[278] See Chapter 2.

Venkatachalam, 2012; McKay Price et al. 2012; Brockman, Li and McKay Price, 2015). The classifier also shows marginal improvements in forecasting long-term CARs, highlighting its robustness across both short- and long-term timeframes (McKay Price et al. 2012).

Finally, Chapter Six reveals that multimodal sentiment has a highly significant relationship and is negatively related to longer-period Cumulative Abnormal Volatility (CAV). In comparison to the only other study that evaluates the relationship between earnings conference call sentiment and CAV over a longer horizon (McKay Price et al. 2012), the findings contained within Chapter Six indicate that multimodal sentiment can more successfully explain longer period fluctuations. Again, this result is identified through stronger R^2 coefficient to the only other study evaluating the same question. These findings reinforce the importance of considering multimodal sentiment in understanding market behaviours over extended periods.

Taken together the technical contributions presented in this thesis represent a significant advancement in the field of financial sentiment analysis, particularly in the context of multimodal data. The creation of the largest multimodal earnings conference call dataset provides a robust foundation for future research, enabling researchers to explore the nuanced interplay of textual, vocal, and behavioural cues in financial communication. The use of state-of-the-art NLP techniques in developing the multimodal sentiment classifier used throughout this thesis not only outperforms existing methods in terms of classification accuracy but also demonstrates its practical relevance by delivering superior predictive accuracy for both financial metrics investigated, cumulative abnormal returns (CAR) and cumulative abnormal volumes (CAV). These findings underscore the value of integrating multimodal sentiment analysis into asset pricing frameworks, offering a more holistic and reliable approach to understanding market behaviours. Collectively, these contributions highlight the transformative potential of multimodal methodologies, paving the way for more accurate, efficient, and contextually aware financial sentiment analysis tools, which have the potential to deepen the areas understanding of financial decision-making within academic research and real-world scenarios.

7.2.2 Theoretical Contributions

The creation of a more accurate sentiment classifier that captures both textual and audio modalities was primarily to allow for a deeper investigation into the impact earnings call content has on financial participant decision making, and consequently financial market behaviours.

The theoretical contributions of this thesis primarily concern the findings of Chapters Five and Six, surrounding investor decision making and market reactions in response to multimodal sentiment. The results of these studies have been framed within the context of behavioural theory. Specifically, by emphasising the role of investor psychology and sentiment, this analysis contributes to a deeper exploration of how behavioural factors drive market dynamics. Chapter Five conducts an event study which evaluates the relationship between multimodal earnings call sentiment and two differing measures of CARs, short- and long-term. Evidence from the analysis indicates that multimodal sentiment has a highly significant positive relationship with short-term abnormal returns whilst returning a highly significant but negative relationship with extended period returns. The behaviour of this relationship aligns with a substantial amount of previous literature (Antweiler and Frank, 2004; Lemmon and Portniaguina, 2006; Tetlock, 2007; Henry, 2008; Tetlock, Saar-Tsechanksy and MacKassy, 2008; Schmeling, 2009; Loughran and McDonald, 2011; Doran et al. 2012; Garcia, 2012; Ho and Hung, 2012; Mayew and Venkatachalam, 2012; McKay Price et al. 2012; Twedt and Rees, 2012; Bathia and Bredin, 2013; Corredor, Ferrer and Santamaria, 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Brockman, Li and McKay Price, 2015; Ferguson et al. 2015; Azar and Lo, 2016; Bannier et al. 2017; Gao and Yang, 2017; Jiang et al. 2019) reinforcing the relationship between financial sentiment and market returns.

The results found in Chapter Five indicate that the market initially overreacts to the information being portrayed on these calls, as proxied by the multimodal sentiment measure, over the immediate term but then returns revert towards a fundamental level in attempts to correct the initial mispricing. This suggests that as the information discussed on an earnings call becomes more positive (negative), the initial reaction in CAR becomes more excessive (dampened), which then subsequently begins to move in the opposite direction over a longer horizon. This relationship aligns with mean reversion (Poterba and Summers, 1988) which, within the context of behavioural theory, states that asset prices revert to fundamental values over an extended period of time following an initial overreaction. Therefore, the results found in the main analysis of Chapter Five align with the expectations of CAR behaviour in relation to earnings call sentiment (Antweiler and Frank, 2004; Lemmon and Portniaguina, 2006; Tetlock, 2007; Henry, 2008; Tetlock, Saar-Tsechanksy and MacKassy, 2008; Schmeling, 2009; Loughran and McDonald, 2011; Doran et al. 2012; Garcia, 2012; Ho and Hung, 2012; Mayew and Venkatachalam, 2012; McKay Price et al. 2012; Twedt and Rees, 2012; Bathia and Bredin, 2013; Corredor, Ferrer and Santamaria, 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Brockman, Li and McKay Price, 2015; Ferguson et al. 2015; Azar and Lo, 2016; Bannier et al. 2017; Gao and Yang, 2017; Jiang et al. 2019).

The main results of Chapter 5 indicate that market behaviour in relation to earnings calls follows a pattern of mean reversion, and behavioural theory indicates that mean reversion arises due to psychological biases. Therefore, Chapter 5 continued this investigation to assess if there were any evident psychological biases driving the relationship found. Building upon the prior literature (Erickson et al. 1978; Conley, O'Barr and Lind, 1978; Brooke and Ng, 1986; Bradac, Mulac and House, 1988; Chattopadhyay et al. 2003; Feinberg et al. 2005; Klofstad, Anderson and Peters, 2012; Martín-Santana et al. 2015; Wang et al. 2018; Chau et al. 2020; Song et al. 2020) showing that specific levels of paralinguistic characteristics in the communication process play a significant role in creating speaker perceptions and impact the persuasiveness of a message, the extended analysis of Chapter 5 evaluates whether call participants conveying information to the market using certain levels of paralinguistic traits have any impact on market behaviour.[280]

The specific psychological bias that is tested for in this additional analysis is the framing effect. McMahon (2005) identifies that the framing psychological bias significantly alters decisions made based on how the information used to make said decision is conveyed or framed. Tversky and Kahnmen (1981) note that the psychological principles that govern the perception of decision problems and the evaluation of probabilities and outcomes produce predictable shifts of preference when the same problem is framed in different ways. Furthermore, the authors highlight that the effects of frames on decision-making can be compared to the effects of perceptions on perceptual appearance. Therefore, due to the multimodal model containing paralinguistic features which give insight into how speakers on earnings calls are framing the information disseminated, along with well-established speaker perceptions in relation to specific paralinguistic traits (Mendoza and Carballo, 1998; Chattopadhyay et al. 2003; Feinberg et al. 2005; Park et al. 2011; Klofstad, Anderson and Peters, 2012; Giddens et al. 2013; Martín-Santana et al. 2015; Wang et al. 2018; Chau et al. 2020; Song et al. 2020), allowed for Chapter Five to investigate whether the framing effect is evident in the financial decision-making process.

From a theoretical perspective, leveraging the paralinguistic modality used within this thesis to identify whether there is a framing bias in financial decision making, allowed this analysis to provide an answer for the psychological driver creating the mean reversion phenomenon found in Chapter 5.5.3. Furthermore, Tversky and Kahnman (1981) identify that rational choice requires that the final outcome to some decision should not be impacted by the way in which the information used to make that decision is framed. Therefore, by analysing whether the way in which information is framed on earnings calls has any significant effect on market behaviour allows for this analysis to contribute to the field of behavioural finance by providing additional evidence in line with the sub-rational characteristics of market participant decisions.

To evaluate the framing effect on earnings calls, the further analysis (Chapter 5.5.4) evaluates the impact differing levels of paralinguistic traits among call participants has on CAR behaviour. The results show that on calls where there are significantly different levels of paralinguistic features among participants, the reaction in CARs is more pronounced than the reaction found in the main results. Although each of the four paralinguistic features used to identify a divergence in the way information is framed on earnings calls – pitch, intensity, jitter and shimmer – all return significant results for short-term CARs, the strongest result is in relation to calls whereby managers have a higher level of jitter in comparison to analysts. Low levels of paralinguistic jitter are associated with stress whilst high levels of jitter are perceived as less stressed. Wang et al (2021) identify that appeals of persuasion are more successful when persuaders vocal characteristics are less stressed due to the perception of calmness and

[280] See Chapters 1.4, 2.2.3 and 5.5.2 for the relevant discussions of the psychology research.

confidence. Hence, when there is a significant difference in paralinguistic stress characteristics (jitter) between managers and analysts on earnings calls (where managers are less stressed) the market experiences a stronger reaction in CARs. Specifically, as the sentiment on earnings calls increases, when managers are calm in comparison to stressed analyst participants, CARs react more positively over the short-term. Evaluating the impact of divergence of paralinguistic features on longer period CARs, the only significant result returned was in relation to calls where managers have low jitter compared to analysts with high jitter paralinguistic characteristics. This suggests that calls where managers are being perceived as more stressed than their call counterparts induce a stronger reversal in returns over a longer horizon.

Contributing to the area of behavioural finance, this thesis finds evidence of framing bias in market participants financial decision-making process in relation to conference call information, which is a significant contributor to the mean reversion behaviour found. This insight contributes to the understanding of information processing by market participants by indicating that market agents react differently dependent on the manner with which information is presented vocally, causing sub-rational decision making and consequently moving prices away from fundamental values.

Chapter Six conducts a similar analysis to that of Chapter Five and again makes contributions to the theoretical understanding of financial decision making however from the perspective of trading volume behaviour. Building on previous literature evaluating similar questions (Antweiler and Frank, 2004; Tetlock, 2007; McKay Price et al. 2012; Garcia, 2013; Sprenger et al. 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014:2017) this analysis evaluates the market reaction to CAV over the short and longer terms in relation to multimodal sentiment. The results indicate that over the short-term multimodal sentiment exhibits a positive but statistically insignificant relationship with CAV. However, over the extended period multimodal sentiment has a negative and highly significant relationship with CAV. These results suggest that heightened sentiment during earnings calls leads to an initial rise in trading volumes, which is then followed by a decline over the subsequent 60 days. While this initial uptick in trading volume aligns with both sentiment and informational theories, it is important to note that the finding lacks statistical significance. In contrast, the longer-term analysis reveals a negative and statistically significant relationship between CAVs and multimodal sentiment, suggesting that trading volume reverts to baseline levels over time, consistent with informational theory. These results contribute to sentiment literature as the results do not fully resonate with either traditional or behavioural theory. Neither theory fully captures the relationship between the absolute values of multimodal sentiment from earnings conference calls and CAVs implying that a new theory could potentially be developed to frame these results.

Following the main analysis of Chapter Six, which evaluates the average level of sentiment on earnings calls, an additional analysis builds upon prior research investigating the impact disagreement among investors has on trading volume (Karpoff, 1986; Harris and Raviv, 1993; Antweiler and Frank, 2004; Banerjee and Kremer, 2010; Atmaz and Basak, 2016). The additional analysis particularly expands on the work of Siganos et al. (2017), by utilising a Divergence of Sentiment (DoS) measure as a proxy for investor disagreement. Focusing on the impact that disagreement has on market behaviour creates a contribution to the area as, to the authors knowledge, this is this is the first study to assess disagreement using both linguistic and paralinguistic elements of communication. The additional analysis specifically explores the link between calls with pronounced levels of participant disagreement and abnormal trading volume. Findings show that earnings calls characterized by higher (or lower) disagreement result in more (or less) substantial short-term market reactions compared to the findings based on absolute levels of sentiment shown in the main analysis. This indicates that the DoS measure, which captures meaningful differences between managers and analysts' tone, acts as a reliable predictor of short-term CAV. Both managerial and analyst optimism are shown to accurately forecast short-term CAV, but calls featuring heightened managerial optimism elicit a particularly strong market reaction in terms of abnormal volume. This pattern is consistently observed across calls demonstrating overall disagreement, as well as those with disagreement specifically in the Q&A segments.

This additional analysis therefore adds to the area of disagreement in financial markets by returning results in line with the theoretical models of investor disagreement which highlight that disagreement of opinions in a market setting induce heightened trading activity (Karpoff, 1986; Harris and Raviv, 1993; Antweiler and Frank, 2004; Banerjee and Kremer, 2010; Atmaz and Basak, 2016). Furthermore, it provides specific insights in regard to which scenarios where disagreement is identified causes markets to react more strongly, showing that when there is a disagreement in sentiment between managers and analysts on these calls (specifically, when managers are significantly more optimistic than their call counterparts) there is a stronger reaction in CAV market behaviour.

The theoretical contributions from this thesis can therefore be summarised as: (i) market reaction in CAR to earnings conference calls directly falls in line with behavioural theory particularly showing mean reversion behaviour; (ii) framing bias appears to drive mean reversion behaviour in relation to earnings calls; (iii) the relationship between multimodal sentiment and CAV does not agree fully with either informational or sentiment theories, which indicates that the results do not fully align with traditional or behaviour theories, and finally; (iv) disagreement on earnings calls, proxied by a divergence in participant sentiment, aligns with theoretical models of investor disagreement (Karpoff, 1986; Harris and Raviv, 1993; Antweiler and Frank, 2004; Banerjee and Kremer, 2010; Atmaz and Basak, 2016) and indicates that disagreement between participants on earnings calls drives heightened trading activity. Overall, the results returned throughout each empirical analysis within this thesis highlight those capabilities of multimodal sentiment classifiers in understanding communication on earnings conference calls and provides the area of research with further insights into market behaviours in reaction to information disseminated on these calls.

7.3 Summary of Findings

This thesis evaluates the area of financial sentiment analysis with a focus on the application of sentiment analysis to the financial disclosure earnings conference calls. Particularly, throughout the empirical chapters of this thesis the applicability and efficacy of a multimodal sentiment classifier applied to earnings calls has been evaluated in relation to the most used sentiment classifiers in previous finance studies. The findings arising from these chapters provide insights not only in the superior abilities of sentiment classification techniques at the forefront of the NLP domain but also in regard to the field of behaviour finance advancing our knowledge of the financial decision-making process and market behaviours in general. This section summarises the theoretical and technical findings arising from each empirical chapter into 5 key themes.

1. **The Multimodal Model introduced in this thesis is more adept at classifying financial sentiment in comparison to long-established methods frequently used within the financial domain.**

Chapter 4 provides evidence towards the theme identified throughout Chapter 2 and in line with prior studies, that more computationally advanced classification methods are better equipped at defining financial sentiment (Kearney and Liu, 2014; Guo, Shi and Tu, 2016; Renault, 2017; Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Alamoudi and Alghamdi, 2021). The results of Chapter 4 displayed in Table 4.3 in relation to the comparative analysis highlight that in comparison to the most used financial sentiment classifiers the multimodal sentiment classifier introduced in this thesis has a better grasp of the language used in financial disclosures and subsequently returns a more robust measure of sentiment. Within the comparative analysis conducted in Chapter 4, eight sentiment classifiers were used to represent different methodologies historically used to identify sentiment in financial disclosures. Of these eight classifiers the multimodal model using a deep learning classifier returned the highest validation accuracy on unseen earnings call sentences of 74.88%. Comparing this classification rate to the most used classifier in previously published studies, Loughran and McDonald's (2011) specific dictionary approach, the multimodal classifier returns an increase in validation accuracy of 27.49%. To put this difference in classification into the context of the full sample of 637,220 sentences, the multimodal model accurately predicts the sentiment of 447,150 sentences in comparison to 301,978 correctly classified sentences using the LM dictionary. The increased number of correctly classified sentences associated with the multimodal model not only shows the models deeper understanding of financial context but also provides beneficial in understanding market behaviours, as shown in the superior forecasting abilities of the model show in Chapter 5.

Interestingly, the findings of Chapter 4 align with that of Loughran and McDonald (2011) who show that finance specific sentiment classification methods are better suited to analyse financial disclosures. The authors make this conclusion due to their finding that 73.8% of words labelled negative in the Harvard-IV dictionary are not negative in a financial context. Tables 4.3-4.5 show the superior performance of the LM dictionary over the Harvard-IV dictionary and FinBERT over general transformers, providing evidence consistent with Loughran and McDonald’s (2011) findings. Furthermore, the collective results in Chapter 4 highlight that the addition of paralinguistic information to text-based classifiers marginally heightens the classification performance and adds to the models understanding of earnings conference calls.

Looking deeper into the abilities of each of the eight classifiers used in Chapter 4, the analysis progressed to evaluate which models have the best understanding of each of the three sentiment categories used in this thesis (positive, negative and neutral) and the classification accuracies of each model when the validation set is disaggregated across call sections and participants. Evaluating which models perform better on each of the three sentiment categories, the Multimodal models (DNN and NN) achieve the highest classification rates across all categories, with slightly better results on positive sentences compared to neutral and negative sentences, reflecting a strong understanding of financial language disseminated with favourable context. Notably, the multimodal model excels in negative sentiment classification, achieving rates 35% higher than the LM dictionary. Prior literature that identifies negative sentiment as a stronger predictor of firm performance (Tetlock et al. 2008; Loughran and McDonald, 2011; Mao and Bollen, 2011; Mayew and Venkatachalam, 2012). Hence, identifying a further potential reason for the models strong forecasting performance in Chapters 5 and 6.

The results outline in Table 4.5 again provide valuable insights into the relative strengths and weaknesses of each classifier however in respect the call sections and participants. The multimodal method consistently achieves strong performance across all categories and outperforms other approaches in classifying sentences from the Q&A section, as well as those delivered by managers during the call. Thus, demonstrating the advantage of incorporating paralinguistic features in sentiment analysis models, particularly for interpreting messages conveyed in the conversational context often seen in the Q&A section of earnings calls. This finding is especially noteworthy, as prior studies have shown that sentiment from the Q&A section has greater predictive power for market outcomes than sentiment from the management discussion section (McKay Price et al. 2012; Borochin et al. 2017; Fu et al. 2019).

2. **Paralinguistic information increases the ability of sentiment classifiers marginally. The increase in classification performance was driven by a set of paralinguistic features that have been shown to increase emotion and sentiment classification across different domains.**

A key objective of this thesis was to identify whether more advance methods to measure sentiment, as shown in the wide NLP domain, returned enhanced results for the classification of financial sentiment (Houjeij et al. 2012; Bhaskar, Sruthi and Nedungadi, 2014; Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Yan, Xu and Gao, 2020; Alamoudi and Alghamdi, 2021; Dair, Donovan and O'Reilly, 2021). Introducing a model that uses transformer architecture and paralinguistic information together in a multimodal fashion allowed the results stemming from this research to give an indication into whether these techniques are more efficient at defining sentiment that historically used methods. The results from the comparative analysis in Chapter 4 indicate that the multimodal method is more robust at classifying financial sentiment over the most used methods in previous publications. Table 4.3 shows that generally pretrained transformer architecture returns a better understanding of financial sentiment in comparison to Loughran and McDonald’s (2011) specific dictionary method (17.78% increase in validation accuracy) whilst the financially pretrained variant FinBERT generates an even greater disparity in classification accuracy (26.07% increase).[281] These results confirm the findings that transformer architecture has revolutionised the capabilities machines have in understanding qualitative information (Munikar, Shakya and Shrestha, 2019; Sun et al. 2020; Gillioz et al. 2020; Alamoudi and Alghamdi, 2021) but further evidence their abilities in defining sentiment within financial context.

Transformer architectures like BERT excel in sentiment analysis by leveraging contextual word embeddings, enabling them to capture nuanced meanings and relationships within text that dictionary-based methods cannot. Unlike static dictionaries, which rely on predefined word lists, BERT dynamically interprets sentiment based on the surrounding context, leading to more accurate and adaptable sentiment classification. The innovative methodology used to create transformers therefore allow even generally pretrained models to substantially outperform finance specific dictionary-based methods. Although a substantial proportion of the increase in validation accuracy of the multimodal model over commonly used methods seems to be driven by the utilisation of the transformer model, the incorporation of the severely under studied paralinguistic modality of earnings conference calls does indeed provide marginal improvements over purely text-based models. Looking directly at the impact paralinguistic features have on sentiment classification, the Multimodal FinBERT NN model, which integrates text and audio modalities, outperforms the text-only FinBERT model by 1.18%. These models are identical minus the incorporate of paralinguistic information, indicating that incorporating audio can enhance the accuracy of earnings call sentiment analysis, though the improvement is marginal. To provide context to this increase in accuracy, the increase in 1.18% translates into an increase of 7,520 correctly classified messages when applied to the overall S&P 100 sample. This finding is notable as it suggests that multimodal approaches may provide a more precise measure of sentiment, potentially improving analyses of the relationship between financial disclosures and trading behaviour.

These findings emphasise the conclusions arising from psychology literature that the consideration of multiple modalities leads to a greater understanding by machines of the sentiment being disseminated in natural language (Morency, Mihalcea and Doshi, 2011; Houjeij et al. 2012; Wollmer et al. 2013; Poria, Cambria and Gelbukh, 2015; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021). Of the ten paralinguistic features used to enhance the multimodal models understanding of earnings call content, each feature has been shown to return informative in the additional permutation

[281] Falling in line with Howard and Ruder (2018) results that domain specific pretraining increases transformer models classification capabilities.

importance analysis. In extant psychology literature four of these features have been extensively in relation to their impact on the persuasion, namely: pitch, intensity, jitter and shimmer (Brooke and Ng, 1986; Wallbott, 1982; Apple et al. 1979; Gelinas-Chebat et al. 1996; Mendoza and Carballo, 1998; Chattopadhyay et al. 2003; Feinberg et al. 2005; Park et al. 2011; Klofstad, Anderson and Peters, 2012; Giddens et al. 2013; Martín-Santana et al. 2015; Wang et al. 2018; Chau et al. 2020; Song et al. 2020). The remaining features - number of periods, fraction of unvoiced, number of voice breaks, mean autocorrelation, mean noise-to-harmonics ratio and audio length – even though studied to a lesser extent in terms of their psychological implications on decision making have been show to significantly improve sentiment classification (Nwe, Foo and De Silv, 2003; Morrison, Wang and De Silva, 2007; Rong, Li and Chen, 2009; Morency, Mihalcea and Doshi, 2011; Lee, Kim and Kang, 2014; Poria, Cambria and Gelbukh, 2015).

From the permutation feature importance analysis, Table 4.6 demonstrates that some features hold greater importance than others. Notably, the fraction of unvoiced features is nearly twice as informative as the third most significant feature, mean pitch. Additionally, shimmer local and the noise-to-harmonics ratio are shown to be particularly influential. These findings align with prior research, which highlights the significance of the fraction of unvoiced features (Morrison, Wang and De Silva, 2007), shimmer (Li et al. 2007; Jacob, 2016), and pitch (Koolagudi and Rao, 2010; Koolagudi and Krothapalli, 2012; Chebbi and Jebara, 2018) in sentiment and emotion classification. Interestingly, mean intensity is identified as the least informative feature, though it still contributes marginally to the predictions made by the multimodal classifier.

3. **Multimodal sentiment returns statistically significant positive (negative) relationship with short term (long term) CARs. These results fall directly in line with the expectations of behavioural theory in relation to Cumulative Abnormal Returns (CARs)**

Chapter 5 applies the multimodal classifier, shown to excel in classifying earnings call sentiment, to the full dataset of S&P100 firms across the period 2005-2021. The results from this chapter provide further clarification that the multimodal classifier has a greater understanding of financial communication but also provided insights into the behavioural tendencies of financial market participants in reaction to the information disseminated on these calls. For each aggregation of call sentiment used in the main analysis of Chapter 5, the coefficient of determination (R^2) and the adjusted coefficient of determination (Adj R^2) exceed 0.71 in predicting short-term CARs. Therefore, further solidifying the multimodal model's strong predictive capability of initial market reactions. Comparing the multimodal models explainability of short-term CARs to previous research the results are significantly more robust. Doran et al. (2012) used both the Harvard-IV4 and Henry (2006) dictionary approaches to classify sentiment in earnings conference calls, with the Henry (2006) dictionary achieving a maximum R^2 of 0.0069 for short-term CAR. Similarly, Mayew and Venkatachalam (2012) reported an adjusted R^2 of 0.0764 when assessing earnings call sentiment based solely on paralinguistic cues from managerial vocal characteristics. McKay Price et al. (2012) achieved an R^2 of just 0.0016 using the Henry (2006) dictionary approach for overall call sentiment. These results, particularly when contrasted against previous research using text-based dictionaries, highlight the valuable insights multimodal sentiment provides into short-term market behaviours.

Providing insights into market behaviour in reaction to earnings calls, the main analysis of chapter 5 identifies a highly significant relationship between multimodal call sentiment with both short- and longer-term CARs. The findings indicate that the multimodal sentiment classifier exhibits a highly significant positive relationship with short-term Cumulative Abnormal Returns (CARs) and a highly significant negative relationship with longer-term CARs. This direction of relationship concurs with the bulk of prior literature that indicates that financial disclosure sentiment has a positive relationship with short term abnormal returns (Antweiler and Frank, 2004; Tetlock, 2007; Henry, 2008; Tetlock, Saar-Tsechanksy and MacKassy, 2008; Loughran and McDonald, 2011; and Garcia, 2012; Doran et al. 2012; Mayew and Venkatachalam, 2012; McKay Price et al. 2012; Twedt and Rees, 2012; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Brockman, Li and McKay Price, 2015; Ferguson et al. 2015; Azar and Lo, 2016; Bannier et al. 2017; Jiang et al. 2019) with a negative association with longer period abnormal returns (Lemmon and Portniaguina, 2006; Tetlock, 2007; Schmeling, 2009; Ho and Hung, 2012; Bathia and Bredin, 2013; Corredor, Ferrer and Santamaria, 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Gao and Yang, 2017; Jiang et al. 2019). The short-term relationship suggests that market participants react quickly to the information presented during earnings calls, aligning their initial responses with the sentiment captured by the multimodal classifier. However, over the longer horizon the negative relationship between sentiment and returns suggests that investors initially overreact to the sentiment conveyed in earnings calls, driving prices away from their fundamental values. This prompts a subsequent market correction as participants reevaluate their positions to address sentiment-driven mispricing. This observed dynamic aligns with behavioural theories, which propose that prices do not simply adjust to and stabilize at fundamental values but instead exhibit initial overreactions followed by reversals.

4. **The cause of the reactions found in relation to CAR from multimodal sentiment indicate that financial market participants are influenced by the framing heuristic.**

Due to the incorporation of the additional paralinguistic modality within this thesis, it allowed Chapter 5 to further investigate the drivers behind the mean reversion dynamic identified in the sentiment returns relationship. Existing research highlights that both the content of our communication and the manner in which we convey it are important (Guyer, Fabrigar and Vaughan-Johnston, 2018). Put simply, the way we speak carries significant information beyond the words themselves. In fact, a substantial body of psychological literature explores the role of vocal characteristics in influencing persuasion and decision-making (Chattopadhyay et al. 2003; Feinberg et al. 2005; Klofstad, Anderson and Peters, 2012; Martín-Santana et al. 2015; Wang et al. 2018; Chau et al. 2020; Song et al. 2020). To understand whether differences in vocal communication impacts the decision-making process and subsequently the market behaviours in response to earnings call content, Chapter 5 evaluated the reaction in abnormal returns on calls displaying significant differences in participant audio characteristics against the full sample of calls.

Based on these additional findings, this thesis argues that the initial overreaction in CARs stems from a framing bias influencing financial market participants' decision-making processes regarding the information presented during earnings calls. By utilizing a multimodal sentiment classifier that incorporates paralinguistic cues—known to affect persuasion and decision-making—the research highlights how nuances in communication style significantly shape market participants' responses. The way information is conveyed during these calls appears to contribute to framing bias, leading to initial overvaluation. Over time, the market reassesses and corrects this mispricing. When framing is present, as evidenced by variations in communication style, discrepancies emerge in the decision-making processes of market participants, further amplifying cycles of overvaluation and subsequent correction. These findings enhance our understanding of how information framing influences market behaviour, shedding light on the interplay between communication and financial decision-making. This supplementary analysis particularly reveals that on calls where analysts exhibit significantly higher intensity compared to managers, and calls where managerial jitter surpasses analyst jitter, short-term CARs react in an amplified manner. These findings, alongside the main results, provide further evidence that market participants are prone to sub-rational decision-making influenced by how information is presented during calls, rather than relying solely on the fundamental data provided.

5. **Multimodal sentiment has a positive relationship with short-term CAV but returns no statistical significance and a negative relationship with longer period CAV that is statistically significant.**

Further investigating the relationship between sentiment and market characteristics, Chapter 6 evaluates the impact of CAV in reaction to earnings call multimodal sentiment following prior research (Antweiler and Frank, 2004; Tetlock, 2007; McKay Price et al. 2012; Garcia, 2013; Sprenger et al. 2013; Siganos, Vagenas-Nanos and Verwijmeren, 2014:2017). In relation to short term CAV, multimodal sentiment returns a positive relationship falling in line with previous literature (Mao and Bollen, 2011; Garcia, 2013; Sprenger et al. 2013; Bochkay et al. 2020; Gu and Kurov, 2020) however this relationship returns insignificant. Furthermore, the findings demonstrate that multimodal sentiment has a significant negative relationship with abnormal trading volumes over longer periods. Compared to the only other study examining the connection between earnings conference call sentiment and trade volumes over extended time horizons (McKay Price et al. 2012), the findings suggest that multimodal sentiment provides a more robust explanation for long-term trading fluctuations. This conclusion is supported by a higher R^2 value, emphasizing the importance of incorporating multimodal sentiment into analyses of market behaviour over prolonged durations.

The initial increase in trading volume corresponds with both sentiment and informational theories, though our findings are not statistically significant. However, the negative and statistically significant association between CAVs and multimodal sentiment over longer time horizons suggests a return of trading volume to baseline levels, aligning with informational theory. Overall, the analysis indicates that neither traditional nor behavioural theories fully explain the observed relationship between the absolute values of multimodal earnings conference call sentiment and trading volume. These results raise the question of whether further investigation into the multimodal sentiment and CAV dynamic could give rise to a fresh asset pricing theory which explains the relationship found here.

6. **Earnings Calls displaying high levels of participant disagreement induce a heightened short-term market reaction in trading volume.**

Building upon the main analysis of Chapter 6 which analysed the relationship between absolute values of earnings call sentiment and CAV, the additional analysis evaluated the relationship between a divergence of sentiment measure and CAV. In doing so Chapter 6 advances the understanding of disagreement in financial disclosures, proxied using a DoS measure, in relation to abnormal trading volume expanding on previous literature (Karpoff, 1986; Harris and Raviv, 1993; Antweiler and Frank, 2004; Banerjee and Kremer, 2010; Atmaz and Basak, 2016). Building on the work of Siganos et al. (2017), this additional analysis is the first to incorporate both linguistic and paralinguistic content in analysing disagreement within earnings calls.

The findings reveal that calls with higher disagreement between managers and analysts lead to significantly larger short-term market reactions in abnormal trading volumes compared to calls with lower disagreement. These findings align with theoretical models of disagreement (e.g., Karpoff, 1986; Harris and Raviv, 1993; Antweiler and Frank, 2004) and emphasize the importance of multimodal sentiment analysis in understanding the dynamics of market behaviour. Furthermore, while both managerial and analyst optimism influence short-term trading volumes, calls with greater managerial optimism elicit more pronounced market responses. This pattern holds true for disagreement detected across the entire call and in the Q&A section specifically, highlighting DoS as a robust predictor of short-term CAV. Over the longer horizon, for both sets of divergent calls when managers are more optimistic or when analysts are more optimistic than their counterparts, as sentiment increases abnormal trading volume is lower. This may entail that disagreement among participants on these calls teases out new fundamental information which gives market participants as a whole a better understanding of future firm performance allowing for better decisions to be made initially, moving prices to fundamental values with a dampening of trading over the longer horizon as prices reside at a correct level.

7.3 Limitations

As with any research project there stands limitations associated with this thesis. This section will outline the main limitations that have arisen over the period of this research. Following the limitations outlined here, the future directions identified from the analysis conducted within this research are discussed. Four main limitations faced when conducting the research for this thesis emerged from the creation and training of the multimodal sentiment classifier: (i) the size of the pre-classified multimodal training dataset, (ii) the accuracy of the text-audio alignment process, (iii) the approach to fusing together textual and paralinguistic information as input for the deep learning classifier and (iv) lack of explainability of the model used. Each of these limitations are discussed below, with appropriate justification for why these limitations have been deemed acceptable for this exploratory study of multimodal analysis.

Firstly, the multimodal classifier is trained on 2,106 manually classified sentences. The only study to the author's knowledge which evaluates the accuracy of a financial sentiment classifier based upon the number of training sentences used is Renault (2020), who evaluates classification accuracy from a model trained on 500 up to 1 million sentences, finding that classification accuracy increases from 59.6% to 73.08% when increasing the training size from 500 sentences to 1 million. Furthermore, Renault (2020) suggest that a training set of between 100,000 to 250,000 is most optimal as the increase in accuracy using larger datasets is marginal. Even though the models understanding of financial information would increase with a larger training set, the results from this analysis still return a comparatively high classification accuracy. However, the ability to manually classify a substantially larger number of sentences would increase the models understanding of the nuances of financial communication and hence increase the probability of deepening the domains knowledge of the sentiment market characteristics dynamic.

Secondly, in the generation of the multimodal sentiment dataset used throughout this thesis, the text-audio alignment presented a limitation. The process attempted to align textual transcripts and corresponding audio of 2.6 million sentences. However, the alignment model returned a full repository of paralinguistic features for 637,220 sentences which translates into roughly 30% alignment rate. From the sample of 100 companies, 98% of calls remained within the final sample. This suggests that many calls are missing paralinguistic data for several sentences, which translates into a substantial amount of information missing from the calls. A sentence missed could be a future looking statement that contains incremental information used by investors and analysts that subsequently moves the fundamental security price. Incremental information surrounding poor performance or bleak outlooks for the future are also potentially missed due to this low level of alignment. Although there are numerous sentences missing from each of the calls used within this analysis, the multimodal dataset is the larger than any other dataset leveraging paralinguistic information from earnings calls (Mayew and Venkatachalam, 2012; Li et al. 2020; Chen, Han and Zhou, 2023). Therefore, the use of this dataset is a step forward from datasets used in previous research. Furthermore, this research should be considered an early exploration into multimodal analysis, with model accuracy constantly improving in tandem with increased computational power in accuracy. However, although this multimodal aligned dataset is the largest of its kind to date in the finance domain, leveraging a more robust method to align textual and paralinguistic content on earnings calls would significantly increase the size of the dataset used and subsequently may lead to a more insightful analysis of multimodal sentiments relationship with market behaviour.

Thirdly, the combination of text and paralinguistic data fed into the deep learning classifier proved to be a limitation. Using an elementary approach the multimodal classifier simply takes into consideration the textual representations from FinBERT and the paralinguistic information produced from the forced alignment process together. No fusion of information was completed when creating this model, however studies developing multimodal models in various other domains have used more advanced methods. For example, Poria et al. (2016) highlight that the fusion of multiple modalities to create a multimodal classifier is an important prerequisite. In their analysis of fusion methods, they evaluate the abilities of two main fusion strategies - feature-level fusion and decision-level fusion.[282] Ghandi et al. (2023) emphasize that there are multiple approaches for integrating different modalities within a multimodal classifier. However, their review of sentiment analysis literature reveals a significant gap in the analysis of text-audio multimodal classifiers. As a result, there is no established consensus on the most effective methods for extracting and fusing these two modalities. Hence, the decision to adopt the elementary approach here was due to a lack of consensus on how to fuse together text and audio features successfully in an established manner like audio and visual fusion. Nonetheless, the creation of a text-audio fusion method has the potential to increase the multimodal classifier's ability to understand financial information and hence the basic method used in this thesis stands a limitation of the model's potential.

The final limitation of this research, and a limitation of all research using advanced methods such as the ones used within this thesis, is the lack of explainability surrounding the model. Xu et al. (2019) highlight that DNNs, such as the one adopted in this research, have advanced prediction capabilities however lack the ability to explain how they arrive at their superior results. While the DNN model effectively integrates textual representations derived from transformer architecture with paralinguistic features to achieve higher classification accuracy over the single modality models used in Chapter 4, the DNN classifier remains a black box, making it difficult to interpret how specific inputs contribute to predictions. This stands as a limitation as it does not allow the research to conclude definitely to which specific modality of information is driving results and within that modality which words or paralinguistic features are most important.

7.4 Future Directions

As the area of computer science expands and grows new capabilities, the methods described and applied throughout this thesis could be made more efficient, cost and time effective contributing to further advancements in the field of financial sentiment analysis. Over the course of this thesis, the literature evaluated that shaped the direction of this research along with the results and conclusions drawn, have identified that further developing the models used to define sentiment create the capacity

[282] Feature level fusion integrates the features from different modalities of data into one singular representation before being fed into a classifier. On the other had decision-level fusion fuses together the outputs from separate classifiers which only take into consideration singular modalities and create a final output based upon singular outputs.

to further understand financial disclosure content and market behaviours. Subsequently, from the perspective of behavioural finance there lies the possibility to further advance our understanding of the decision-making process of market participants in relation to the dissemination of information from financial disclosures. Hence, although the conclusions drawn from this research add to our understanding of behavioural finance, the research conducted here is still ongoing and the field of research still contains questions that require investigation.

Throughout this thesis, earnings conference calls have been shown to be useful in evaluating the presence of heuristics and behavioural biases, particularly as these calls offer the opportunity to study market reactions to the understudied paralinguistic modality on these disclosures. While a substantial body of research has already explored the relationship between earnings conference call sentiment and market reactions (Davis et al. 2012; Doran et al. 2012; Larcker and Zakolyukina, 2012; Mayew and Venkatachalam, 2012; McKay Price et al. 2012; Davis et al. 2015; Blau, DeLisle and McKay Price, 2015; Brockman, Li and McKay Price, 2015; Wang and Hua, 2014; Borochin et al. 2017; Chen, Nagar and Schoenfeld, 2018; Fu et al. 2019; Amoozegar et al. 2020; Bochkay et al. 2020), significant scope for further investigation remains. This thesis underscores the advantages of utilizing multiple modalities to deepen our understanding of the communication dynamics during earnings calls. Although the application of advanced Natural Language Processing (NLP) techniques within finance is still in its infancy, the results presented here underscore the promising insights into behavioural finance that such approaches may yield in the future.

The technical and theoretical contributions associated with this thesis create a solid base for further research of multimodal sentiment techniques within the domain of financial sentiment analysis. The research conducted here offers several promising avenues for future research into the decision-making behaviours of financial market participants in reaction to earnings conference call content. As shown throughout this thesis, the abilities of forecasting and understanding behaviours in market characteristics is increased as the capabilities of the models being used to classify sentiment increases (Loughran and McDonald, 2011; Kearney and Liu, 2014; Guo, Shi and Tu, 2016; Renault, 2017; Hiew et al. 2019; McGurk, Nowak and Hall, 2020; Jayaraman and Dennis, 2020). One of the contributions to literature this thesis makes, which aided in the multimodal model having a deeper understanding of earnings call content, was the incorporation of the seldom used paralinguistic modality of these calls (Mayew and Venkatachalam, 2012; Li et al. 2020; Chen, Han and Zhou, 2023). Creating and working with paralinguistic information was a difficult process which highlighted ways in which the use of multimodal sentiment techniques can be streamlined in the future. Particularly, the process of aligning the text and audio data of earnings calls and generating paralinguistic features can be improved upon in future analysis.

A potential future direction arising from this thesis is therefore the creation of a more robust way to align text and audio information. Streamlining this process would allow for a greater number of sentences to be accurately aligned and therefore generate a larger dataset that contains a full repository of paralinguistic features. Chapter 3 of this thesis identifies that the text-audio alignment process was roughly 30% accurate i.e., only 30% of the 2.6 million sentences were aligned and a full list of features were generated. In line with previous literature (Chen and Lin, 2014; Korotcov et al. 2017; Sun et al. 2017) deep learning methods perform substantially better on larger datasets. Therefore, increasing the size of the dataset used to train the classifier used throughout this thesis may be improved upon with a more efficient method for aligning and generating paralinguistic features. Furthermore, a more accurate method of generating paralinguistic features would allow for more sentences from each individual earnings call to be used in the analysis of market behaviours, meaning there is a lower probability that sentences conveying important fundamental information on these calls are missed in a multimodal analysis.

The fusion of text and paralinguistic data also generates another promising area for future research. Ghandi et al. (2023) indicate that there are multiple ways in which information from differing modalities can be fused together in their review of sentiment analysis literature. However, they note that there is a lack of consensus on the best approach to fuse textual and audio data. Furthermore, the development of text-audio based fusion methods lag behind that of tri-modality fusion techniques. Creating a robust method to fuse together textual and paralinguistic information will grant the sentiment classifier with a deeper understanding of financial context and allow for more robust investigations into market reactions to be conducted.

Another promising area for future research is the application of multimodal analysis to different financial information sets. This thesis focuses on calls from the S&P 100 index however the application of the of multimodal analysis to earnings calls from less visible firms and mediums of information beyond earnings calls has the potential to unearth further insights into the decision-making process of market participants in response to financial information dissemination. Previous literature (Jiang et al. 2019; Seok, Cho and Ryu, 2019; Maidiya and Kresta, 2024) suggests that sentiment forecasting power is more robust when there is a higher degree of informational asymmetry, which is more prominent in less visible firms. Therefore, applying the methods used here to less visible firms may produce further insights into market participants processing, understanding and use of the information portrayed on earnings calls. Furthermore, future research would benefit from extending the application of financial multimodal sentiment beyond the scope of earnings conference calls to include information disclosures that are disseminated using video. The gold standard of multimodal sentiment analysis considers the combination of textual, audio and visual data with multiple studies showing that the performance of tri-modality models is greater than that of models only considering dual modalities (Morency, Mihalcea and Doshi, 2011; Houjeij et al. 2012; Wollmer et al. 2013; Bhaskar, Sruthi and Nedungadi, 2014; Poria, Cambria and Gelbukh, 2015; Yan, Xu and Gao, 2020; Dair, Donovan and O'Reilly, 2021). Hence, the application of a tri-modal model to financial information has the potential to further understand the nuances in the financial communication process beyond that of the bi-modal model used here and consequently provide new insights into market behaviours.

Furthermore, the application of the multimodal sentiment classifier used throughout this thesis has promising avenues for future research on the dataset generated and used throughout this research. The dataset generated within this thesis returns 637,220 earnings conference call sentences with accompanying paralinguistic features. Of these sentences, this thesis classified each into three sentiment categories of positive, negative and neutral and analysed how the market behaved in reaction to a sentiment indicator based on these categories. Future research may look to create more insightful categories of sentiment that directly align with previous psychology literatures insights on decision-making. A common theme discussed throughout prior psychology literature is the persuasive abilities associated with specific vocal traits. It is thought that paralinguistic information which conveys qualities such as credibility, tranquillity, trustworthiness, maturity and confidence consequently increase the persuasive abilities of the speaker. A consensus of research indicates that the traits which heighten persuasiveness, that are included within the repository of paralinguistic features used in this analysis, are lower levels of pitch, higher intensity and higher levels of jitter and shimmer (Mendoza and Carballo, 1998; Chattopadhyay et al. 2003; Feinberg et al. 2005; Park et al. 2011; Klofstad, Anderson and Peters, 2012; Giddens et al. 2013; Martín-Santana et al. 2015; Wang et al. 2018; Chau et al. 2020; Song et al. 2020). Creating sentiment categories which align with the qualities shown in prior psychology literature to impact persuasiveness such as credibility, tranquillity, trustworthiness, maturity and confidence (or any emotion category which has been previously shown to impact decision-making) may allow for the paralinguistic information to be leveraged more efficiently and provide the analysis with a more accurate insight into the reactions of market participants in relation to call content.

Finally, in line with Xu et al. (2019) and Dwivedi et al. (2023) future research could build upon this analysis by focusing on enhancing the interpretability of multimodal sentiment analysis models to address the challenges posed by their black-box nature. Given that the proposed model integrates textual embeddings from transformer architectures with paralinguistic features before classification, understanding how each modality contributes to predictions remains difficult. To improve transparency, future work could build upon this analysis by applying new explainable AI methods to develop definitive understanding of how this classifier achieves greater classification accuracy. Subsequently, understanding the driving textual and paralinguistic features behind this model would play a significant role in understanding market behaviours in reaction to earnings call content.

Bibliography

Abirami, A.M. and Gayathri, V., 2016. A survey on sentiment analysis methods and approach. *2016 IEEE Eighth International Conference on Advanced Computing (ICoAC)*. Chennai, India: Institute of Electrical and Electronics Engineers, pp. 72–76.

Ackert, L.F., Church, B.K. and Deaves, R., 2003. Emotion and financial markets. *Federal Reserve Bank of Atlanta Economic Review*, *88*(2).

Akbas, F., Boehmer, E., Erturk, B. and Sorescu, S.M., 2013. Short interest, returns, and fundamentals. *Returns, and Fundamentals.*

Alamoudi, E. and Alghamdi, N., 2021. Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *Journal of Decision Systems*, 30(2-3), pp.259-281.

Allen, E., O'Leary, D.E., Qu, H. and Swenson, C.W., 2021. Tax specific versus generic accounting-based textual analysis and the relationship with effective tax rates: Building context. Journal of Information Systems, 35(2), pp.115-147.

Amoozegar, A., Berger, D., Cao, X. and Pukthuanthong, K., 2020. Earnings Conference Calls and Institutional Monitoring: Evidence from Textual Analysis. Journal of Financial Research, 43(1), pp.5-36.

Andersen, P.A. and Blackburn, T.R., 2004. An experimental study of language intensity and response rate in e mail surveys.

Antweiler, W. and Frank, M., 2004. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance*, 59(3), pp.1259-1294.

Apple, W., Streeter, L.A. and Krauss, R.M., 1979. Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology,* 37(5), pp. 715–727.

Armitage, S., 1995. Event study methods and evidence on their performance. Journal of economic surveys, 9(1), pp.25-52.

Arthur, W.B., 1994. Inductive reasoning and bounded rationality. The American economic review, 84(2), pp.406-411.

Arthur, W.B., 1995. Complexity in economic and financial markets. Complexity, 1(1), pp.20-25.

Asur, S. and Huberman, B.A., 2010. Predicting the Future with Social Media. ACM International Conference on Web Intelligence and Intelligent Agent Technology, Institute of Electrical and Electronics Engineers, pp. 492–499.

Atmaz, A., Basak, S. and Ruan, F., 2024. Dynamic equilibrium with costly short-selling and lending market. The Review of Financial Studies, 37(2), pp.444-506.

Audrino, F. and Tetereva, A., 2019. Sentiment spillover effects for US and European companies. Journal of Banking & Finance, 106, pp.542-567.

Aune, R.K. and Kikuchi, T., 1993. Effects of language intensity similarity on perceptions of credibility relational attributions, and persuasion. *Journal of Language and Social Psychology*, *12*(3), pp.224-238.

Ayers, B.C., Li, O.Z. and Yeung, P.E., 2011. Investor trading and the post-earnings-announcement drift. *The Accounting Review*, *86*(2), pp.385-416.

Azar, P. and Lo, A., 2016. The Wisdom of Twitter Crowds: Predicting Stock Market Reactions to FOMC Meetings via Twitter Feeds. *The Journal of Portfolio Management*, 42(5), pp.123-134.

Baker, H.K. and Nofsinger, J.R., 2010. Behavioral finance: an overview. Behavioral finance: Investors, corporations, and markets, pp.1-21.

Baker, M. and Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. The journal of Finance, 61(4), pp.1645-1680.

Ball, R., and P. Brown. 1968. An empirical evaluation of accounting income numb Accounting Research 6 (2): 1

Baltussen, G., 2009. Behavioral finance: an introduction. Available at SSRN 1488110.

Banerjee, S. and Kremer, I., 2010. Disagreement and learning: Dynamic patterns of trade. The Journal of Finance, 65(4), pp.1269-1302.

Bannier, C., Pauls, T. and Walter, A., 2017. CEO-Speeches and Stock Returns. Center for Financial Studies, (583).

Bao, W., Li, Y., Gu, M., Yang, M., Li, H., Chao, L. and Tao, J., 2014, October. Building a chinese natural emotional audio-visual database. In 2014 12th International conference on signal processing (ICSP) (pp. 583-587). IEEE.

Barber, B.M. and Odean, T., 2008. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. The review of financial studies, 21(2), pp.785-818.

Barberis, N., 2003. A Survey of Behavioral Finance. Handbook of the Economics of Finance, 1.

Bathia, D. and Bredin, D., 2016. An examination of investor sentiment effect on G7 stock market returns. In Contemporary Issues in Financial Institutions and Markets (pp. 99-128). Routledge.

Bernanke, B. and Kuttner, K., 2005. What Explains the Stock Market's Reaction to Federal Reserve Policy? The Journal of Finance, 60(3), pp.1221-1257.

Bhaskar, J. and Sruthi, K., 2014. Prof. Prema Nedungadi,"Enhanced sentiment analysis of informal textual communication in social media by considering objective words and intensifiers". In IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE).

Bhaskar, J., Sruthi, K. and Nedungadi, P., 2014. Enhanced sentiment analysis of informal textual communication in social media by considering objective words and intensifiers. International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014).

Bhonde, R., Bhagwat, B., Ingulkar, S. and Pande, A., 2015. Sentiment Analysis Based on Dictionary Approach. International Journal of Emerging Engineering Research and Technology, 3(1), pp.51-54.

Black, F. and Scholes, M., 1973. The pricing of options and corporate liabilities. Journal of political economy, 81(3), pp.637-654.

Blau, B., DeLisle, J. and Price, S., 2015. Do sophisticated investors interpret earnings conference call tone differently than investors at large? Evidence from short sales. Journal of Corporate Finance, 31, pp.203-219.

Blitz, D. and Fabozzi, F.J., 2017. Sin stocks revisited: Resolving the sin stock anomaly.

Blonde, J. and Girandola, F., 2016. Revealing the elusive effects of vividness: A meta-analysis of empirical evidences assessing the effect of vividness on persuasion. Social Influence, 11(2), pp.111-129.

Blume, L., Easley, D. and O'Hara, M., 1994. Market Statistics and Technical Analysis: The Role of Volume. The Journal of Finance, 49(1), pp.153-181.

Bochkay, K., Hales, J. and Chava, S., 2020. Hyperbole or Reality? Investor Response to Extreme Language in Earnings Conference Calls. The Accounting Review, 95(2), pp.31-60.

Bollen, J., Mao, H. and Zeng, X., 2011. Twitter Mood Predicts the Stock Market. Journal of Computational Science, 2(1), pp.1-8.

Bond Jr, C.F., Kahler, K.N. and Paolicelli, L.M., 1985. The miscommunication of deception: An adaptive perspective. Journal of Experimental Social Psychology, 21(4), pp.331-345.

Borochin, P., Cicon, J., DeLisle, R. and Price, S., 2017. The effects of conference call tones on market perceptions of value uncertainty. Journal of Financial Markets, 40, pp.75-91.

Borth, D., Ji, R., Chen, T., Breuel, T. and Chang, S.F., 2013, October. Large-scale visual sentiment ontology and detectors using adjective noun pairs. Proceedings of the 21st ACM international conference on Multimedia (pp. 223-232).

Boussaidi, R., 2013. Representativeness heuristic, investor sentiment and overreaction to accounting earnings: The case of the Tunisian stock market. Procedia-Social and Behavioral Sciences, 81, pp.9-21.

Bowden, J., 2019. Examining Relationships between Online Financial Message Board Interactions and Trading Activity of Securities Listed on the AIM Submarket of the London Stock Exchange. University of Dundee.

Bowden, J., Kwiatkowski, A. and Rambaccussing, D., 2019. Economy through a lens: distortions of policy coverage in UK national newspapers. Journal of Comparative Economics, 47(4), pp.881-906.

Bowen, R.M., Davis, A.K. and Matsumoto, D.A., 2002. Do conference calls affect analysts' forecasts? The Accounting Review, 77(2), pp.285-316.

Bowers, J.W., 1963. Language intensity, social introversion, and attitude change. Communications Monographs, 30(4), pp.345-352.

Bradac, J.J., Mulac, A. and House, A., 1988. Lexical diversity and magnitude of convergent versus divergent style shifting: Perceptual and evaluative consequences. *Language and Communication*, *8*(3-4), pp.213-228.

Brockman, P., Li, X. and Price, S., 2015. Differences in Conference Call Tones: Managers vs. Analysts. *Financial Analysts Journal*, 71(4), pp.24-42.

Brooke, M.E. and Ng, S.H., 1986. Language and social influence in small conversational groups. *Journal of Language and Social Psychology*, *5*(3), pp.201-210.

Brown, G.W. and Cliff, M.T., 2005. Investor sentiment and asset valuation. *The Journal of Business*, *78*(2), pp.405-440.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. Advances in neural information processing systems, 33, pp.1877-1901.

Buelow, M.T., Hupp, J.M., Porter, B.L. and Coleman, C.E., 2020. The effect of prosody on decision making: Speech rate influences speed and quality of decisions. Current Psychology, 39, pp.2129-2139.

Burgoon, J.K., Birk, T. and Pfau, M., 1990. Nonverbal behaviors, persuasion, and credibility. Human communication research, 17(1), pp.140-169.

Cambria, E. and White, B., 2014. Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. IEEE Computational Intelligence Magazine, 9(2), pp.48-57.

Cao, S., Jiang, W., Yang, B. and Zhang, A., 2020. How to Talk When a Machine is Listening: Corporate Disclosure in the Age of AI. National Bureau of Economic Research.

Chan, C., Bajjalieh, J., Auvil, L., Wessler, H., Althaus, S., Welbers, K., van Atteveldt, W. and Jungblut, M., 2021. Four best practices for measuring news sentiment using 'off-the-shelf' dictionaries: a large-scale p-hacking experiment. Computational Communication Research, 3(1), pp.1-27.

Chan, S.W. and Chong, M.W., 2017. Sentiment analysis in financial texts. Decision Support Systems, 94, pp.53-64.

Chan, W.S., 2003. Stock price reaction to news and no-news: drift and reversal after headlines. Journal of Financial Economics, 70(2), pp.223-260.

Chattopadhyay, A., Dahl, D., Ritchie, R. and Shahin, K., 2003. Hearing voices: The impact of announcer speech characteristics on consumer response to broadcast advertising. Journal of Consumer Psychology, 13(3), pp. 198–204.

Chebbi, S. and Ben Jebara, S., 2018. On the use of pitch-based features for fear emotion detection from speech. *2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)* [Preprint].

Chen, J., Nagar, V. and Schoenfeld, J., 2018. Manager-analyst conversations in earnings conference calls. *Review of Accounting Studies*, 23(4), pp.1315-1354.

Chen, Y., Han, D. and Zhou, X., 2023. Mining the emotional information in the audio of earnings conference calls: A deep learning approach for sentiment analysis of securities analysts' follow-up behavior. International Review of Financial Analysis, 88, p.102704.

Chomsky, N., 2003. On nature and language. Cambridge: Cambridge University Press.

Chowdhary, K. and Chowdhary, K.R., 2020. Natural language processing. Fundamentals of artificial intelligence, pp.603-649.

Chua, G.Y.P., Er, H., Liaw, S. and He, T., 2020. Pitch Right: The Effect of Vocal Pitch on Risk Aversion. Economics Bulletin, 40(4), pp. 3131–3139.

Chen, X.W. and Lin, X., 2014. Big data deep learning: challenges and perspectives. *IEEE access*, *2*, pp.514-525.

Chung, J., Gulcehre, C., Cho, K. and Bengio, Y., 2014 ‘Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modelling’, in Deep Learning and Representation Learning Workshop. Montreal: Neural Information Processing Systems.

Cieslak, A., Morse, A. and Vissing-Jorgensen, A., 2014. Stock Returns Over the FOMC Cycle. NBER Working Paper.

Clementson, D.E., Pascual-Ferra, P. and Beatty, M.J., 2016. How language can influence political marketing strategy and a candidate's image: Effect of presidential candidates' language intensity and experience on college students' ratings of source credibility. Journal of Political Marketing, 15(4), pp.388-415.

Clough, P. and Nutbrown, C., 2012. A student's guide to methodology.

Collins, R.L., Taylor, S.E., Wood, J.V. and Thompson, S.C., 1988. The vividness effect: Elusive or illusory?. Journal of Experimental Social Psychology, 24(1), pp.1-18.

Conley, J.M., O’Barr, W.M. and Lind, E.A., 1978. The power of language: Presentational style in the courtroom. Duke Law Journal, 1978(6), p. 1375.

Core, J.E., 2001. A review of the empirical disclosure literature: discussion. Journal of accounting and economics, 31(1-3), pp.441-456.

Corrado, C.J., 2011. Event studies: A methodology review. Accounting & Finance, 51(1), pp.207-234.

Corredor, P., Ferrer, E. and Santamaria, R., 2013. Investor sentiment effect in stock markets: Stock characteristics or country-specific factors?. International Review of Economics & Finance, 27, pp.572-591.

Coval, J.D. and Shumway, T., 2005. Do behavioral biases affect prices?. The Journal of Finance, 60(1), pp.1-34.

Craig, T.Y. and Blankenship, K.L., 2011. Language and persuasion: Linguistic extremity influences message processing and behavioral intentions. Journal of Language and Social Psychology, 30(3), pp.290-310.

Crotty, M., 1998. The foundations of social research: Meaning and perspective in the research process.

D’Andrea, A., Ferri, F., Grifoni, P. and Guzzo, T., 2015. Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications, 125(3), pp. 26–33.

Dair, Z., Donovan, R. and O'Reilly, R., 2021. Classification of Emotive Expression Using Verbal and Non-Verbal Components of Speech. 2021 32nd Irish Signals and Systems Conference (ISSC).

D'Andrea, E., Ducange, P., Bechini, A., Renda, A. and Marcelloni, F., 2019. Monitoring the public opinion about the vaccination topic from tweets analysis. Expert Systems with Applications, 116, pp.209-226.

Das, S.R. and Chen, M.Y., 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management science, 53(9), pp.1375-1388.

Daudert, T., 2021. Exploiting textual and relationship information for fine-grained financial sentiment analysis. Knowledge-Based Systems, 230, p.107389.

Davis, A. and Tama-Sweet, I., 2012. Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A*. Contemporary Accounting Research, 29(3), pp.804-837

Davis, A., Ge, W., Matsumoto, D. and Zhang, J., 2015. The effect of manager-specific optimism on the tone of earnings conference calls. Review of Accounting Studies, 20(2), pp.639-673.

Davis, A.K., Piger, J.M., Sedor, L.M., 2008. Beyond the numbers: managers' use of optimistic and pessimistic tone in earnings press releases. AAA 2008 Financial Accounting and Reporting Section (FARS) Paper

De Amicis, C., Falconieri, S. and Tastan, M., 2021. Sentiment analysis and gender differences in earnings conference calls. Journal of Corporate Finance, 71, p.101809.

De Long, J.B., Shleifer, A., Summers, L.H. and Waldmann, R.J., 1990. Noise trader risk in financial markets. Journal of political Economy, 98(4), pp.703-738.

Demers, E. and Vega, C., 2008. Soft information in earnings announcements: News or noise?.

DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K. and Cooper, H., 2003. Cues to deception. Psychological bulletin, 129(1), p.74.

Devlin, J., Chang, MW., Lee, K. and Toutanova, K., 2019. Bert: Pre-training of deep bidirectional Transformers for language understanding. [online] arXiv.org.

Dewally, M., 2003. Internet investment advice: Investing with a rock of salt. *Financial Analysts Journal*, *59*(4), pp.65-77.

Dey, L., Chakraborty, S., Biswas, A., Bose, B. and Tiwari, S., 2016. Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier. *International Journal of Information Engineering and Electronic Business*, 8(4), pp.54-62.

Diesner, J. and Evans, C., 2015. Little Bad Concerns. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*.

Doran, J., Peterson, D. and Price, S., 2012. Earnings Conference Call Content and Stock Price: The Case of REITs. *The Journal of Real Estate Finance and Economics*, 45(2), pp.402-434.

Duan, H.K., Hu, H., Yoon, Y. and Vasarhelyi, M., 2022. Increasing the utility of performance audit reports: Using textual analytics tools to improve government reporting. Intelligent Systems in Accounting, Finance and Management, 29(4), pp.201-218.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G. and Ranjan, R., 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), pp.1-33.

Edinger, J.A. and Patterson, M.L., 1983. Nonverbal involvement and social control. Psychological bulletin, 93(1), p.30.

El-Haj, M., Rayson, P., Walker, M., Young, S. and Simaki, V., 2018. Computational Analysis of Financial Narratives: Overview, Critique, Resources and Future Directions. JBFA Capital Markets Conference, Dublin.

Engelberg, J., 2008. Costly information processing: Evidence from earnings announcements. In AFA 2009 San Francisco meetings paper.

Epley, N. and Gilovich, T., 2002. Putting adjustment back in the anchoring and adjustment heuristic.

Erickson, B., Lind, E.A., Johnson, B.C. and O'Barr, W.M., 1978. Speech style and impression formation in a court setting: The effects of "powerful" and "powerless" speech. Journal of experimental social psychology, 14(3), pp.266-279.

Fabozzi, F.J., Ma, K.C. and Oliphant, B.J., 2008. Sin stock returns. The Journal of Portfolio Management, 35(1), pp.82-94.

Fama, E.F., 1970. Efficient capital markets. Journal of finance, 25(2), pp.383-417.

Fama, E.F., Fisher, L., Jensen, M.C. and Roll, R., 1969. The adjustment of stock prices to new information. International economic review, 10(1), pp.1-21.

Fehr, E. and Schmidt, K.M., 2006. The economics of fairness, reciprocity and altruism–experimental evidence and new theories. Handbook of the economics of giving, altruism and reciprocity, 1, pp.615-691.

Feinberg, D.R., Jones, B.C., Little, A.C., Burt, D.M. and Perrett, D.I., 2005. Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. Animal behaviour, 69(3), pp.561-568.

Ferguson, N., Philip, D., Lam, H. and Guo, J., 2015. Media Content and Stock Returns: The Predictive Power of Press. Multinational Finance Journal, 19(1), pp.1-31.

Fisher, I.E., Garnsey, M.R. and Hughes, M.E., 2016. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. Intelligent Systems in Accounting, Finance and Management, 23(3), pp.157-214.

Frankel, R., Jennings, J. and Lee, J., 2022. Disclosure sentiment: Machine learning vs. dictionary methods. Management Science, 68(7), pp.5514-5532.

Frankel, R., Johnson, M. and Skinner, D.J., 1999. An empirical examination of conference calls as a voluntary disclosure medium. *Journal of Accounting Research, 37*(1), pp.133-150.

Frankel, R., Mayew, W.J. and Sun, Y., 2010. Do pennies matter? Investor relations consequences of small negative earnings surprises. *Review of Accounting Studies, 15*, pp.220-242.

Fridman, L., 2019. 'Noam Chomsky: Language, Cognition, and Deep Learning', Lex Fridman Podcast. [Podcast]

Friestad, M. and Wright, P., 1994. The persuasion knowledge model: How people cope with persuasion attempts. Journal of consumer research, 21(1), pp.1-31.

Fu, X., Wu, X. and Zhang, Z., 2019. The Information Role of Earnings Conference Call Tone: Evidence from Stock Price Crash Risk. Journal of Business Ethics 173, 643–660.

Gaertig, C. and Simmons, J.P., 2018. Do people inherently dislike uncertain advice?. Psychological Science, 29(4), pp.504-520.

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E. and Hussain, A., 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 91, pp.424-444.

Gao, B. and Yang, C., 2017. Forecasting stock index futures returns with mixed-frequency sentiment. International Review of Economics & Finance, 49, pp.69-83.

Garcia, D., 2013. Sentiment during Recessions. The Journal of Finance, 68(3), pp.1267-1300.

Gélinas-Chebat, C., Chebat, J.-C. and Vaninsky, A., 1996. Voice and advertising: Effects of intonation and intensity of voice on source credibility, attitudes toward the advertised service and the intent to buy. *Perceptual and Motor Skills*, 83(1), pp. 243–262.

Gentzkow, M., Kelly, B. and Taddy, M., 2019. Text as data. Journal of Economic Literature, 57(3), pp.535-574.

Ghahfarrokhi, A. and Shamsfard, M., 2020. Tehran stock exchange prediction using sentiment analysis of online textual opinions. Intelligent Systems in Accounting, Finance and Management, 27(1), pp.22-37.

Giddens, C.L., Barron, K., Byrd-Craven, J., Clark, K. and Winter, A., 2013. Vocal indices of stress: A Review. *Journal of Voice*, 27(3).

Giles, H. and Smith, P. 1979 Accommodation theory: optimal levels of convergence. In H. Giles and R. St. Clair (Eds), Language and Social Psychology, pp. 45565. Baltimore, MD.

Gillioz, A., Casas, J., Mugellini, E. and Abou Khaled, O., 2020, September. Overview of the Transformer-based Models for NLP Tasks. In 2020 15th Conference on computer science and information systems (FedCSIS), pp. 179-183. IEEE.

Given, L., 2008. The SAGE Encyclopedia of Qualitative Research Methods. DICTION (Software).

Goel, S. and Gangolly, J., 2012. Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. Intelligent Systems in Accounting, Finance and Management, 19(2), pp.75-89.

Goel, S. and Uzuner, O., 2016. Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. Intelligent Systems in Accounting, Finance and Management, 23(3), pp.215-239.

González-Bailón, S. and Paltoglou, G., 2015. Signals of Public Opinion in Online Communication. *The ANNALS of the American Academy of Political and Social Science*, 659(1), pp.95-107.

Gräbner, D., Zanker, M., Fliedl, G. and Fuchs, M., 2012. Classification of customer reviews based on sentiment analysis. *Information and communication technologies in tourism 2012*, pp. 460-470. Springer, Vienna.

Grimmer, J. and Stewart, B., 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), pp.267-297.

Groß-Klußmann, A. and Hautsch, N., 2011. When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance*, 18(2), pp.321-340.

Gu, C. and Kurov, A., 2020. Informational role of social media: Evidence from Twitter sentiment. *Journal of Banking & Finance*, 121, p.105969.

Guo, L., Shi, F. and Tu, J., 2016. Textual analysis and machine leaning: Crack unstructured data in finance and accounting. *The Journal of Finance and Data Science*, 2(3), pp.153-170.

Guth, W., Schmittberger, R. and Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization*, *3*(4), pp.367-388.

Guyer, J.J., Fabrigar, L. and Vaughan-Johnston, T., 2018. The counterintuitive influence of vocal affect on the efficacy of affectively-based persuasive messages. *Journal of Experimental Social Psychology*, 74, pp. 161–173.

Hajek, P. and Munk, M., 2023. Speech emotion recognition and text sentiment analysis for financial distress prediction. Neural Computing and Applications, 35(29), pp.21463-21477.

Harris, M. and Raviv, A., 1993. Differences of Opinion Make a Horse Race. *Review of Financial Studies*, 6(3), pp.473-506.

Henry, E., 2006. Market Reaction to Verbal Components of Earnings Press Releases: Event Study Using a Predictive Algorithm. *Journal of Emerging Technologies in Accounting*, 3(1), pp.1-19.

Henry, E., 2008. Are Investors Influenced By How Earnings Press Releases Are Written? *Journal of Business Communication*, 45(4), pp.363-407.

Hiew, J., Huang, X., Mou, H., Li, D., Wu, Q. and Xu, Y., 2019. BERT-based Financial Sentiment Index and LSTM-based Stock Return Predictability. Cornell University Working Paper,

Hirshleifer, D. and Teoh, S.H., 2003. Limited attention, information disclosure, and financial reporting. Journal of accounting and economics, 36(1-3), pp.337-386.

Hirshleifer, J., 1977. Economics from a biological viewpoint. The Journal of Law and Economics, 20(1), pp. 1–52.

Ho, J.C. and Hung, C.H.D., 2012. Predicting stock market returns and volatility with investor sentiment: Evidence from eight developed countries. Available at SSRN 2279339.

Hoard, J.E., 2002. Language understanding and the emerging alignment of linguistics and natural language processing. In Using Computers in Linguistics (pp. 197-230). Routledge.

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural Computation, 9(8), pp. 1735–1780.

Hollandsworth Jr, J.G., Kazelskis, R., Stevens, J. and Dressel, M.E., 1979. Relative contributions of verbal, articulative, and nonverbal communication to employment decisions in the job interview setting. Personnel Psychology, 32(2), pp.359-367.

Hong, H. and Stein, J.C., 2007. Disagreement and the stock market. Journal of Economic perspectives, 21(2), pp.109-128.

Houjeij, A., Hamieh, L., Mehdi, N. and Hajj, H., 2012. A novel approach for emotion classification based on fusion of text and speech. *2012 19th International Conference on Telecommunications (ICT)*.

Howard, J. and Ruder, S., 2018. Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

Huang, A.H., Wang, H. and Yang, Y., 2023. FinBERT: A large language model for extracting information from financial text. Contemporary Accounting Research, 40(2), pp.806-841.

Humpherys, S.L., Moffitt, K.C., Burns, M.B., Burgoon, J.K. and Felix, W.F., 2011. Identification of fraudulent financial statements using linguistic credibility analysis. Decision Support Systems, 50(3), pp.585-594.

Irani, A.J. and Karamanou, I., 2004. A study of the economic consequences of regulation FD (fair disclosure). Research in Accounting Regulation, 17, pp.191-207.

Jacob, A., 2016. International Conference on Communication and Signal Processing. Speech emotion recognition based on minimal voice quality features.

Jayaraman, J.D. and Dennis, A., 2020. Can Earnings Call Sentiment Predict Stock Price Movement? Proceedings of the Northeast Business & Economics Association.

Jegadeesh, N. and Wu, A., 2012. Word Power: A New Approach for Content Analysis. Journal of Financial Economics, 3(110).

Jiang, F., Lee, J., Martin, X. and Zhou, G., 2019. Manager sentiment and stock returns. *Journal of Financial Economics,* 132(1), pp.126-149.

Jiang, W., 2021. Applications of deep learning in stock market prediction: recent progress. *Expert Systems with Applications*, *184*, p.115537.

Johnman, M., Vanstone, B. and Gepp, A., 2018. Predicting FTSE 100 returns and volatility using sentiment analysis. *Accounting & Finance*, 58(S1), pp.253-274.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H. and Mikolov, T., 2016. Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N. and Wu, Y., 2016. Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410.

Kahneman, D. and Tversky, A., 1984. Choices, values, and frames. American psychologist, 39(4), p.341.

Kahneman, T., 1979. D. kahneman, a. tversky. Prospect theory: An analysis of decisions under risk, pp.263-291.

Kandel, E. and Pearson, N.D., 1995. Differential interpretation of public signals and trade in speculative markets. Journal of Political Economy, 103(4), pp.831-872.

Karpoff, J.M., 1986. A theory of trading volume. The journal of finance, 41(5), pp.1069-1087.

Kartik, N., Ottaviani, M. and Squintani, F., 2007. Credulity, lies, and costly talk. Journal of Economic Theory, 134(1), pp. 93–116.

Katz, J.J., 2002. Mathematics and metaphilosophy. The Journal of philosophy, 99(7), pp.362-390.

Kaushik, L., Sangwan, A. and Hansen, J.H., 2013, May. Sentiment extraction from natural audio streams. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 8485-8489). IEEE.

Kearney, C. and Liu, S., 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33, pp.171-185.

Kennedy, J.A., Anderson, C. and Moore, D.A., 2013. When overconfidence is revealed to others: Testing the status-enhancement theory of overconfidence. Organizational Behavior and Human Decision Processes, 122(2), pp.266-279.

Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S. and Shah, M., 2022. Transformers in vision: A survey. ACM computing surveys (CSUR), 54(10s), pp.1-41.

Kim, S., Lee, J.W., and Kang, H.-G., 2014. Detecting pathological speech using contour modeling of harmonic-to-noise ratio. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [Preprint].

Kimbrough, M.D., 2005. The effect of conference calls on analyst and market underreaction to earnings announcements. The Accounting Review, 80(1), pp.189-219.

Kirchler, E., Maciejovsky, B. and Weber, M., 2010. Framing effects, selective information and market behavior: An experimental analysis. In Handbook of behavioral finance. Edward Elgar Publishing.

Kliger, D. and Kudryavtsev, A., 2010. The availability heuristic and investors' reaction to company-specific events. The journal of behavioral finance, 11(1), pp.50-65.

Klofstad, C.A., Anderson, R.C. and Peters, S., 2012. Sounds like a winner: Voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences*, 279(1738), pp. 2698–2704.

Koolagudi, S.G. and Krothapalli, S.R., 2012. Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features. International Journal of Speech Technology, 15(4), pp. 495–511.

Koolagudi, S.G. and Rao, K.S., 2010. Real life emotion classification using VOP and pitch based spectral features. 2010 Annual IEEE India Conference (INDICON) [Preprint].

Koolagudi, S.G. and Rao, K.S., 2012. Emotion recognition from speech: A Review. International Journal of Speech Technology, 15(2), pp. 99–117.

Korotcov, A., Tkachenko, V., Russo, D.P. and Ekins, S., 2017. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Molecular pharmaceutics*, *14*(12), pp.4462-4475.

Larcker, D. and Zakolyukina, A., 2012. Detecting Deceptive Discussions in Conference Calls. Journal of Accounting Research, 50(2), pp.495-540.

Lee, J.W., Kim, S. and Kang, H.-G., 2014. Detecting pathological speech using contour modeling of harmonic-to-noise ratio. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [Preprint]

Lemmon, M. and Portniaguina, E., 2006. Consumer confidence and asset prices: Some empirical evidence. The Review of Financial Studies, 19(4), pp.1499-1529.

Leetham, A. (2023) *S&P 100 index, Moneyzine. Available at: https://moneyzine.com/investments/sp-100-index/ (Accessed: 21 September 2023).*

Levi, S., 2008. Voluntary disclosure of accruals in earnings press releases and the pricing of accruals. Review of Accounting Studies, 13(1), pp.1-21.

Li, F., 2010. The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach. Journal of Accounting Research, 48(5), pp.1049-1102.

Li, J., Yang, L., Smyth, B. and Dong, R., 2020. MAEC: A Multimodal Aligned Earnings Conference Call Dataset for Financial Risk Prediction. Proceedings of the 29th ACM International Conference on Information &amp; Knowledge Management, pp. 3063–3070.

Li, X., Tao, J., Johnson, M., Soltis, J., Savage, A., Leong, K. and Newman, J., 2007. Stress and emotion classification using Jitter and Shimmer Features. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07 [Preprint].

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In: International Conference on Learning Representations.

London, H., Meldman, P.J. and Lanckton, A.V.C., 1970. The jury method: Some correlates of persuading. Human Relations, 23(2), pp.115-121.

Lopez-Cabarcos, M., M Pérez-Pico, A. and López Perez, M.L., 2020. Investor sentiment in the theoretical field of behavioural finance. Economic research-Ekonomska istraživanja, 33(1), pp.2101-2228.

Loughran, T. and McDonald, B., 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance, 66(1), pp.35-65.

Louis, H., Robinson, D. and Sbaraglia, A., 2008. An integrated analysis of the association between accrual disclosure and the abnormal accrual anomaly. Review of Accounting Studies, 13(1), pp.23-54.

Lucca, D. and Moench, E., 2015. The Pre-FOMC Announcement Drift. The Journal of Finance, 70(1), pp.329-371.

Luong, M.T., Pham, H. and Manning, C.D., 2015. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

MacKinlay, A.C., 1997. Event studies in economics and finance. Journal of economic literature, 35(1), pp.13-39.

Maidiya, B. and Kresta, A., 2024. Impact of Investor Sentiment on Stock Characteristics in Big vs. Small Companies. *Managing and Modelling of Financial Risks,* p.45.

Mairesse, F., Polifroni, J. and Di Fabbrizio, G., 2012, March. Can prosody inform sentiment analysis? experiments on short spoken reviews. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5093-5096). IEEE.

Malkiel, B.G., 2003. The efficient market hypothesis and its critics. Journal of economic perspectives, 17(1), pp.59-82.

Mandrekar, J.N., 2010. Receiver operating characteristic curve in diagnostic test assessment. Journal of Thoracic Oncology, 5(9), pp. 1315–1316.

Mäntylä, M.V., Graziotin, D. and Kuutila, M., 2018. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review, 27, pp.16-32.

Mao, H., Counts, S. and Bollen, J., 2011. Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv preprint arXiv:1112.1051.

Markowitz, H., 1952. The utility of wealth. Journal of political Economy, 60(2), pp.151-158.

Martín-Santana, J.D., Muela-Molina, C., Reinares-Lara, E. and Rodriguez-Guerra, M., 2015. Effectiveness of radio spokesperson's gender, vocal pitch and accent and the use of music in radio advertising. BRQ Business Research Quarterly, 18(3), pp. 143–160.

Matsumoto, D., Pronk, M. and Roelofsen, E., 2011. What Makes Conference Calls Useful? The Information Content of Managers' Presentations and Analysts' Discussion Sessions. The Accounting Review, 86(4), pp.1383-1414.

Mayew, W. and Venkatachalam, M., 2012. The Power of Voice: Managerial Affective States and Future Firm Performance. The Journal of Finance, 67(1), pp.1-43.

Mayew, W.J., Parsons, C.A. and Venkatachalam, M., 2013. Voice pitch and the labor market success of male chief executive officers. Evolution and Human Behavior, 34(4), pp.243-248.

McCroskey, J.C. and Mehrley, R.S., 1969. The effects of disorganization and nonfluency on attitude change and source credibility. Communications Monographs, 36(1), pp.13-21.

McGurk, Z., Nowak, A. and Hall, J., 2020. Stock returns and investor sentiment: textual analysis and social media. Journal of Economics and Finance, 44(3), pp.458-485.

McKay Price, S., Doran, J., Peterson, D. and Bliss, B., 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. Journal of Banking & Finance, 36(4), pp.992-1011.

McMahon, R., 2005. Behavioural finance: a backround briefing. Pusan University National Press.

Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), pp. 1093–1113.

Mehrabian, A. and Williams, M., 1969. Nonverbal concomitants of perceived and intended persuasiveness. Journal of Personality and Social psychology, 13(1), p.37.

Mehrabian, A., 1968. Inference of attitudes from the posture, orientation, and distance of a communicator. *Journal of Consulting and Clinical Psychology*, 32(3), pp. 296-308.

Mendoza, E. and Carballo, G., 1998. Acoustic analysis of induced vocal stressby means of cognitive workload tasks. *Journal of Voice*, *12*(3), pp.263-273.

Merton, R.C., 1973. An intertemporal capital asset pricing model. Econometrica: Journal of the Econometric Society, pp.867-887.

Mian, G.M. and Sankaraguruswamy, S., 2012. Investor sentiment and stock market response to earnings news. The Accounting Review, 87(4), pp.1357-1384.

Milian, J. and Smith, A., 2017. An Investigation of Analysts' Praise of Management During Earnings Conference Calls. Journal of Behavioral Finance, 18(1), pp.65-77.

Miller, E.M., 1977. Risk, uncertainty, and divergence of opinion. The Journal of finance, 32(4), pp.1151-1168.

Miller, N., Maruyama, G., Beaber, R.J. and Valone, K., 1976. Speed of speech and persuasion. Journal of personality and social psychology, 34(4), p.615.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L.T. and Trajanov, D., 2020. Evaluation of sentiment analysis in finance: from lexicons to transformers. IEEE access, 8, pp.131662-131682.

Moffitt, K. and Burns, M.B., 2009. What does that mean? Investigating obfuscation and readability cues as indicators of deception in fraudulent financial reports. AMCIS 2009 Proceedings, p.399.

Morency, L.-P., Mihalcea, R. and Doshi, P., 2011. Towards multimodal sentiment analysis. Proceedings of the 13th international conference on multimodal interfaces [Preprint].

Morrison, D., Wang, R. and De Silva, L.C., 2007. Ensemble methods for spoken emotion recognition in call-centres. Speech Communication, 49(2), pp. 98–112.

Muller, M. (2007) Information Retrieval for Music and Motion Vol 2. Berlin: Sprenger.

Munikar, M., Shakya, S. and Shrestha, A., 2019. Fine-grained Sentiment Classification using BERT. 2019 Artificial Intelligence for Transforming Business and Society (AITB).

Nardo, M., Petracco-Giudici, M. and Naltsidis, M., 2016. Walking down wall street with a tablet: A survey of stock market predictions using the web. Journal of Economic Surveys, 30(2), pp.356-369.

Niederhoffer, V. and Osborne, M.F.M., 1966. Market making and reversal on the stock exchange. Journal of the American Statistical Association, 61(316), pp.897-916.

NIRI (2004). Vienna, VA: National Investor Relations Institute.

Nisbett, R.E. and Ross, L., 1980. Human inference: Strategies and shortcomings of social judgment.

Nogueira, R. and Cho, K., 2020. Passage Re-ranking with BERT. Cornell University Working Paper.

Noroozi, F., Sapiński, T., Kamińska, D. and Anbarjafari, G., 2017. Vocal-based emotion recognition using random forests and decision tree. *International Journal of Speech Technology*, *20*(2), pp.239-246.

Nwe, T.L., Foo, S.W. and De Silva, L.C. 2003. Speech emotion recognition using Hidden Markov models. Speech Communication, 41(4), pp. 603–623.

O'Leary, D.E., 2011. Blog mining-review and extensions: “From each according to his opinion”. Decision support systems, 51(4), pp.821-830.

O'Leary, D.E., 2016. On the relationship between number of votes and sentiment in crowdsourcing ideas and comments for innovation: A case study of Canada's digital compass. Decision Support Systems, 88, pp.28-37.

Pang, B., Lee, L. and Vaithyanathan, S., 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing – EMNLP ‘02 [Preprint].

Park, C.K., Lee, S., Park, H.J., Baik, Y.S., Park, Y.B. and Park, Y.J., 2011. Autonomic function, voice, and mood states. *Clinical Autonomic Research*, *21*, pp.103-110.

Pena, V.A. and Gomez-Mejia, A., 2019. Effect of the anchoring and adjustment heuristic and optimism bias in stock market forecasts. Revista Finanzas y Política Económica, 11(2), pp.389-409.

Peng, L. and Xiong, W., 2006. Investor attention, overconfidence and category learning. Journal of Financial Economics, 80(3), pp.563-602.

Pereira, J., Luque, J. and Anguera, X., 2014. Sentiment retrieval on web reviews using spontaneous natural speech. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4583-4587). IEEE.

Pérez-Rosas, V. and Mihalcea, R., 2013. Sentiment analysis of online spoken reviews. In INTERSPEECH (pp. 862-866).

Petty, R.E., Cacioppo, J.T. and Kasmer, J.A., 2015. The role of affect in the elaboration likelihood model of persuasion. In Communication, social cognition, and affect (PLE: Emotion) (pp. 117-146). Psychology Press.

Poria, S., Gelbukh, A. and Cambria, E., 2015. Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-Level Multimodal Sentiment Analysis. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal: Association for Computational Linguistics., pp. 2539–2544.

Poterba, J.M. and Summers, L.H., 1988. Mean reversion in stock prices: Evidence and implications. *Journal of financial economics*, *22*(1), pp.27-59.

Prakash, J., 2012. A study of weak, semi-strong and strong forms of market efficiency: review of literature. Journal of global research & analysis, 1, p.98.

Qaiser, S., Yusoff, N., Remli, M.A. and Adli, H.K., 2021. A comparison of machine learning techniques for sentiment analysis. Turkish Journal of Computer and Mathematics Education.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1), pp.5485-5551.

Renault, T., 2017. Intraday online investor sentiment and return patterns in the U.S. stock market. Journal of Banking & Finance, 84, pp.25-40.

Renault, T., 2020. Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. Digital Finance, 2(1-2), pp.1-13.

Reyes, R.M., Thompson, W.C. and Bower, G.H., 1980. Judgmental biases resulting from differing availabilities of arguments. Journal of Personality and Social Psychology, 39(1), p.2.

Ribeiro, F., Araújo, M., Gonçalves, P., André Gonçalves, M. and Benevenuto, F., 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science, 5(1).

Rong, J., Li, G. and Chen, Y.-P.P., 2009. Acoustic feature selection for automatic emotion recognition from speech. Information Processing & Management, 45(3), pp. 315–328.

Ryan, S. and Mi, C., 2018. The contribution of confucius to virtue epistemology. Epistemology for the rest of the world, pp.65-76.

Sadique, S., In, F.H. and Veeraraghavan, M., 2008. The impact of spin and tone on stock returns and volatility: Evidence from firm-issued earnings announcements and the related press coverage. Available at SSRN 1121231.

Schmeling, M., 2009. Investor sentiment and stock returns: Some international evidence. Journal of empirical finance, 16(3), pp.394-408.

Scholes, M., 1969. A Test of the Competitive Hypothesis: The Market for New Issues and Secondary Offerings, unpublished Ph. D. DD D (Doctoral dissertation, thesis, Graduate School of Business, University of Chicago).

Seok, S.I., Cho, H. and Ryu, D., 2019. Firm-specific investor sentiment and daily stock returns. *The North American Journal of Economics and Finance*, *50*, p.100857.

Shefrin, H., 2001. Behavioral corporate finance. Journal of applied corporate finance, 14(3), pp.113-126.

Shiller, R., 1989. Market Volatility Cambridge.

Shiller, R.J., 2003. From efficient markets theory to behavioral finance. Journal of economic perspectives, 17(1), pp.83-104.

Siganos, A., Vagenas-Nanos, E. and Verwijmeren, P., 2014. Facebook's daily sentiment and international stock markets. Journal of Economic Behavior & Organization, 107, pp.730-743.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. and Pantic, M., 2017. A survey of multimodal sentiment analysis. Image and Vision Computing, 65, pp. 3–14.

Song, S., Baba, J., Nakanishi, J. and Yoshikawa, Y., 2020. Mind The Voice!: Effect of Robot Voice Pitch, Robot Voice Gender, and User Gender on User Perception of Teleoperated Robots. CHI EA '20: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, New York: Association for Computing Machinery, pp. 1–8.

Sorokowski, P., Puts, D., Johnson, J., Żółkiewicz, O., Oleszkiewicz, A., Sorokowska,A., Kowal, M., Borkowska, B., and Pisanski, K., 2019. Voice of authority: Professionals lower their vocal frequencies when giving expert advice. Journal of Nonverbal Behaviour, 43(2), pp.257–269.

Sprenger, T., Tumasjan, A., Sandner, P. and Welpe, I., 2013. Tweets and Trades: the Information Content of Stock Microblogs. European Financial Management, 20(5), pp.926-957.

Stice, E., 1991. The Market Reaction to 10-K and 10-Q Filings and to Subsequent The Wall Street Journal Earnings Announcements. The Accounting Review, 66(1), pp.42-55.

Stone, P. and Hunt, E., 1963. A computer approach to content analysis. Proceedings of the May 21-23, 1963, spring joint computer conference on - AFIPS '63 (Spring).

Strong, N., 1992. Modelling abnormal returns: A review article. Journal of Business Finance & Accounting, 19(4), pp.533-553.

Sun, C., Shrivastava, A., Singh, S. and Gupta, A., 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision* (pp. 843-852).

Sun, C., Qiu, X., Xu, Y. and Huang, X., 2020. How to Fine-Tune BERT for Text Classification? Chinese Computational Linguistics, pp.194-206.

Sun, L., Najand, M. and Shen, J., 2016. Stock return predictability and investor sentiment: A high-frequency perspective. Journal of Banking & Finance, 73, pp.147-164.

Tabari, N., Seyeditabari, A., Peddi, T., Hadzikadic, M. and Zadrozny, W., 2019. A comparison of neural network methods for accurate sentiment analysis of stock market tweets. In ECML PKDD 2018 Workshops: MIDAS 2018 and PAP 2018, Dublin, Ireland, September 10-14, 2018, Proceedings 3 (pp. 51-65). Springer International Publishing.

Ten Brinke, L., Stimson, D. and Carney, D.R., 2014. Some evidence for unconscious lie detection. Psychological science, 25(5), pp.1098-1105.

Tenney, E.R., Meikle, N.L., Hunsaker, D., Moore, D.A. and Anderson, C., 2019. Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of confidence. Journal of personality and social psychology, 116(3), p.396.

Tetlock, P., 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance, 62(3), pp.1139-1168.

Tetlock, P., Saar-Tsechansky, M. and Macskassy, S., 2008. More Than Words: Quantifying Language to Measure Firms' Fundamentals. The Journal of Finance, 63(3), pp.1437-1467.

Todd, A., Bowden, J. and Moshfeghi, Y., 2024. Text-based sentiment analysis in finance: Synthesising the existing literature and exploring future directions. Intelligent Systems in Accounting, Finance and Management, 31(1), p.e1549.

Troussas, C., Virvou, M., Espinosa, K.J., Llaguno, K. and Caro, J., 2013, July. Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. In *IISA 2013* (pp. 1-6). IEEE.

Tumasjan, A., Sprenger, T., Sandner, P. and Welpe, I., 2010, May. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of the international AAAI conference on web and social media (Vol. 4, No. 1, pp. 178-185).

Turing, A.M., 1950. Mind. Mind, 59(236), pp.433-460.

Tversky, A. and Kahneman, D., 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. science, 185(4157), pp.1124-1131.

Tversky, A. and Kahneman, D., 1981. The framing of decisions and the psychology of choice. *science*, *211*(4481), pp.453-458.

Tversky, A., Kahneman, D. and Slovic, P., 1982. Judgment under uncertainty: Heuristics and biases (pp. 3-20).

Twedt, B. and Rees, L., 2012. Reading between the lines: An empirical examination of qualitative attributes of financial analysts’ reports. Journal of Accounting and Public Policy, 31(1), pp.1-21.

U.S. Securities and Exchange Commission, 2023. Available at: https://www.sec.gov/about/mission.

Van Zant, A.B. and Berger, J., 2020. How the voice persuades. Journal of Personality and Social Psychology, 118(4), pp. 661–682.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I., 2017. Attention Is All You Need. [online] arXiv.org.

Wallbott, H.G., 1982. Contributions of the German "Expression Psychology" to nonverbal communication research: Part III: Gait, gestures, and body movement. Journal of Nonverbal Behavior, 7, pp.20-32.

Wang, N., Kosinski, M., Stillwell, D. and Rust, J., 2012. Can Well-Being be Measured Using Facebook Status Updates? Validation of Facebook's Gross National Happiness Index. Social Indicators Research, 115(1), pp.483-491.

Wang, T.Y., Kawaguchi, I., Kuzuoka, H. and Otsuki, M., 2018. Effect of manipulated amplitude and frequency of human voice on dominance and persuasiveness in audio conferences. Proceedings of the ACM on human-computer interaction, 2(CSCW), pp.1-18.

Wang, X., Lu, S., Li, X.I., Khamitov, M. and Bendle, N., 2021. Audio mining: The role of vocal tone in persuasion. *Journal of Consumer Research*, *48*(2), pp.189-211.

Wang, W. and Hua, Z., 2014. A Semiparametric Gaussian Copula Regression Model for Predicting Financial Risks from Earnings Calls. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. and Rush, A., 2020. Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp.38-45.

Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K. and Morency, L.P., 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intelligent Systems, 28(3), pp.46-53.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. and Klingner, J., 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D. and Zhu, J., 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural language processing and Chinese computing: 8th CF international conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8* (pp. 563-574). Springer International Publishing.

Yan, K., Xu, H. and Gao, K., 2020. CM-BERT. *Proceedings of the 28th ACM International Conference on Multimedia*.

Yang, Y., Mark Christopher, M. and Huang, A., 2020. *Finbert: A pretrained language model for Financial Communications, [online] arXiv.or*

Appendix 3.1 Breakdown of Standard & Poor's 100 Companies

Appendix 3.1 Breakdown of Standard & Poor's 100 Companies

Ticker	Market Cap (Millions)	Share of Market Cap	Number of Calls	ICB Industry	ICB Sector	Location
AAPL.O	2676060	10%	62	Technology	Technology Hardware & Equipment	USA
ABBV.K	281586	1%	33	Health Care	Pharmaceuticals and Biotechnology	USA
ABT	215938	1%	55	Health Care	Medical Equipment and Services	USA
ACN	214596	1%	50	Industrials	Industrial Support Services	ROI
ADBE.O	213836	1%	61	Technology	Software & Computer Services	USA
ADP.O	90180	0%	59	Industrials	Industrial Support Services	USA
AMAT.O	119382	0%	58	Technology	Technology Hardware & Equipment	USA
AMD.O	184641	1%	62	Technology	Technology Hardware & Equipment	USA
AMGN.O	131598	0%	60	Health Care	Pharmaceuticals and Biotechnology	USA
AMT	110566	0%	55	Real Estate	Real Estate Investment Trusts	USA
AMZN.O	1641030	6%	63	Consumer Discretionary	Retailers	USA
ANTM.K	115348	0%	26	Health Care	Health Care Providers	USA
AVGO.O	249219	1%	36	Technology	Technology Hardware & Equipment	USA
AXP	144430	1%	57	Industrials	Industrial Support Services	USA
BA	113844	0%	61	Industrials	Aerospace and Defense	USA
BAC	345982	1%	53	Financials	Banks	USA
BKNG.O	89393	0%	59	Consumer Discretionary	Travel and Leisure	USA
BLK	112297	0%	52	Financials	Investment Banking and Brokerage Services	USA
BMY	154803	1%	53	Health Care	Pharmaceuticals and Biotechnology	USA
C	113121	0%	60	Financials	Banks	USA
CAT	118383	0%	56	Industrials	Industrial Engineering	USA
CHTR.O	99727	0%	44	Telecommunications	Telecommunications Service Providers	USA
CI	76780	0%	56	Health Care	Health Care Providers	USA
CMCSA.O	212654	1%	63	Telecommunications	Telecommunications Service Providers	USA
COST.O	248804	1%	60	Consumer Discretionary	Retailers	USA
CRM	216612	1%	59	Technology	Software & Computer Services	USA
CSCO.O	232509	1%	63	Telecommunications	Telecommunications Equipment	USA
CVS	140832	1%	61	Consumer Staples	Personal Care, Drug and Grocery Stores	USA
CVX	314978	1%	57	Energy	Oil, Gas and Coal	USA
DE	126748	0%	55	Industrials	Industrial Engineering	USA
DHR	209019	1%	53	Health Care	Medical Equipment and Services	USA
DIS	255435	1%	61	Consumer Discretionary	Media	USA
EL	99304	0%	63	Consumer Discretionary	Personal Goods	USA
FB.O	589273	2%	36	Technology	Software & Computer Services	USA
FIS	58411	0%	54	Industrials	Industrial Support Services	USA
GE	105316	0%	61	Industrials	General Industrials	USA
GILD.O	74506	0%	66	Health Care	Pharmaceuticals and Biotechnology	USA
GOOGL.O	1803770	7%	64	Technology	Software & Computer Services	USA
GS	116713	0%	55	Financials	Investment Banking and Brokerage Services	USA
HD	355814	1%	61	Consumer Discretionary	Retailers	USA

HON.O	133477	0%	58	Industrials	General Industrials	USA
IBM	115795	0%	62	Technology	Software & Computer Services	USA
INTC.O	193169	1%	58	Technology	Technology Hardware & Equipment	USA
INTU.O	135713	1%	58	Technology	Software & Computer Services	USA
ISRG.O	104206	0%	54	Health Care	Medical Equipment and Services	USA
JNJ	459762	2%	56	Health Care	Pharmaceuticals and Biotechnology	USA
JPM	413689	2%	54	Financials	Banks	USA
KO	260533	1%	58	Consumer Staples	Beverages	USA
LIN	158244	1%	7	Basic Materials	Chemicals	UK
LLY	273933	1%	58	Health Care	Pharmaceuticals and Biotechnology	USA
LMT	113592	0%	54	Industrials	Aerospace and Defense	USA
LOW	157899	1%	55	Consumer Discretionary	Retailers	USA
LRCX.O	75977	0%	59	Technology	Technology Hardware & Equipment	USA
MA	342205	1%	57	Industrials	Industrial Support Services	USA
MCD	177657	1%	60	Consumer Discretionary	Travel and Leisure	USA
MDLZ.O	85396	0%	55	Consumer Staples	Food Producers	USA
MDT	147811	1%	53	Health Care	Medical Equipment and Services	ROI
MMM	84706	0%	55	Industrials	General Industrials	USA
MO	93916	0%	57	Consumer Staples	Tobacco	USA
MRK	199969	1%	59	Health Care	Pharmaceuticals and Biotechnology	USA
MS	168903	1%	56	Financials	Investment Banking and Brokerage Services	USA
MSFT.O	2252280	8%	62	Technology	Software & Computer Services	USA
MU.O	88922	0%	60	Technology	Technology Hardware & Equipment	USA
NEE	161671	1%	15	Utilities	Electricity	USA
NFLX.O	168972	1%	62	Consumer Discretionary	Media	USA
NKE	207529	1%	59	Consumer Discretionary	Personal Goods	USA
NOW	115976	0%	33	Technology	Software & Computer Services	USA
NVDA.O	663970	2%	63	Technology	Technology Hardware & Equipment	USA
ORCL.K	217935	1%	61	Technology	Software & Computer Services	USA
PEP.O	225212	1%	59	Consumer Staples	Beverages	USA
PFE	307860	1%	56	Health Care	Pharmaceuticals and Biotechnology	USA
PG	359919	1%	62	Consumer Staples	Personal Care, Drug and Grocery Stores	USA
PLD	118518	0%	54	Real Estate	Real Estate Investment Trusts	USA
PM	145529	1%	51	Consumer Staples	Tobacco	USA
PNC	80770	0%	53	Financials	Banks	USA
PYPL.O	138368	1%	21	Industrials	Industrial Support Services	USA
QCOM.O	173547	1%	62	Technology	Technology Hardware & Equipment	USA
RTX	145331	1%	4	Industrials	Aerospace and Defense	USA
SBUX.O	103067	0%	62	Consumer Discretionary	Travel and Leisure	USA
SPGI.K	148849	1%	62	Financials	Finance and Credit Services	USA
SYK	100933	0%	55	Health Care	Medical Equipment and Services	USA
T	165858	1%	61	Telecommunications	Telecommunications Service Providers	USA
TGT	104530	0%	59	Consumer Discretionary	Retailers	USA
TJX	74483	0%	55	Consumer Discretionary	Retailers	USA
TMO	230447	1%	58	Health Care	Medical Equipment and Services	USA
TMUS.O	158885	1%	47	Telecommunications	Telecommunications Service Providers	USA

TSLA.O	935727	3%	40	Consumer Discretionary	Automobiles and Parts	USA
TXN.O	166210	1%	76	Technology	Technology Hardware & Equipment	USA
UNH	476208	2%	56	Health Care	Health Care Providers	USA
UNP	166434	1%	54	Industrials	Industrial Transportation	USA
UPS	191295	1%	54	Industrials	Industrial Transportation	USA
USB	83949	0%	53	Financials	Banks	USA
V	459434	2%	52	Industrials	Industrial Support Services	USA
VZ	213249	1%	62	Telecommunications	Telecommunications Service Providers	USA
WFC	195478	1%	54	Financials	Banks	USA
WMT	400219	1%	50	Consumer Discretionary	Retailers	USA
XOM	333057	1%	61	Energy	Oil, Gas and Coal	USA
ZTS	92459	0%	32	Health Care	Pharmaceuticals and Biotechnology	USA

Notes: This table shows the firms contained within the S&P 100 sample set used for this thesis along with the company's market capitalisation, industry, sector, location of company's headquarters, number of calls associated with each company and the date range these calls are selected from.

Appendix 4.1 Examples of Manually Classified Sentences

Examples of Manually Classified Sentences.

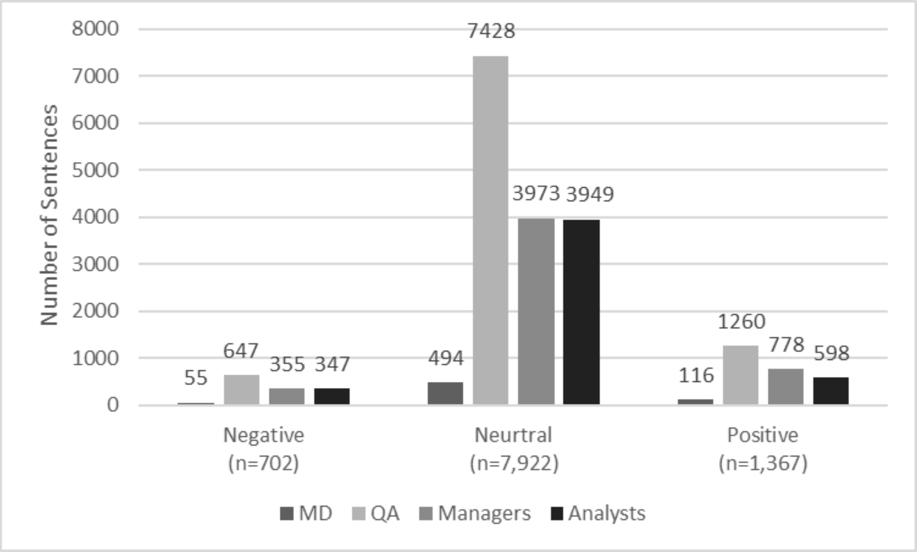
Category	Message
Negative	So, working down some of these bullets here you see the loss for the quarter of $371 million.
Neutral	I wonder, back to the topic of this China sovereign business, if you could give us some colour on the number of customers?
Positive	When we look our overall insulin franchise, we had very good volume growth worldwide.
Complex	You've emphasized the last month or two that (Inaudible). And the reduction we reported at the 35-milligram dose was around 60% reduction in the cerebrospinal fluid A-beta.

Notes: This table shows example messages relating to all three sentiment categories to give the reader an insight of what type of messages are contained within each sentiment category. It also includes two complex examples that related to messages that were hard to classifying during the annotation process.

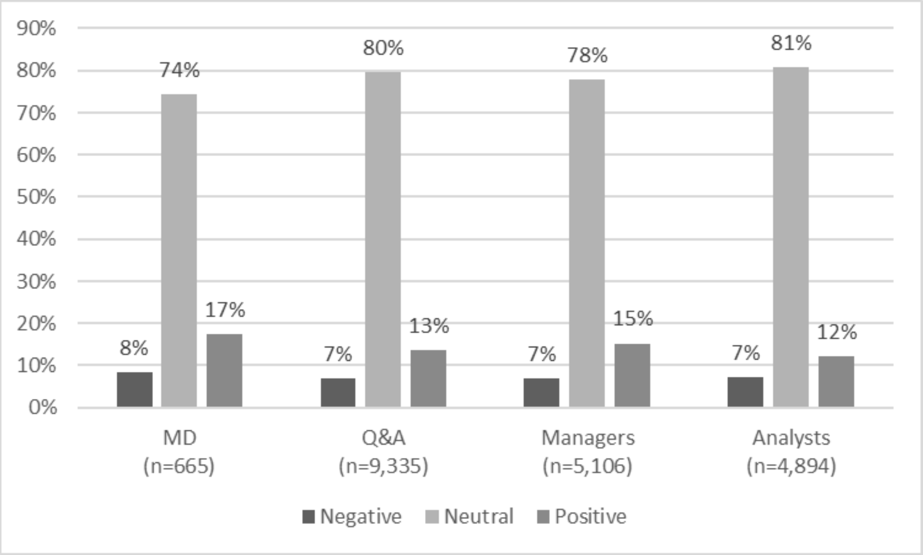
Appendix 4.2 A Breakdown of Sentence Sentiment Relating to Call Sections and Participants

Figure 4.2: Breakdown of Sentiment

Panel A: Breakdown of sentiment identifying the number of sentences in each sentiment category.



Panel B: Breakdown of sentiment identifying the proportion of sentences classified in each category relating to sections and participants on the call.



Notes: These Figures represent the breakdown of sentiment across different sections (Management Discussion and Question & Answer) and participants (Managers and Analysts) on earnings conference calls.

Appendix 4.3 Full List of All Paralinguistic Features Prior to Multicollinearity Testing

Full List of Paralinguistic Features

Feature Type	Features
Pitch	Median Pitch Mean Pitch Minimum Pitch Maximum Pitch
Pulses	Number of Pulses Number of Periods Mean Period Standard Deviation of Period
Voice	Fraction of Unvoiced Number of Voice Breaks Degree of Voice Breaks
Jitter	Jitter (local) Jitter (absolute) Jitter (rap) Jitter(ppq5) Jitter (ddp)
Shimmer	Shimmer (local) Shimmer (dB) Shimmer (apq5) Shimmer (apq11) Shimmer (dda)
Harmonics	Mean Autocorrelation Mean Noise-to-Harmonics Ratio Mean Hormonics-to-Noice-Ratio

Notes: This table gives a full list of all paralinguistic features collected for each sentence in the sample set before the removal of some features due to their high correlation with the final variable set.

Appendix 4.4 Supplementary Overall Call Classification Results

Overall Call Classification Accuracy Results (10K Sample)

	Accuracy	
Model	Training	Validation
Harvard IV4	37.51%	38.05%
Loughran and McDonald (2011)	64.38%	64.25%
Audio Only Classifier	66.94%	66.80%
Naïve Bayes Classifier	89.15%	79.95%
BERT	80.64%	80.85%
FinBERT	82.42%	83.05%
Multimodal FinBERT (NN)	82.51%	83.45%
Multimodal FinBERT (DNN)	86.52%	**84.35%**

Notes: This table outlines the accuracy for each sentiment classifier when the classifier is used to classify the training dataset of 8,000 sentences and the validation dataset of 2,000 sentences. The highest accuracy achieved in each case is signified using bold text.

Overall Call Classification Accuracy Results (2K Sample)

	Accuracy	
Model	Training	Validation
Harvard IV4	40.25%	41.53%
Loughran and McDonald (2011)	66.46%	66.95%
Audio Only Classifier	76.34%	69.49%
Naïve Bayes Classifier	**93.50%**	81.07%
BERT	79.24%	81.36%
FinBERT	79.94%	83.62%
Multimodal FinBERT (NN)	78.95%	84.75%
Multimodal FinBERT (DNN)	85.03%	**86.44%**

Notes: This table outlines the accuracy for each sentiment classifier when the classifier is used to classify the training dataset of 1,600 sentences and the validation dataset of 400 sentences. The highest accuracy achieved in each case is signified using bold text.

Appendix 4.5 Research Paradigms

Prior to evaluating the selected methods employed in this thesis, it is crucial to acknowledge that each individual approach is often rooted in specific values and assumptions that inevitably impact the study. Consequently, it becomes essential to thoroughly examine these methods to elucidate the research decisions that have been made (Clough and Nutbrown, 2012). The two competing theories being evaluated differ in their description of financial markets but also in their philosophical positioning. In evaluating the philosophical positions that underpin these theories it is pertinent to evaluate their ontological and epistemological assumptions. These assumptions can be thought of as pre-commitments surrounding various features of a scholar's work that allow for further understanding and explanation of phenomena (Katz, 2002). Crotty (2003) defines ontology as "the study of being". For the purposes of this analysis a researcher can relate to two ontological positions - realism or relativism.[284] Realists believe that phenomena already exist, and its existence is there to be discovered. However, relativists postulate that the world depends on how an individual experiences it and is different among each person i.e., there is not one 'true' reality.

Epistemology in simple terms is the branch of philosophy related to knowledge. Ryan (2018) specifies that a researcher's epistemological position can take an objective or subjective form. An objective researcher believes that there is only one reality which can be unearthed through the analysis of credible data, further noting that researcher perspective does not influence the study. Opposing these beliefs, a subjective standpoint considers multiple realities, considering each to understand the truth. Thus, inferring that our perceptions, experiences and feelings define our reality.

It is clear from the research conducted on market efficiency that the theory stems from a positivist philosophical position. From an ontological perspective the reality already exists (markets) and Fama (1970) is merely discovering a phenomenon. Hence, the scholar is adhering to a realist position. Furthermore, there is evidence of an objective epistemology. Market efficiency is validated on the basis of results from hypothesis testing. The ontological and epistemological undertones seen in this theory aligns with positivism. Additionally, the nature of the theory strongly aligns with positivism as the theory itself implies that the reality of markets is the same for all agents and that these agents are unable to affect market prices that already reflect all available fundamental information.

Behavioural finance not only contradicts market efficiency in theory but also in its philosophical positioning. Behavioural theory resonates with a relativism ontology unlike traditional theory. Arguing that not all market agents are rational[285] and in fact are sub-rational infers that not all agents are the same. Hence, they do not view financial markets in the same way. Each individual views the market differently based upon their individual views and experiences, aligning with a relativism ontology. The

[284] The author understands there are more than two ontological positions. However, these two ontologies directly relate to the two theories in discussion and hence only these have been mentioned.

[285] See Chapter 1.4 for a definition of a rational market agent otherwise known as "homo-economicus".

epistemological position of behavioural finance can be thought of as subjective. Evidently, with the theory implying that each agent views the market differently, it considers differing and varied perspectives to understand the overall phenomena. Behavioural finance concludes that not all agents are rational and base decisions upon emotions, heuristics and cognitive biases. This falls in line with the subjectivism standpoint, which reiterates that perceptions, experiences and feelings define reality. Behavioural theory therefore aligns with the interpretivism philosophy.

Within finance there are two relevant theories which attempt to explain asset pricing behaviour. This thesis applies sentiment analysis to earnings conference calls to further understand whether sentiment in these calls has any relationship with market prices. The results will return evidence towards either traditional or behavioural theory. If there is a relationship between sentiment and prices that can be profited upon then markets could be described as inefficient. Inferring that the sentiment expressed in earnings call communication is not a factor incorporated into fundamental market prices and that the way in which sentiment information is received by financial market agents does in fact impact their future financial decisions. Hence, agreeing with behavioural theory that agents are not fully rational and calculate emotional and other supposedly irrelevant factors into their decision-making process. On the contrary, the following analysis could produce statistically and economically insignificant results, therefore agreeing with traditional theory, potentially implying sentiment stemming from communication on these calls is already incorporated into a securities fundamental price. Alternatively, sentiment could be an irrelevant factor that does not produce any material information and therefore does not impact security pricing, again agreeing with a traditional perspective.

For the following methods and analysis, it is useful to state that I resonate closer towards positivism than interpretivism. I do believe that each individual agent of the market is not fully rational and that each individual views the market differently. However, in terms of this analysis I believe that there is one reality to be studied and that is the overall market reaction to sentiment. The markets evidently already do exist, and the aim of this research is to better understand them through the use of credible data and hypothesis testing. Furthermore, the results in each of the following chapters will stand or fall by their statistical significance and these results will not impact the financial market being studied. Hence, this research’s ontological and epistemological positions lie closer to realism and objectivism than their counterparts. These positions are commonly associated with the positivism paradigm and allow for this study to investigate investor behaviour and financial market dynamics from an external perspective.