

Social Media Information Credibility in the Context of  
Dementia

By Fatimah Alhayan

Department of Computer and Information Sciences  
University of Strathclyde, Glasgow

A thesis presented in partial fulfilment of the requirements  
for the degree of Doctor of Philosophy

July, 2022

## 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.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Some parts of the work presented in Chapter 5 and Chapter 7 have been published in two different conference proceedings.

Signed: Fatimah Alhayan

Date: July, 2022

## Acknowledgements

First of all, I would like to thank Allah for giving me the strength to complete this journey. Alhumdulillah!

My deepest appreciation to my principal supervisor, Dr.Diane Pennington, for her insightful guidance, support, inspiration and patience. I can not thank her enough for all the opportunities to build my research skills. Without her support, this thesis would not have been achievable. Special thanks to the esteemed Prof.Ian Ruthven for his invaluable input. He always made time to provide me with feedback, and went the extra mile to help me improve.

My deep and sincere gratitude to my parents Nasser and Haifa for their continuous love, care, prayers and motivation throughout my life. Many thanks to my sisters: Noura, Nouf, and Latifa; for their love and support. I would like to express my heartfelt gratitude to my lovely kids, my daughter Raghad, and my son Talal, who has shared this journey since before his birth. A big thank you to my husband for his understanding and encouragement.

A special word of appreciation to all the great friends I have met during this journey, especially Dr.Bayan Al-abdullah, Dr.Dina Al-hammadi, Dr.Maha Alsweilem and Dr.Hanan Himdi. Thank you for having my back, for inspiring me and providing me with motivation when I needed it most.

Sincere thanks to Dr.George Weir for offering access to the Posit tool, which has been used in this research. Thank you to my internal supervisor, Dr.Sara Ayouni for her mentorship and unlimited support.

I would like to thank the Scottish Dementia Research Consortium (SDRC) for financially supporting part of this research, as well as all the dementia organizations in the UK who helped to recruit participants for this study. Last but not the least, I would like to thank my employer, Princess Nourah bint Abdulrahman University for the scholarship enabling me to pursue my PhD degree.

## Publications

Parts of the information covered in this thesis have been published in peer reviewed publications including:

- Alhayan, F., & Pennington, D. (2020, July). Twitter as health information source: exploring the parameters affecting dementia-related tweets. In *International Conference on social media and Society* (pp. 277-290). (Appeared in Chapter 5)
- Alhayan, F., Pennington, D. R., & Ruthven, I. (2022, February). “She Seems More Human”: Understanding Twitter Users’ Credibility Assessments of Dementia-Related Information. In *International Conference on Information* (pp. 292-313). Springer, Cham. (Appeared in Chapter 7)

Other publications during the research that are not directly linked to this thesis include:

- Alhayan, F., Pennington, D. R., & Ayouni, S. (2022). Twitter use by the dementia community during Covid-19: A user classification and social network analysis. *Online Information Review*. Emerald

## Abstract

Credibility is a major concern of social information retrieval (SIR). Absence of editorial oversight and increased bot presence on social media (SM) have led to the spread of low-quality health information, which could be misinterpreted by vulnerable populations (e.g., people with dementia) and their caregivers, affecting their lives negatively. Several studies have proposed automated solutions to evaluate information quality on SM, yet most ignore the role of bots. Another limitation of previous research is that there is a lack of understanding of the features that affect user perception and automation solutions. Automation solutions also often rely on human annotation. Also, previous studies have mainly focused on social events and political topics, with limited focus on the health context. Therefore, the purpose of this research is to explore the credibility aspects, information quality and perceived credibility by humans, in a health context, focusing on dementia information on Twitter.

This research employed a sequential explanatory mixed methods design and conducted three empirical studies in two phases. In the first phase, the research explored bot features in the context of dementia to evaluate the feasibility of using these features to automatically assess the quality of dementia information. Then, different annotated dementia related myth datasets were used to examine the usefulness of varying combinations of features developed in the first study and gleaned from the literature in assessing information quality using a quantitative approach with several supervised machine learning (ML) algorithms. The compiled classification ML model reached an accuracy score of 84% using 28 different linguistic and domain features. These promising results indicate that using the identified features in automatic assessment is feasible. In the second phase, a qualitative approach (using the think-aloud method) was used to identify the most crucial features from user perspectives by analysing people's as-

assessment of information credibility on SM when they were provided access to all the available features on the platform. The findings demonstrate the importance of the qualitative approach to expand understanding of perceived credibility from the user perspective. Users reported unique credibility factors associated with the particular context, and some of these factors are explained using Sundar's MAIN model (Sundar, 2008). Employing this mixed methods design provided a holistic picture of the research problem.

The research findings provide insight into the dissimilarity between information quality evaluated by automated methods and perceived credibility of information evaluated by information consumers. Evaluation by automated methods appears to be based mainly on static features (linguistic cues), whereas user evaluation reveals combinations of static and dynamic features influenced by consumer related characteristics, like prior knowledge and relevance of the information to the consumer. The outcome of this research points to the need for future research to close the gap between human and machine interpretation of credibility. This research concludes by proposing a framework that includes both features evaluated using ML and features based on consumer perception. The framework can be used to develop an automatic assessment model of health information credibility on SM.

# Contents

<b>List of Figures</b>	<b>xi</b>
<b>List of Tables</b>	<b>xii</b>
<b>List of Acronyms</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Research Motivation . . . . .	7
1.2 Research Context . . . . .	9
1.3 Rationale for Selection of the Research Context . . . . .	11
1.4 Research Objectives . . . . .	17
1.5 Research Significance . . . . .	18
1.6 Thesis Structure . . . . .	19
<b>2 Credibility</b>	<b>21</b>
2.1 The Credibility Concept . . . . .	22
2.2 Credibility in the Computer and Information Science Field . . . . .	30
2.3 Credibility Assessment Models and Frameworks . . . . .	31
<b>3 Social Information Retrieval (SIR) Credibility</b>	<b>39</b>
3.1 Introduction . . . . .	39
3.1.1 Low-Quality Information Types . . . . .	41
3.1.2 Types of Sources of Low-Quality Information . . . . .	43
3.2 Approaches Applied in SIR Credibility Literature . . . . .	45
3.2.1 The Human Approach . . . . .	45

## Contents

3.2.2	The Machine Approach . . . . .	52
3.2.2.1	Propagation-Based Methods . . . . .	53
3.2.2.2	Classification-Based Methods . . . . .	54
3.2.3	Hybrid Approach . . . . .	57
3.3	The Health SIR Credibility Literature . . . . .	60
<b>4</b>	<b>Methodology</b>	<b>73</b>
4.1	Research Design . . . . .	73
4.1.1	Mixed Methods Approach Overview . . . . .	75
4.1.2	Rationale for Mixed Methods Design . . . . .	77
4.1.3	Mixed Methods Design Types . . . . .	78
4.1.4	Explanatory Sequential Mixed Methods Design . . . . .	79
4.2	Methods . . . . .	82
4.2.1	Machine Learning Experiment (Phase 1b) . . . . .	82
4.2.1.1	Machine Learning Algorithms . . . . .	82
4.2.1.2	Feature Selection Techniques . . . . .	85
4.2.1.3	Model Evaluations . . . . .	88
4.2.2	Think-Aloud Interviews (Phase 2) . . . . .	92
4.2.3	Content Analysis (Phase 1a and Phase 2) . . . . .	94
4.3	Tools . . . . .	96
4.3.1	Linguistic Inquiry and Word Count (LIWC) . . . . .	96
4.3.2	Posit Toolset . . . . .	98
4.3.3	Botometer as Bot Detection Tool . . . . .	99
4.4	Research Ethics Statement . . . . .	103
<b>5</b>	<b>Study 1: Exploratory Study</b>	<b>105</b>
5.1	Introduction . . . . .	105
5.2	Methods . . . . .	107
5.2.1	Data Collection . . . . .	108
5.2.2	Profile Categorisation . . . . .	109
5.2.3	Profile Feature Analysis . . . . .	111



## Contents

5.2.4	Content Feature Analysis . . . . .	113
5.3	Results . . . . .	115
5.3.1	Descriptive Analysis . . . . .	115
5.3.2	Profile Feature Analysis Result . . . . .	116
5.3.3	Content Features Analysis Result . . . . .	120
5.3.3.1	Linguistic Feature Analysis . . . . .	120
5.3.3.2	Domain-Specific Feature (URLs) Analysis . . . . .	123
5.4	Discussion . . . . .	126
<b>6</b>	<b>Study 2: Machine Learning Experiments</b>	<b>130</b>
6.1	Introduction . . . . .	130
6.2	Methods . . . . .	131
6.2.1	Dementia Myths Selection . . . . .	131
6.2.2	Data Collection . . . . .	133
6.2.3	Data Labelling Guidelines . . . . .	134
6.2.4	Bot Evaluation . . . . .	135
6.2.5	Feature Extraction, Standardisation, and Selection . . . . .	136
6.3	Analysis and Results . . . . .	139
6.3.1	Descriptive Analysis . . . . .	139
6.3.2	Feature Importance Selection . . . . .	144
6.3.3	Machine Learning Classification Analysis . . . . .	152
6.4	Discussion . . . . .	157
<b>7</b>	<b>Study 3 : User Study</b>	<b>161</b>
7.1	Introduction . . . . .	161
7.2	Methods . . . . .	162
7.2.1	Study Design . . . . .	162
7.2.2	Participants and Recruitment . . . . .	164
7.2.3	Study Procedure . . . . .	165
7.3	Data Coding and Analysis . . . . .	165
7.4	Findings . . . . .	166

## Contents

7.4.1	Source . . . . .	167
7.4.1.1	Authority . . . . .	168
7.4.1.2	Identity . . . . .	168
7.4.1.3	Social Presence . . . . .	169
7.4.1.4	Collective Endorsement . . . . .	170
7.4.2	Content . . . . .	171
7.4.2.1	Content Type . . . . .	171
7.4.2.2	Variety . . . . .	172
7.4.2.3	Evidence . . . . .	172
7.4.3	User . . . . .	173
7.4.3.1	Relevance . . . . .	173
7.4.3.2	Prior Knowledge . . . . .	173
7.5	Discussion . . . . .	174
<b>8</b>	<b>Conclusions</b>	<b>178</b>
8.1	Key Research Findings . . . . .	178
8.1.1	The Proposed Framework . . . . .	183
8.1.2	Conversion Model . . . . .	185
8.2	Contributions . . . . .	189
8.3	Practical implications . . . . .	190
8.4	Limitations and Possible Future Work . . . . .	192
	<b>References</b>	<b>195</b>
	<b>Appendices</b>	<b>228</b>
	<b>Appendix Appendices</b>	<b>229</b>
	<b>Appendix A List of LIWC features</b>	<b>229</b>
	<b>Appendix B Example of posit run command</b>	<b>230</b>
	<b>Appendix C Example of Posit output file</b>	<b>231</b>

## Contents

<b>Appendix D</b>	<b>Posit features</b>	<b>232</b>
<b>Appendix E</b>	<b>Botometer versions</b>	<b>233</b>
<b>Appendix F</b>	<b>Example of Botometer API response</b>	<b>234</b>
<b>Appendix G</b>	<b>Ethics approval</b>	<b>235</b>
<b>Appendix H</b>	<b>Consent form</b>	<b>237</b>
<b>Appendix I</b>	<b>Information sheet</b>	<b>238</b>
<b>Appendix J</b>	<b>High-low level search keywords</b>	<b>241</b>
<b>Appendix K</b>	<b>User categorisation codebook</b>	<b>242</b>
<b>Appendix L</b>	<b>Features selected by Anova</b>	<b>243</b>
<b>Appendix M</b>	<b>Features selected by REF</b>	<b>244</b>
<b>Appendix N</b>	<b>Features selected by RF</b>	<b>245</b>
<b>Appendix O</b>	<b>Example of high bot score profile</b>	<b>246</b>
<b>Appendix P</b>	<b>Questionnaire</b>	<b>247</b>
<b>Appendix Q</b>	<b>Task session homepage</b>	<b>250</b>

# List of Figures

1.1	Example of dementia community thread on Twitter (Names and profile pictures of personal accounts are masked for privacy). . . . .	6
1.2	Example of Twitter home timeline . . . . .	14
1.3	Example of a Twitter bot participating in a dementia thread. . . . .	17
4.1	Research phases. . . . .	75
4.2	The typical flow of the sequential explanatory design process. . . . .	80
4.3	Ten-fold cross-validation in ML. . . . .	89
4.4	Confusion matrix in ML. . . . .	90
5.1	Study 1 method overview. . . . .	108
5.2	Bot score-wise distribution in each category. . . . .	117
6.1	Study 2 method overview . . . . .	131
6.2	Examples of Myth 2 from astroturf bot. . . . .	143
6.3	Example of spammer bot profiles . . . . .	144
6.4	Boxplot for selected features . . . . .	147
7.1	Qualitative results. . . . .	167
8.1	Proposed framework for automatically assessing the credibility of dementia information. . . . .	184

# List of Tables

2.1	Theoretical frameworks of web credibility assessment (Choi & Stvilia, 2015). . . . .	37
3.1	User studies in credibility literature (quantitative). . . . .	46
3.2	Relevant studies on quality of health related information on SM (2017-2020). . . . .	63
3.3	Summary of health-related studies in the SIR credibility literature (machine-based). . . . .	69
4.1	Quantitative and qualitative research design characteristics. . . . .	76
4.2	MCC interpretations. . . . .	92
5.1	Descriptions of SMLR measures. . . . .	113
5.2	User categories and bot scores. . . . .	116
5.3	Results of SMLR analysis of total dataset . . . . .	118
5.4	Coefficient and constant values for regression equation . . . . .	118
5.5	Results of SMLR analysis for the profiles with low bot scores . . . . .	119
5.6	Results of SMLR analysis for the profiles with bot-likely profiles . . . . .	119
5.7	Summary of SMLR analysis results of the all three datasets. . . . .	119
5.8	Linguistic analysis of tweets from profiles with non-extreme high/low bot scores. . . . .	121
5.9	Linguistic analysis of tweets from profiles with extreme high/low bot scores. . . . .	122
5.10	URL analysis of tweets. . . . .	124
6.1	Dementia myths. . . . .	133

List of Acronyms

6.2 Descriptive statistical analysis of myths and bots. . . . . 140

6.3 Results of the Pearson correlation coefficient between bot score and bot-  
type score. . . . . 141

6.4 Selected features. . . . . 146

6.5 Summary statistics for class 0 . . . . . 150

6.6 Summary statistics for class 1 . . . . . 151

6.7 Performance of the ML algorithms using different features sets . . . . . 156

6.8 RF classification performance on each topic eliminated from training the  
classifier and then used for testing using 27 selected features. . . . . 157

7.1 Participants' credibility ratings. . . . . 167

8.1 Proposed score (weight) calculation for a feature. . . . . 187

8.2 Proposed criteria for profile credibility evaluation established in Phase 2 188

# Acronyms

## Acronym Definition

<b>AD</b>	Alzheimer's disease
<b>ANOVA</b>	Analysis of Variance
<b>API</b>	Application Programming Interface
<b>CDC</b>	Centers for Disease Control and Prevention
<b>CAP</b>	Complete Automation Probability
<b>CMC</b>	Computer Mediated Communication
<b>CTA</b>	Concurrent Think-Aloud
<b>COVID-2019</b>	Coronavirus Disease of 2019
<b>DT</b>	Decision Tree
<b>ELM</b>	Elaboration Likelihood Model
<b>FN</b>	False Negative
<b>FP</b>	False Positive
<b>HONcode</b>	Health on the Net Foundation's Code of Conduct for Medical Websites
<b>HSM</b>	Heuristic-Systematic Model
<b>IR</b>	Information Retrieval
<b>IS</b>	Information Seeking
<b>KNN</b>	K-Nearest Neighbour
<b>LASSO</b>	Least Absolute Shrinkage and Selection Operator
<b>LIWC</b>	Linguistic Inquiry and Word Count
<b>LR</b>	Logistic Regression
<b>ML</b>	Machine Learning
<b>MCC</b>	Matthews' Correlation Coefficient

## List of Acronyms

<b>POS</b>	Part-of-Speech
<b>PWD</b>	People with Dementia
<b>P-I</b>	Prominence-Interpretation
<b>RF</b>	Random Forest
<b>REF</b>	Recursive Feature Elimination
<b>RTA</b>	Retrospective Think-Aloud
<b>SIR</b>	Social Information Retrieval
<b>SM</b>	Social Media
<b>SPSS</b>	Statistical Package for the Social Sciences
<b>SMLR</b>	Stepwise Multiple Linear Regression
<b>SVM</b>	Support Vector Machine
<b>TMC</b>	Traditional Media Content
<b>TN</b>	True Negative
<b>TP</b>	True Positive
<b>UGC</b>	User-Generated Content



# Chapter 1

## Introduction

‘The internet in general and social media in particular are changing health care’

---

David Perlmutter

Health information credibility on the web is a major concern among healthcare professionals and policy makers (Dalmer, 2017), because low-quality information and possible misinterpretations can have a severe impact, especially on cases of health information related to vulnerable populations such as people with dementia (PWD) (Robillard, Johnson, Hennessey, Beattie, & Illes, 2013). Despite abundant literature considering different aspects of information credibility of early forms of the web (e.g., websites), some aspects remain to be investigated, in particular credibility aspects related to health information on social platforms. There is a lack of knowledge about the features that enable automatic assessment of health information quality on social platforms, the criteria users apply to reach credibility judgements, and how they complement each other. This research fills this gap in credibility research by focusing on quality and consumers’ perception of the credibility of dementia information on Twitter.

The internet has evolved from a static web where users can only consume information to a web where users can also generate information, the social web (Bouadjenek,

Hacid, & Bouzeghoub, 2016). This includes the development of various types of social network sites with different technological affordances, which gained momentum with launching of MySpace in 2003 and continues to this day. Social networks are defined as web-based services that allow users to create a public or semi-public profile within a constrained system, build a list of other users with whom they are connected, and observe and navigate their own list of connections as well as those made by others in the system (Boyd & Ellison, 2007). Other terms, such as “social platforms” and “social media” (SM), are often used to refer to social networks. Thus, throughout this thesis, the terms “social platforms”, “social media”, and “social networks” are used interchangeably.

Social platforms have revolutionised society, culture, lifestyle, and people’s perspective of the world. In 2020, SM was used by almost half of the world population (hootsuite, 2020). The number of social platforms has increased and they have attracted millions of users. In 2021, Facebook, YouTube, WhatsApp, Instagram, and Twitter were the most popular SM sites (Statista, 2022). Social platform usage is an integral part of most people’s daily lives (Y. Zhao & Zhang, 2017). People access information more effectively and personally through SM platforms than they can using traditional search engines (Y. Zhao & Zhang, 2017). However, people are overwhelmed by the amount of information on these social platforms and find it increasingly challenging to verify the credibility of information for their specific needs. In this situation, it is essential plight to allow users to locate relevant information regarding their interests and needs, this task being referred to as information retrieval (IR). Information retrieval is an activity that does not just take place on SM, but also on the wider internet and it is a daily occurrence (Bouadjenek et al., 2016). Classic models of IR are generalised to the internet as a whole and fail to take into account the social context of users and resources (Bouadjenek et al., 2016). The joint study of IR and social networks is the study of social information retrieval (SIR) (Bouadjenek et al., 2016). Social information retrieval is: ‘the incorporation of information about social networks and relationships in the information retrieval process: Social Information Retrieval = Social Networks + Information Retrieval’ (Kirsch et al., 2005, p. 34).

## Chapter 1. Introduction

Information retrieval systems provide users with required information in answer to a query. Searchers evaluate the usefulness or relevance of the information that the IR system retrieves (Tombros, Ruthven, & Jose, 2005). Social information retrieval, on the other hand, refers to a group of IR techniques that assist users in getting the information they need through other users' expert knowledge or search experience (Goh & Foo, 2007). Bouadjenek et al. (2016, p. 3) defined SIR as a 'process of leveraging social information (both social relationships and the social content), to perform an IR task with the objective of better satisfying the users' information needs'. The most significant distinction between the two systems is that traditional IR deals with documents and queries and their interrelationships (Kirsch, Gnasa, & Cremers, 2006). Other users are not involved (Goh & Foo, 2007). Social information retrieval is characterised by the existence of all three: documents, queries, and individuals and their connections with each other (Kirsch et al., 2006).

Information retrieval is assessed for accuracy and the capability to retrieve high-quality information that optimises user satisfaction and meets expectations (Bouadjenek et al., 2016). Relevance and utility concepts are commonly used as a measure of IR effectiveness. Relevance is a topic-relatedness evaluation, which is concerned with whether the topic of a search query matches the topic of a document. Utility is a broad concept that encompasses not only topic relevance but also quality, innovation, importance, credibility, and other factors (Rieh, 2010). With the emergence of SIR, credibility has been recognised as an important concept, due to two unique characteristics of SIR: 1) pinpointing the source of information accurately is challenging and 2) advanced algorithms and machine learning (ML) make decisions in the place of consumers, with limited scrutiny from consumers themselves (Ginsca, Popescu, & Lupu, 2015). Another challenge regarding credibility, however, is that there is no consensus on what constitutes credibility; it is ultimately a complex, intuitive (Rieh, 2010; Flanagan & Metzger, 2008) and multi-dimensional concept (Rieh, Kim, Yang, & St. Jean, 2010).

Over the last few years, interest in the credibility of social information has developed in various academic disciplines, including information science, communications, and psychology. Trustworthiness, expertise, quality and reliability are generally listed as

## Chapter 1. Introduction

components of credibility (Ginsca et al., 2015). Credibility has been viewed objectively or subjectively in various research domains. For example, researchers in fields such as psychology and communication often approach credibility as a subjective concept, a perceived characteristic based on the information consumer’s perspective (Flanagin & Metzger, 2008). Conversely, some researchers in information science, for example, view credibility as an objective attribute of information “quality”, or the degree to which information may be deemed accurate, as determined by established standards or by experts on a given topic (Flanagin & Metzger, 2008). Different concepts related to credibility are discussed in more detail in Chapter 2.

As mentioned earlier, the credibility of social information is a concern in various domains (e.g., marketing, communication, healthcare). In the current study, the focus lies on the credibility of social information related to health. In fact, providing health information on social platforms has several advantages. Social networks benefit healthcare by being a platform where various patients, researchers, caregivers and practitioners connect and share their ideas, independent of their physical locations. For example, social platforms are used for healthcare surveillance and outbreak prediction (Gupta & Katarya, 2020), because the spread of information via social platforms to the public about the outbreak of a disease is fast, cost effective and global compared to traditional methods. Traditional methods are accurate, but suffer from long delays, limited coverage areas and high costs (St Louis & Zorlu, 2012; Ji, Chun, Wei, & Geller, 2015). HealthMap is an example of a real-time digital health surveillance system developed to search and scrape information from news and social platform sites to detect signs of growing threats to public health (Freifeld, Mandl, Reis, & Brownstein, 2008). This provides the information needed to prevent, anticipate, and deal with epidemics. Twitter data has been utilised by various researchers for outbreak predictions of illnesses such as influenza (Byrd, Mansurov, & Baysal, 2016) and syphilis (Young, Mercer, Weiss, Torrone, & Aral, 2018).

Social platforms (e.g., Twitter, Facebook) have been utilised to gauge public reaction to health-related events or to reveal topics discussed during outbreaks (Laaksonen, Jalonen, & Paavola, 2014; Gupta & Katarya, 2020). A study by Fu et al. (2016) ex-

## Chapter 1. Introduction

pressed Twitter users' desire to share Zika virus information, including symptoms, anecdotes of Zika-infected pregnant mothers, and parental concerns. Similarly, Lyu, Le Han, and Luli (2021) examined public topics and sentiments toward COVID-19 vaccination through tweet analysis. The tweets were divided into 16 categories. Opinions on vaccination were the most frequently tweeted out of all 16 topics. Another advantage of using social platforms in healthcare is that it can satisfy patient needs. Twitter, for example, has been shown to be effective in asking people for blood donations (Abbasi et al., 2018). Furthermore, physicians use social platforms for professional, ongoing education and development; research presented at medical conferences; and the continuing exchange of information, efficiently shared within the healthcare community via Twitter (Dizon et al., 2012). Twitter benefits the fields of haematology and oncology by notifying physicians of the publication of new articles and journals (Attai et al., 2017). Similarly, medical organisations post the follow-ups of meetings and findings along with debates on shared interests. Online collaboration on SM, specifically on Twitter, plays a major role in collaboration among participants as well. Twitter hashtags are used by physicians, professionals and users with common interests. A hashtag is a keyword preceded by the # symbol. It is used to index keywords or topics on Twitter to enable people to easily follow topics that interest them. As an illustration, the Radiation Oncology Journal club uses the hashtag #radonc and the International Urology Journal club uses #urojc. They schedule specific times to meet online to discuss articles and their professional practice patterns (Attai et al., 2017). Two other hashtags, #AlzChat and #DiverseAlz, are used by people with different types of dementia and the dementia community at large to discuss dementia related topics. These hashtags chats can function as an efficient source of information for patients and the public (Talbot, O'Dwyer, Clare, & Heaton, 2021). #DiverseAlz is a hashtag enabling PWD, caregivers and researchers to participate in meetings once a week. #DiverseAlz discusses dementia topics such as care, rights, and inclusion. The current researcher was invited to talk about this research in these discussions and to recruit participants for the second study in this research. A screenshot of the thread on Twitter is shown in Figure 1.1.

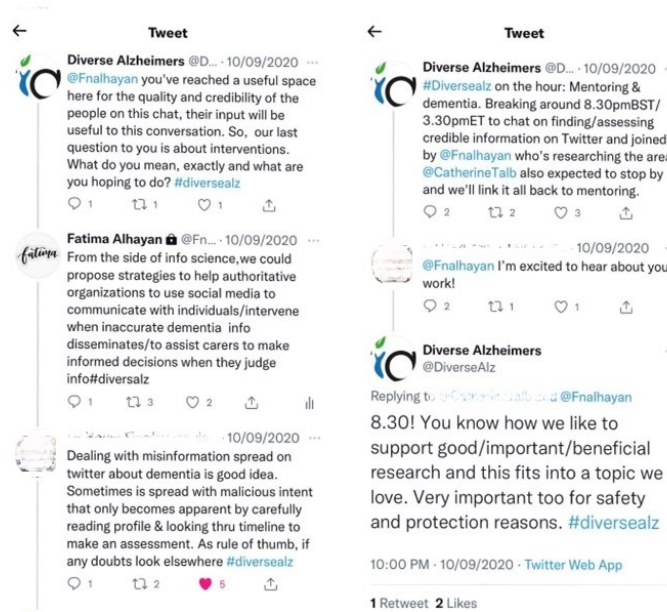


Figure 1.1: Example of dementia community thread on Twitter (Names and profile pictures of personal accounts are masked for privacy).

In the realm of public health education, social platforms have enormous potential to raise awareness; disseminate legitimate, evidence-based information; and counter inaccurate material on the internet (Dizon et al., 2012). Patients with chronic conditions and caregivers seek out information through social platforms. A physician’s presence in the community on social platforms is a vital component that plays an essential role in the treatment process by increasing patient knowledge and decreasing anxiety (Attai et al., 2017). Physicians can benefit from patient experience, correct misconceptions, and ensure accuracy of information (Attai et al., 2017). However, real concerns may hold some medical experts back from embracing these platforms. There is a possibility of loss of control of the information available to patients, or fear about the lack of safety because of inaccurate information. Another issue is regulations and associated risks related to professional or personal reputations on the internet (Dizon et al., 2012). Despite these concerns, it is important for physicians to be present on these platforms.

The healthcare domain thus exploits the power of social platforms for many purposes: to keep track of public health concerns, to develop a picture of the current state

of public health, to detect outbreaks, and to spread information and raise awareness related to health-related issues. Hence, the credibility of information retrieved from these social spaces is crucial, as misleading health information could have serious consequences. Importantly, misinformation usually spreads faster and wider than facts. Certain sources on these platforms, such as commercial interest groups, individuals promoting biased viewpoints (e.g., “fake news”) and bots, can take advantage of this possibility (Attai et al., 2017) and artificially encourage their videos to go viral, by promoting, upvoting, etc., in order to promote certain types of content on a large scale (S. Kumar & Shah, 2018). This further emphasises the need to investigate credibility assessments. The current research’s intent is to provide a comprehensive understanding of credibility, particularly of dementia information on Twitter.

## 1.1 Research Motivation

Firstly, this research uses ML to assess the quality of dementia information on Twitter. The utilisation of ML techniques to assess information quality on social platforms has increased tremendously during the past few years and has proven the potential power of ML in contexts such as events, politics and news. Given that automation capacity is different for each dataset, based on diverse contextual factors, the findings of many empirical case studies are difficult to generalise. The necessity of recognising the fundamental features of health misinformation was emphasised by (Afsana, Kabir, Hassan, & Paul, 2020). Specific features of ML can assist in finding the distinguishing traits between low and high-quality health information, which will facilitate devising effective countermeasures to tackle health misinformation (Afsana et al., 2020). However, in a health context, not enough studies have been undertaken to empirically verify the possibilities to automatically assess health information quality on social platforms. A recent systematic review of studies using SM data and ML algorithms in the health-care domain by (Gupta & Katarya, 2020) indicated that few publications focused on widespread false information posted on SM. Moreover, according to Pasi and Viviani (2020), few publications have addressed the credibility of health information on SM.

## Chapter 1. Introduction

A limited number of studies have begun to investigate health information quality on SM, particularly on topics related to vaccination and health crises (Y. Wang, McKee, Torbica, & Stuckler, 2019), such as the Zika outbreak (Ghenai & Mejova, 2017) or COVID-19 (Abdelminaam et al., 2021).

Most studies also do not address the particular characteristics of those whose goal it is to spread low-quality information regarding health (Y. Wang et al., 2019). For example, the goal of political misinformation may be to gain power, but the goal of medical misinformation could be to increase income for a company (Afsana et al., 2020). Malicious actors such as bots are among the less investigated agents in credibility assessment methods (Qureshi, Malick, & Sabih, 2021) and information quality assessment, despite their significant impact on credibility (Qureshi et al., 2021).

Malicious bots have the ability to deliver content to large audiences, giving them significant opportunity to spread inaccurate information. Although many bot-detection studies have been carried out (Kantepe & Ganiz, 2017; Sayyadiharikandeh, Varol, Yang, Flammini, & Menczer, 2020; Chavoshi, Hamooni, & Mueen, 2016; Beskow & Carley, 2020), these did not focus on credibility (Qureshi et al., 2021). There is thus a clear gap in the study of the automatic assessment of health information quality on social platforms and the role of bots as a likely source of low-quality information on a wide scale. Hence, the current research builds on these common findings that it is imperative to understand the content creator type (bot vs. human) and their participation in sharing information to identify features distinguishing bots from other sources and to utilise these features in information quality assessment.

Secondly, ML is increasingly used to classify or predict information credibility, but credibility cues, that is, features that contribute to user perception of credibility, have not been analysed (Qureshi et al., 2021). Understanding the factors that influence users' assessment of digital health information is important not only for the design of information education programmes, health information content and systems, and patient-provider interactions (Sbaffi, Rowley, et al., 2017), but also to use these features to create algorithms to predict what a typical consumer will perceive as credible or not (Ginsca et al., 2015). Users from different backgrounds, including medical re-



searchers, physicians and caregivers, are susceptible to false information and they may also unintentionally be involved in the dissemination of false information by following fake accounts and liking or reposting messages from malicious bots. This can severely harm their reputation, as well as having bad consequences for individuals' health. The review of (Y. Wang et al., 2019) stated that it is impossible to measure the seriousness and the negative impacts of inaccurate health information, however, it is abundantly clear that more exploration and interdisciplinary research is needed to identify vulnerable (susceptible) users to misinformation. In the health domain particularly, researchers have indicated that only a few publications have focused on how users assess the reliability of health information on SM (Dalmer, 2017; Sbaffi et al., 2017; Keshavarz, 2020), as opposed to many publications on the use of databases and static websites or online searches, which were the topic of studies such as (Eysenbach & Köhler, 2002; Sillence, Briggs, Fishwick, & Harris, 2004; Kammerer, Bråten, Gerjets, & Strømsø, 2013; Liao & Fu, 2014; Klawitter & Hargittai, 2018; Ghenai, Smucker, & Clarke, 2020). As a result, there is a need for research that considers judgements on specific health information on dynamic platforms, such as social networks.

Thus, the current research proposes a framework with comprehensive criteria for evaluating health information credibility on SM. These criteria are derived through automatic methods and criteria reflected by what users consider most important when assessing the credibility of health information on a social platform. The next section discusses the research context.

## 1.2 Research Context

Context has a significant influence on credibility assessment. Contexts can both direct and limit information selection and create a “boundary” around web activities, as well as around users' judgement about what is found (Rieh, 2014). Two contextual factors have been determined to influence credibility judgements, namely topic and medium. The medium is the type of technology platforms (e.g., websites, blogs). Prior research has suggested that credibility evaluation of health information varies from

platform to platform (T. J. Ma & Atkin, 2017). People have a wide range of information needs, hence the information topics vary. People judge credibility based on the type of information and the context (e.g., entertainment or commercial) (Flanagin & Metzger, 2007).

Rieh (2014) studied people's perception about the credibility of a particular topic on platforms with traditional media content (TMC) and user-generated content (UGC). User-generated content differs from TMC in terms of the creation and exchange of content. Content on traditional websites is usually created by certain number of professional users (e.g., [www.flu.gov](http://www.flu.gov)), whereas UGC is usually created by public users. It refers to different types of online platforms such as SM, wikis, and forums (e.g., [www.healthexpertadvice.org/forum](http://www.healthexpertadvice.org/forum)). The authors found that as far as health information is concerned, users considered content on TMC platforms as more trustworthy, accurate and reliable than content on UGC platforms. However, this disparity in believability does not hold for other information domains such as news, travel, or products. Thus, people's belief in the accuracy and reliability of UGC is influenced by the topic in question. Trust in TMC is less affected by the topic (Rieh, 2014). People assign more value to official papers from the government, as an established power, yet in other fields like tourism, people readily rely on reviews of others and personal experience that is not affiliated with certain organisations (Schulz et al., 2022).

Research also pointed to the differences in seeking, sharing and assessing online health information behaviours in particular. De Choudhury, Morris, and White (2014), for instance, investigated users' seeking and sharing behaviours across two different mediums, namely search engines (e.g., Bing) and SM (e.g., Twitter). It was found that there is a relation between consumers' source preferences and the type and severity of health conditions. For example, for severe medical conditions (e.g., multiple sclerosis), users prefer search engines, whereas for mild health conditions and symptoms (e.g., a headache), they use SM.

In general, the two contextual factors; topic and medium, are fundamental in credibility assessment. A credibility assessment model is not completely applicable in all contexts; however, features generalisation can be investigated in different contexts. For

this research, the platform Twitter (as medium) and the topic dementia were selected.

### **1.3 Rationale for Selection of the Research Context**

The rapid increase in the ageing population worldwide has made dementia a major public health concern, with new cases occurring every three seconds, affecting 55 million people worldwide (WHO, 2021). People with dementia access social platforms like blogs, Facebook, and Twitter to connect with others, seek support and share information (Rodriquez, 2013; Craig & Strivens, 2016; Talbot, O'Dwyer, Clare, Heaton, & Anderson, 2020; Mackie, Mitchell, & Marshall, 2019). Caregivers also take part in decision making (Mackie et al., 2019) and increasingly turn to the web for information and support. Sixteen out of 31 studies show the positive impact of SM interventions and tools (e.g., increased knowledge, satisfaction, and involvement) on informal caregivers of critically ill patients (Cherak et al., 2020). In the UK, 66% of dementia caregivers utilise the internet for dementia information, while 76% use SM to combat isolation by keeping in touch with friends and family (French, 2016). According to the findings of a systematic review conducted by Egan et al. (2018) to identify studies of internet based interventions for informal caregivers of PWD (e.g., programs designed to support or train caregivers such as icare), there is evidence that Internet-based interventions can enhance different outcomes for informal caregivers of PWD mental health and supportive outcomes include: depressive and anxiety symptoms, caregiver knowledge, self-efficacy. Another study by Egan et al. (2022) found social networking sites in the top three out of 16 technologies used by caregivers who cared for people with different conditions, with dementia as the most common condition. Dementia caregivers use social media such as blogs to seek and share information and social support (Anderson, Hundt, Dean, Keim-Malpass, & Lopez, 2017). On Facebook, dementia related groups allow caregivers and PWD to share daily activities, situations, emotions and experiences (Bachmann, 2020). On Twitter, dementia is among the top five most discussed health conditions (Z. Zhang & Ahmed, 2019). People with dementia use it for fundraising, lobbying, awareness raising, educating, providing support, challenging stigma, shar-

ing their lived experiences, and advocating for social change (Talbot, O'Dwyer, Clare, Heaton, & Anderson, 2020). Caregivers use Twitter to expand their social networks, obtain support, learn about support services (Danilovich, Tsay, Al-bahrani, Choudhary, & Agrawal, 2018), and share their caregiving experience (Al-bahrani, Danilovich, Liao, Choudhary, & Agrawal, 2017). The findings of (Alhayan & Pennington, 2020) emphasised the engagement of stakeholders, from patients to physicians, in the Twitter dementia community. Twitter has emerged as a powerful tool for professional communication and for disseminating medical information to the public (Oltulu, Mannan, & Gardner, 2018). However, there are impacts in the quality of available information regarding health and care of older adults, potential misinterpretation (Robillard et al., 2013) and associated economic burdens. Therefore, research is required to comprehensively understand the credibility of dementia information on social platforms, and how typical dementia information consumers (dementia caregivers) assess the credibility of dementia-related information.

### **Dementia**

Dementia and Alzheimer's disease are a leading cause of death in the UK, comprising 12.8% of all deaths in 2018 (ONS, 2019). Dementia is an umbrella term under which various illnesses and conditions resulting in progressive and irreversible diminishing of cognitive abilities are grouped (Astell, Dove, & Hernandez, 2019). Alzheimer's disease, defined as the increasing inability to acquire new memories, is the most frequent cause of dementia. People with dementia can recall past incidents and people better than those of the present, and cannot learn new information easily (Astell et al., 2019). There are many dementia subtypes, including frontotemporal dementia (FTD) and primary progressive aphasia (PPA). Each subtype has its own cognitive and behavioural outline (Astell et al., 2019). For instance, FTD has two forms: frontal and temporal. The former exhibits more behavioural than cognition changes: disinhibition, risk taking and indifference to others. The latter, also referred to as semantic dementia, is the inability to recognise everyday items or word meanings. This means that individuals with various forms of dementia suffer from different challenges in their daily lives, which necessitates a wide range of solutions and interventions to assist them (Astell et al., 2019). There

is currently no cure or treatment available for these dementias in general (Astell et al., 2019). People with dementia eventually need increasing amounts of assistance and care as they become unable to live alone (S. Shu & Woo, 2021). Researchers have developed a wide range of technological interventions to assist PWD and their caregivers through all stages (S. Shu & Woo, 2021). These range from wearable technology aids, like heart rate monitors and fall detection, to GPS, to smart home technologies to simplify daily routines, for example controlling lights and switches, set thermostats, view security cameras and many more. Social media platforms are frequently discussed as another important technological tool for facilitating education and awareness (S. Shu & Woo, 2021; L. I. Castillo, Hadjistavropoulos, & Brachaniec, 2021).

### **Twitter**

Social media platforms have a range of technical features in common; however, each has some unique features. This section describes Twitter and the most important Twitter technical features and terminologies used in this research.

- Twitter is a social platform that allows people to communicate by exchanging short, frequent messages called tweets<sup>1</sup>. A tweet may contain a mixture of text, photos, and videos. In 2018, the length of a Tweet was doubled from 140 to 280 characters: still brief yet enabling more expression. A tweet may include specific Twitter features such as a mention, which is to include another person's username. This can be added anywhere in the body of the tweet by adding the @ symbol in front of the username. A hashtag, with the # symbol placed in front of a keyword, serves to index keywords or topics on Twitter to enable people to easily follow topics that interest them. A tweet may also include the user's actual location (i.e., GPS coordinates) from where the tweet was sent.
- The Twitter profile or Twitter user (in this study sometimes referred to as Twitter source) uses account metadata for features such as profile description, location, and picture, while choosing which information to share publicly. A Twitter user can post tweets to their profile, which is set to be public by default unless the

---

<sup>1</sup><https://help.twitter.com/en/resources/new-user-faq>



Figure 1.2: Example of Twitter home timeline <sup>2</sup>

user chooses to make it private to those they follow only. A Twitter user may also “retweet” another user’s tweet to their own followers, either by simply reposting the exact tweet, or by attaching it to another tweet with their comments. Users may also “reply” to another user’s tweet. A user can find and “follow” accounts whose tweets interest them. A source can also “quote” other users’ tweets and be mentioned in tweets and be grouped into lists.

- The timeline of a user is a homepage that shows a stream of tweets that have been published by the users they have chosen to follow. An example of a user’s home timeline is shown in Figure 1.2.
- Twitter can also show recommendations of whom to follow. Twitter recommends tweets to users based on people or topics they already follow. These recommendations can appear in the form of notifications or be added to profile’s home timeline <sup>3</sup>.
- A Twitter bot is a software program that automates the process of tweet posting,

<sup>2</sup><https://twitter.com/twittersupport/status/1501989523588358145>

<sup>3</sup><https://help.twitter.com/en/using-twitter/twitter-timeline>

retweeting and replying to tweets or following users. These automated programs have the primary aim to diffuse information in the form of news, business promotions, political dissemination, and help during emergencies (Chu, Gianvecchio, Wang, & Jajodia, 2012). Apart from humans and bots, a third type of Twitter user is called a cyborg, which refers to human-assisted bots or bot-assisted humans. Cyborgs are common on Twitter and may be set to post tweets in the absence of humans; this is different from bots in the sense that bots are completely automated, whereas cyborgs have characteristics of both manual and automated behaviour (Chu et al., 2012). Bots serve a range of purposes, for example, legitimate bots usually deliver news and update feeds, while malicious bots spread spam, try to influence public perception about a topic, or spread misinformation. A bot may also promote websites and tempt users to click on links. While some types of bots, “legitimate bots”, such as for instance weather bots and help or chat bots, are permitted on Twitter, they should not violate Twitter’s policies and rules. Twitter’s policies, for example, expressly forbid the use of its service for spam or abusive activity, such as making threats to others or impersonating other accounts (Twitter, 2017a).

Twitter strongly opposes malicious bots, which are set to spread malicious content, and suspends them. In 2016, Twitter announced that it had suspended 235,000 accounts for violating its policies and promoting terrorism (Twitter, 2016). In 2018, Twitter suspended more than 70 million accounts which it had identified as fake (Timberg & Dvoskin, 2019). Twitter usually suspends accounts based on a variety of spam-fighting tools and user reports (Aleroud, Abu-Alsheeh, & Al-shawakfa, 2020); Twitter offers an option to users to report profiles or tweets that violate its rules and policies (Twitter, 2017b). Despite some victories in bot detection by Twitter, there is still much more to be done against malicious bots (Golberg, 2017; Albadi, Kurdi, & Mishra, 2019). In a study by Saha Roy et al. (2020) shows that Twitter fails to detect tweets that include phishing URLs. The authors created multiple dummy Twitter accounts that continually posted malicious URLs for almost a month before they were manually reported by other

Twitter accounts, indicating that Twitter failed to detect them (Saha Roy et al., 2020). Manual reporting may take a long time, putting users at risk prior to the suspension of malicious accounts. The malicious content of these accounts, whether produced by humans or bots, should be deleted before their followers are exposed to it and fall victim to it (Saha Roy et al., 2020).

An example of a malicious Twitter bot participating in the dementia community specifically is shown in Figure 1.3. This researcher had a personal encounter with a bot participating in the dementia community while conducting this research. When the researcher posted a recruitment flyer for the third study of this research on Twitter in 2020 with the hashtag #dementia, a reply to that tweet was posted within seconds. The message in the reply was an invitation to join a Facebook group called “Obviate dementia with diet”. It was clear that it was a bot, tracking the #dementia hashtag to post automatic replies. This happened despite Twitter’s policy prohibiting the sending of automated replies to Tweets based on keyword searches or specific hashtags and automatic likes of Tweets (Twitter, 2017a). A glance at the Facebook profile revealed what looked like a new profile without any posts. A screen capture of the reply was taken at the time. The researcher looked up the bot profile on Twitter again in 2021, and found that the profile name, picture and post topics had completely changed since the previous time it was checked (Figure 1.3). This illustrates how complicated bots can quickly change behaviour. This confirms that strategies currently employed by Twitter appear to be insufficient. What makes the detection and mitigation more complex is that bots are continuously changing and evolving to evade detection (Sayyadiharikandeh et al., 2020). Therefore, computer and data science researchers have been designing advanced methods to automatically detect bots or differentiate between humans and bots, with various degrees of accuracy. The most common bot detection tool is Botometer, introduced in Chapter 2.



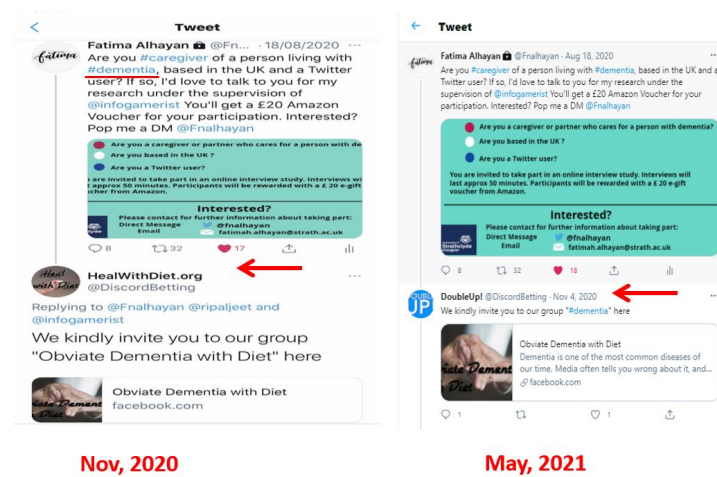


Figure 1.3: Example of a Twitter bot participating in a dementia thread.

## 1.4 Research Objectives

The main goal of this research is to develop a sophisticated and comprehensive understanding of information credibility related to dementia on SM. More precisely, it investigates two credibility aspects, namely quality and perceived credibility, of dementia related information on Twitter. The research proposes a framework that combines features of both aspects that help to assess the credibility of health information on SM. In this context, the research objectives are as follows:

1. To explore the bot role in information dissemination, since the presence of a bot can be a quality indicator, and to identify features of bot-like profiles.
2. To investigate the possibility of using bot-like features together with other features previously defined in the literature to assess the quality of dementia related content using the ML approach.
3. To enhance understanding of credibility factors that influence consumers of dementia information.

To address these objectives, this research answers the following research questions:

**RQ1:** What profile types participate in dementia-related discussions on Twitter?

(Study 1)

**RQ2:** Are there bot activities in the context of dementia information dissemination on Twitter? If so, what is the relationship between bot patterns and different profile types? (Study 1)

**RQ3:** What profile features and content features contribute most to demonstrating bot-like behaviour? (Study 1)

**RQ4:** To what extent, if at all, do the most active bots contribute to spreading low-quality dementia-related information on Twitter? Which bot types have the greatest involvement in the spread of low-quality dementia-related information on Twitter? (Study 2)

**RQ5:** What are the most effective features to improve the automated assessment of dementia information quality? (Study 2)

**RQ6:** What are the factors used by information consumers to assess the credibility of dementia information on Twitter? (Study 3)

## 1.5 Research Significance

Due to the variance in the linguistic, semantic and structural patterns of information features among different domains, there is a need to characterise the credibility of domain-specific information to address the underlying features of health information in comparison to other types of information (e.g., politics) (Afsana et al., 2020). Moreover, the persuasion disparities in the nature of parties associated with misinformation; Some types being more- and others less persuasive, what they have in common and where they differ, should also be considered, so that appropriate action can be taken to combat misinformation (Afsana et al., 2020). Therefore, the importance of this study lies in its attempt to enrich the current literature by investigating two main aspects of credibility of SIR: information quality and perceived credibility, in the little investigated context of dementia.

The results of this research lead to a framework to develop automatic methods for assessing dementia information credibility, incorporating digital features of both information quality and user perception features. Most of the existing automatic solutions

using ML have two major limitations: bot features tend not to be utilised to assess and explain the information quality, and criteria affecting user perceptions are rarely integrated into automatic methods. Only once there is a better understanding of how credibility is evaluated and which features affect this process can an effective technical solution for SM consumers be provided (Meinert, Aker, & Krämer, 2019). Knowing the features that end users employ to make credibility decisions will improve automatic credibility classification by, for instance, recommending changes to feature weightings in ML approaches or guiding the choice of features to emphasise (or downplay) the task of labelling training data for supervised learning (Morris, Counts, Roseway, Hoff, & Schwarz, 2012).

The development of automation methods will make it possible to investigate a greater scope of information to gain a complete picture of the information available on a specific health-related topic. Additionally, it is essential for health organisations to build an observation tool to track the credibility of information that PWD, caregivers and the general public are exposed to.

### 1.6 Thesis Structure

The remainder of the thesis is organised as follows:

**Chapter 2** starts by reviewing the main credibility related components, namely expertise, trustworthiness, and information quality, followed by the specific components studied in this research and how they are defined. The chapter also discusses the research directions in credibility literature, followed by the existing theoretical frameworks and models on credibility developed mainly for information on websites.

**Chapter 3** presents a detailed review of the research in the SIR credibility assessment literature in general and the health context in particular.

**Chapter 4** introduces and justifies the methodology of this research, which is mainly based on sequential explanatory mixed methods using different data collection methods. Ethical considerations of the research are also provided.

**Chapter 5** presents the first study, investigating Twitter profile types that were

## Chapter 1. Introduction

involved in the transmission of dementia information. It assesses the existence of bots among various groups and the distinguishable features of bots. This is done using different statistical tests and analysis.

**Chapter 6** presents the second study, discussing the role of bots in disseminating dementia related myths. The chapter also presents an empirical evaluation of features revealed in the first study and features from the prior literature to assess the information quality using ML algorithms.

**Chapter 7** presents the third study, investigating the essential features that make consumers consider dementia information on Twitter credible. This is done using a qualitative approach.

**Chapter 8** concludes this thesis and summarises the main contributions. It draws on the full thesis to propose a framework for evaluating health information on SM. It also discusses the research's limitations and some directions for future work.

### Chapter Summary

This chapter introduces an overview of the problem of evaluating the credibility of information on social platforms. It highlights the motivation of this research and describes the research context. Additionally, it discusses the objectives, questions, and importance of the research.

## Chapter 2

# Credibility

Credibility is an integral component of both SIR and information seeking (IS). As mentioned earlier in Chapter 1, if IR processes information from social platforms, it is called SIR (Bouadjenek et al., 2016). The difference between IR and SIR lies in the fact that traditional IR is about the interaction of a user with an information system (other users are not involved), whereas SIR takes advantage of the knowledge of other users when obtaining information. The former may be compared to a solo endeavour, the latter is more of a collaborative effort (Goh & Foo, 2007). The IR system mainly refers to the ranking method and it involves indexing and similarity scoring of documents (Ginsca et al., 2015). It is critical to ensure people have access to both relevant and credible information that does not corrupt their perception of reality (Petrocchi & Viviani, 2022; Lioma, Simonsen, & Larsen, 2017). Therefore, advances in IR are required to examine and address the problem of false information, by providing users with automatic tools to assess the credibility of the information they are accessing (Basu, Ghosh, & Ghosh, 2018; Petrocchi & Viviani, 2022; Lioma et al., 2017). On the other hand, IR research has emphasised that user perceptions and their credibility judgements during IS are prejudiced in many ways (Kattenbeck & Elswailer, 2019). Understanding how user judgements are reached and what affects their biases is also crucial, for example, to design systems that improve critical assessment of information. Therefore, as mentioned in Chapter 1, the primary goal of this research is to investigate credibility from both an automatic perspective and a user perspective. This chapter presents the back-

ground information for this research by introducing the credibility concept and related aspects like trustworthiness, expertise, quality, and perceived credibility, and briefly discusses how these aspects can be measured by computer and information science researchers. Contrasting perspectives on the notion of credibility among researchers are then presented. Lastly, a definition is adopted to conceptualise credibility in the current research (Section 2.1). This is followed by the different directions the credibility literature has taken (Section 2.2). The chapter also reviews and compares existing credibility frameworks and models (Section 2.3).

### 2.1 The Credibility Concept

Credibility is a complex and intuitive concept that does not have one clear definition and is closely related to several other concepts, such as ‘believability, trustworthiness, fairness, accuracy, trustfulness, factuality, completeness, precision, freedom from bias, objectivity, depth, and informativeness’ (Rieh, 2010, p.1337). The two concepts most prominently linked to credibility and clearly documented as relevant in literature are expertise and trustworthiness (Rieh, 2010; Ginsca et al., 2015), which are related to the source of the information.

Expertise refers to the skill, competence, and experience level of the source (B. J. Fogg & Tseng, 1999). Hovland, Janis, and Kelley (1953) defined expertise as the degree to which the source is perceived to be able to provide valid information. Source expertise is mainly focused on the knowledge of the source (Ginsca et al., 2015). Various methods have been proposed in the literature to use computational methods to find expert users (Gonçalves & Dorneles, 2019) in blogs, on forums, and in online question and answer (Q&A) communities by analysing user content or examining direct or indirect links between participants, which is usually done by graph modelling. For example, V. Kumar and Pedanekar (2016) used a graph-partitioning approach to find experts in the Stack-Exchange Q&A community and categorised the expertise level into expert, beginner, and novice. A graph of user-to-question interactions was constructed, indicating that a user answering questions posed by other users is regarded as a user with expertise. The

study gives a normalised weight to each user as a fragment of the total of best answers that the user provides. The answer with the greatest number of votes is considered as the best answers. A study by L. Yang et al. (2013) combined topical interests of users and link analysis to measure users' topical expertise. Topical interests are modelled based on the history of user posts using algorithms such as Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), whereas link analysis is based on the relationships between users, questions and answers in the Q&A community using algorithms such as PageRank (Page, Brin, Motwani, & Winograd, 1999). It is important to note that most attempts to measure expert users regard expertise as the ability of a source to answer inquiries or have an opinion on a specific topic regardless of their goal or the quality of a single content item (Ginsca et al., 2015). Studies assessing source expertise on social platforms using machine approaches are discussed in more detail in Section 3.2.2.

Trustworthiness, on the other hand, has a moral dimension and refers to the goodness of the source (Tseng & Fogg, 1999). Trustworthiness refers to whether the source is unbiased and truthful (B. J. Fogg & Tseng, 1999; Rieh, 2010). It mainly focuses on the intent of the source (Ginsca et al., 2015). Hovland et al. (1953) defined trustworthiness as the willingness of the source to provide valid information. Trustworthy sources are an indicator of the credibility of information. In computer science, most methods to assess trustworthiness place an emphasis on authority (Gallwitz & Kreil, 2021) as well as influence. However, in the world of SM, explicit, external authority may be absent, so other cues, such as links and content, are needed to estimate the trustworthiness of a source. The terms "trust" (Ginsca et al., 2015), "influence" and "opinion leaders" are usually used in the assessment of the source trustworthiness on SM (H. Zhao et al., 2018). Ranking users in social networks using algorithms such as PageRank (Page et al., 1999) is an example of a method commonly used to measure the authority, influence and trustworthiness of nodes in the network (H. Zhao et al., 2018). For example, H. Zhao et al. (2018) used motif-based PageRank to measure the authority of the nodes and then ranked them correspondingly. Motifs refer to sub-graphs containing few nodes in complex networks. More studies assessing source trustworthiness on social platforms

using machine approaches are reviewed in Section 3.2.2.

The assessment of online information credibility is regarded as an indirect evaluation of information quality, solidly grounded in source expertise and trustworthiness (Choi & Stvilia, 2022). However, Ginsca et al. (2015) emphasises the significance of both the content and the observable behaviour of the source and extends the credibility concept to include content-related components such as quality and reliability as essential. This is especially the case in today’s dynamic digital terrain, where information sources are much more difficult to pinpoint with absolute accuracy (Sundar, 2008).

Information quality is defined as the degree to which information is useful and “fit for use” in a given task or context (Juran et al., 1992; R. Y. Wang & Strong, 1996). A piece of information is “good” if it is fit for purpose and “bad” if it is not (Ginsca et al., 2015). This definition has been applied to information on social platforms by scholars such as (Washha, 2018). Assessing “good” or “bad” information is subject to quality measurement. Information quality assurance, checking, or filtering is a process that ensures that information fulfils certain context-specific quality requirements. The filtering phase goes one step further and removes the data that does not fit the requirements (Ivanov, 1972). As a result, information that does not fulfil set quality requirements is labelled as low quality; otherwise, it is regarded as high quality. On Twitter, a post containing a phishing link, for example, is considered low quality because it does not match Twitter’s posting information requirements (Washha, 2018). In the context of social platforms, Washha, Qaroush, Mezghani, and Sèdes (2017) staggered the information quality process in three phases. First, selection of a dataset that requires improvement (e.g., tweets). Next, identifying the type of noise (e.g., spam, rumour) to be removed, and finally, based on the noise type identified, pre-design algorithms to generate noise-free data sets. The computer science literature on information quality has focused on text quality analysis and on spam as an indicator of bad quality (Ginsca et al., 2015). Text quality analysis includes the complexity of the writing style or the readability level, quantifying the syntactic and lexical features. Spam does not only refer to email, but to any form of unwanted communication (Ginsca et al., 2015), including SM posts by malicious bots. Since part of the current research is concerned



with information quality, a text quality analysis is conducted and the presence of bots is considered as an indicator of low quality. Different text quality assessment methods were applied on social platforms, like ML and natural language processing (NLP), which is discussed in detail in Section 3.1.2. Bot detection approaches and tools are also discussed in detail in Section 4.2.3.

On the other hand, reliability usually indicates the extent to which something is interpreted as dependable and consistent in quality (Lankes, 2007). As defined by Ginsca et al. (2015), a temporal aspect is added to the content quality via reliability; in other words, consistency or predictability is content quality observed over time. For example, web pages' content reliability assessment is usually estimated according to the quality of their information over time.

To conclude, four general components of credibility are identified in literature: expertise and trustworthiness, referring to the source, and quality and reliability, referring to content.

Aside from the field of computer science, credibility is also studied in many other fields and it is to a great extent interdisciplinary (Flanagin & Metzger, 2008; Metzger & Flanagin, 2015). Scholars in various fields, including information science, management information systems (MIS), human computer interaction (HCI), communication studies, marketing, and psychology have studied credibility assessment for a variety of objectives (Danielson & Rieh, 2007). Credibility research began in the 20th century, when psychologists studied persuasion as part of propaganda operations during the World Wars (Rieh, 2010). Psychology researchers study how source credibility affects persuasion. In the field of communication, researchers have been particularly interested in the demands imposed by the medium, as various types of modalities (e.g., text, audio, video) limit attention and memory in different ways (Danielson & Rieh, 2007). Media credibility evolved from professional news organisations' investigation of perceived credibility of newspapers vs. television (Rieh, 2010). In information science, particularly in the research of IS and IR, information and source evaluation have often been discussed in relation to relevance judgements (Rieh, 2002; Danielson & Rieh, 2007; Mierzecka, Wasilewski, & Kisilowska, 2019). Credibility is one relevant criterion that is

frequently reported by users; users decide whether to accept or disregard information based on whether they believe it is relevant to their information problem (Danielson & Rieh, 2007). Researchers have adopted ideas that are comparable to relevance yet emphasise various aspects of the information assessment process, such as credibility, information quality, and trustworthiness (Mierzecka et al., 2019). In MIS research, credibility assessment is related to advice provided by information systems, including decision support systems and expert systems, the extent to which users rely on this advice, and the effect of the advice on decision making (Danielson & Rieh, 2007).

It is important to note that there are two different perspectives on credibility among researchers. Fields such as psychology and communication regard credibility as a subjective concept and view it as a perceived characteristic based on the perspectives of information consumers (Flanagin & Metzger, 2008). In fields such as computer and information science credibility is regarded as subjective and objective. Information retrieval researchers focus on how information receivers assess the retrieved document's quality (Danielson & Rieh, 2007). Different people may assess the credibility of the same source or piece of information differently. Credibility judgement is a highly subjective procedure which relies on a person's knowledge, experience, and skill (Rieh et al., 2010).

In general, researchers who regard credibility as a subjective concept have investigated the impact of one or more credibility components (e.g., trustworthiness) on user perceptions. Trustworthiness and expertise have been identified as the two key credibility components in most literature (Tseng & Fogg, 1999). Consequently, a large amount of previous research has examined user credibility perceptions towards source expertise, source trustworthiness or both (McGinnies & Ward, 1980). That is, there is no credibility if either dimension is missing. Users see a website as a credible online source when they believe it has both the intent and the ability to deliver information on a certain topic (Choi & Stvilia, 2022). Trustworthiness is the evaluation of the receiver based on subjective factors. Expertise is based on the recipient's evaluation of the source based largely on objective factors (e.g., the message's source credentials) (Flanagin & Metzger, 2008). Some studies have examined both source aspects, trust-

worthiness and expertise, together in perceived credibility. For example, one study has examined whether source expertise or trustworthiness affect user perception on correcting erroneous inferences about political misinformation (Guillory & Geraci, 2013) and another has done the same for vaccination (Pluviano, Della Sala, & Watt, 2020). The findings of these two studies show the source trustworthiness aspect to be more important than source expertise.

Yet, there is consensus among different scholars that the perception of credibility depends on the evaluation of a variety of aspects simultaneously (Tseng & Fogg, 1999) and it is not limited to only trustworthiness or expertise. People’s credibility assessments can be based on cognitive authority. Cognitive theory, developed by Wilson (1983), shows that successful influence relies on source credibility and that competence and trustworthiness are the two foundations and equally weighted dimensions of that credibility. People can have cognitive authority without being an expert. A person might become a cognitive authority for a certain individual or group of people on a given topic or group of topics. For example, people may ask friends for movie reviews if they trust their judgement, yet their friends are not professional movie critics (Froehlich, 2019). Cognitive authority states that humans can only become aware of a larger world beyond their own experience through being told about it by another. The theory in this field addresses the multi-layered question of who is given the right to speak and consequently, who is deemed worthy of believing. In a global society, the big issue is: Who do we believe and why? Cognitive authority is more about the relationships between people, not who the “experts” are (Wilson, 1983). Those who possess cognitive authority have the power to influence, they are the ones who are believed, and therefore they are thought to be credible sources of information.

Hilligoss and Rieh (2008) conducted information activity diaries and individual interviews with participants to gain a better understanding of the way individuals assess credibility in a broad spectrum of IS scenarios in daily life contexts (e.g., work, school, and personal pursuits). The goal was to develop a credibility assessment framework in relation to people’s IS behaviours. People’s IS behaviours involve the use of many information resources (books, peer-reviewed journal articles, the web, blogs, and

libraries). One of the findings of the study made it clear that different people conceptualise credibility in different ways. Participants conceptualised information credibility referencing multiple aspects, such as truthfulness, trustworthiness, objectivity, and reliability. Participants applied particular constructs of credibility based on the situation or information type they encountered. Another study, by Rieh et al. (2010), found that, while utilising SM, information seekers did not value the author's authority and expertise, but their trustworthiness, reliability, accuracy, and completeness. Kang, Höllerer, and O'Donovan (2015) confirmed that underlying "credibility" of an entity and its perceived credibility are not always the same. Perceived credibility may be considered an subjective aspect of credibility. Based on how the entity is depicted and the qualities of the individual assessing credibility, perceived credibility might differ from (inherent) credibility (Kang et al., 2015). Therefore, credibility is not a characteristic of the information or the source, it depends more on a person's evaluation and perception (Jeon & Rieh, 2014).

Despite the presumed subjectivity of the credibility concept, some researchers in the field of computer and information science do view it as 'an objective property of information "quality", or the degree to which information can be considered accurate, as judged by accepted standards or by experts in a particular domain' (Flanagin & Metzger, 2008, p.141).

As mentioned earlier in this section, quality needs to fulfil certain context-specific quality requirements. In health domain, an example of a standard to regulate the quality of online health information is the Health on the Net Foundation's Code of Conduct for Medical Websites (HONcode) (Boyer, Selby, Scherrer, & Appel, 1998). The HON code is a quality certification given by the HON foundation, a Swiss non-profit organisation. A website is accredited when it meets eight criteria: authorship, complementarity, confidentiality, attribution, justifiability, transparency, financial disclosure, and advertising policy. These criteria are defined as follows: Authorship: medical advice on this site will only be provided by medically qualified professionals. Complementary: information on this site is to support, not replace, relationships between a patient/site visitor and physician. Confidentiality: data and identity of patients and visitors are respected

by the medical website. Attribution: information on this site will be supported by clear references to source data and have specific web links. Dates of modification of pages will be clearly displayed. Justifiability: statements about the benefits or performance of particular treatments, products, or services should be supported by suitable, balanced evidence. Transparency: clear presentation of information and contact addresses will be provided for users needing more information and support. Financial disclosure: clear identification of commercial and non-commercial parties who provided funding, services, or material to this website. Advertising policy: advertising materials will be presented in a manner distinguishable from the original material. Unfortunately, little has been done with this code on the social web, which may reflect the complexity of controlling these fast-growing social networks (Dalmer, 2017).

In the context of credibility of online health information in particular, quality and perceived credibility have been identified as the two most significant concerns since the evolution of the internet (Danielson & Rieh, 2007). Danielson and Rieh (2007, p. 331) stated ‘there are at least two significant issues regarding the credibility of online health information from the consumer’s point of view: One has to do with the quality of online health information and the other with the consumer’s ability to understand the information’. Evaluating the quality of online health information involves the application of criteria to information; in addition to criteria, there is need to understand how customers receive online content (Sun, Zhang, Gwizdka, & Trace, 2019). The quality evaluation checklists is primarily dependent on an expert’s opinion and may not adequately meet the needs of consumers; therefore, consumer behaviour must be considered when designing interventions to enhance quality evaluation in internet searches (Sun et al., 2019).

Online health information quality varies and there is no consensus on what constitutes quality, especially on social platforms. This results in broadly different strategies between users for assessing online health information (Danielson & Rieh, 2007). For example, lay people’s statements of their experiences with a condition may not match the traditional standards for authority, yet they might be deemed authoritative by a person attempting to emotionally cope with a disease (Neal & McKenzie, 2011; Neal,

2010).

Therefore, in this dissertation, credibility is viewed as both an objective attribute of information quality and a subjective attribute of how consumers assess information. The term “quality” is used to assess text features and bot features. The term “perceived credibility” is used to refer to consumers’ credibility assessments, which could result from evaluating different credibility components. (e.g., trustworthiness, expertise) at the same time.

## **2.2 Credibility in the Computer and Information Science Field**

At the point of convergence of computer science, information science and credibility, Ginsca et al. (2015) has found four different research directions in prior studies: analysing, predicting, informing, and effect in IS. Analysing refers to studies that aim to understand features or cues that contribute to users’ perception to assess the credibility of a source or information. These features can be exploited in predicting. Human approaches are usually used in this area; Section 3.2.1 discusses these approaches further. Predicting studies aim to develop models to identify or predict the credibility of the source or content. Machine approaches are usually used in this area; Section 3.2.2 discusses these approaches further. Informing refers to providing credibility information to the user in a way that is both understandable and believable. For instance, Yamamoto and Tanaka (2011) designed a system that displays graphs as radar charts, depicting scores based on different credibility aspects (e.g., authority) on Google’s search engine result pages. Effect is where IR and IS collide. In the context of IS, credibility is seen as a filter or qualifier between the task and user behaviour to achieve the task (Ginsca et al., 2015). Credibility affects how users will use and engage with the information. Therefore, scholars have researched both the characteristics of users searching for information and the characteristics of information sources to comprehend how credibility factors influence a user’s IS behaviour. For example, in the context of health IS behaviour, Crawford, Guo, Schroeder, Arriaga, and Mankoff (2014) used both a survey

and search tasks to understand the impact of individual differences on trust. The study tested different related hypotheses such as the effect of users' preference for institution or peer-produced data on their search task. Information on people's search habits was gathered by asking them about a recent health search. A high correlation was found between trust in webpages and trust in forums.

The current research investigates credibility in two directions: predicting and analysing, thus, Section 3.2 in the next chapter explores human and machine approaches as the most common methods applied in the literature to investigate these two credibility research directions.

### **2.3 Credibility Assessment Models and Frameworks**

The study of web credibility has a long history in the literature. The credibility of earlier forms of the web, in which there were fewer websites, with fewer content creators (e.g., blogs), is well identified and studied. Several models and theories have been developed to understand not only user perception, but also the processes of credibility assessment in IS. This section reviews the six theoretical frameworks that were developed to understand people's perceptions of information credibility. These frameworks were developed to investigate credibility in the context of websites, not specifically for SM.

Wathen and Burkell's Judgment Model (2002) considers web credibility assessment as a staged process. Two processing stages of receiver judgements were identified: surface evaluation and message evaluation. Upon launching a web page, a user evaluates surface aspects in terms of usability (e.g., download speed, website appearance), design (e.g., colours) and information organisation. If this initial evaluation meets a user's satisfaction, the user will move on to the next evaluation stage, considering both the message and its source. Message aspects (accuracy, currency, information breadth, and relevance) combined with source characteristics (expertise and trustworthiness) aid the credibility evaluation. In the final stage, an individual's credibility perception depends on the interaction of the message evaluation and the user's cognitive state.

Another theory about credibility assessment of websites, prominence-interpretation (P-I) theory, was developed by (B. Fogg, 2002). B. Fogg (2002) argues that all known web credibility studies can be understood in the context of P-I. The theory posits that two things take place when credibility is assessed: an individual observes a certain element (prominence) and evaluates it (interpretation). Factors such as user involvement, content, task, experience, and individual differences can affect the prominence, whereas user assumptions (e.g., culture and heuristics), the skill/knowledge of a user (e.g., a user's level of competency in the site's subject matter), and context affect the interpretation. The processes of prominence and interpretation occur multiple times, because new features of the site are constantly observed and interpreted in the process of generating credibility judgements overall.

The dual processing model, by Metzger (2007), considers user motivation and ability in the process of web credibility assessment. The dual processing model follows the main idea of dual models in persuasive contexts such as the elaboration likelihood model (ELM) (Petty & Cacioppo, 1986) and the heuristic-systematic model (HSM) (Chaiken, 1980). In ELM, information is processed via two different routes: (1) a central route, where the user is highly motivated and has the cognitive ability to process the message systematically and (2) a peripheral route, where the user has a low personal motivation and relies on cues of message or source. Some of the cues are the reputation of the source, text colour and the source itself. The HSM makes a similar distinction, where heuristic processing depends on mental shortcuts to heuristics that are previously mentally retained, yet systematic processing concerns a comprehensive analysis of judgement related information.

A unifying framework of credibility assessment was developed by Hilligoss and Rieh (2008), who identified three levels of credibility judgements: construct, heuristics, and interaction. The construct level pertains to how a user defines credibility, where credibility is constructed in terms of reliability, truthfulness, believability, trustworthiness, and objectivity. Heuristics levels involve general rules of thumb that are employed if a person is unmotivated or unable to assess content for quick judgement. Interaction is based on specific source or content cues, peripheral from the information object



(e.g., information appearance) or from the source (e.g., affiliation). All three levels are interlinked and affect each other.

Lucassen and Schraagen's (2011) 3S model was developed to understand how users form judgements about the credibility of information. The model explains three strategies users employ when assessing the information. The model suggested that trust judgements rely on three user characteristics: source experience, domain expertise, and information skills. Applying any of these three characteristics results in different features (strategies) of the information being utilised in forming trust judgements. These features are source features (e.g., authority), semantic features of the information (e.g., accuracy or neutrality), and surface features of the information (e.g., website design), hence the model being named 3S. Domain expertise (or topic familiarity) refers to the user's knowledge of the topic at hand. Information skills are 'the skills required to identify information sources, access information, evaluate it, and use it effectively, efficiently, and ethically' (Julien & Barker, 2009, p. 12). However, the relationship between user and information is difficult to pinpoint. Therefore, the initial version of the model was improved to propose a Revised 3S model (Lucassen, Muilwijk, Noordzij, & Schraagen, 2013) by further investigating the effect of two key user characteristics (domain expertise and information skills of the user) on information evaluation. Domain expertise and information skills were manipulated and systematically controlled in a think-aloud experiment, to gain a deeper understanding of their relationship to trust. Findings showed that those with a higher level of knowledge on a topic concentrate on the semantic features, yet individuals who are not familiar with the topic rely more on surface features. Similarly, information quality is evaluated more by those with higher information skills compared to those with lower skills. The revised 3S model is similar to P-I theory (B. Fogg, 2002), however, an essential addition is that user characteristics are attributed to specific information features.

Source credibility has long been seen as the most crucial factor in determining whether information is deemed believable (Choi & Stvilia, 2015). However, with the development of the web, many information creators have become interconnected, and the lines around credibility have blurred (Choi & Stvilia, 2015). Online, users have to

assess the credibility of both the message and the medium, and they are confronted with an abundance of information and a lack of consistency in content quality (Choi & Stvilia, 2015). In this context, the modality, agency, interactivity and navigability (MAIN) model (Sundar, 2008) focuses on technology affordance that can trigger various cognitive heuristics which affect credibility judgements. Whereas content features, such as headlines, have cues that prompt heuristics, technological features have their own cues that influence user perceptions and processing of content. Affordance provided by technology refers to a certain medium's capability to enable a certain action, and affordances may be found in various degrees in most digital media. The MAIN model divides technological affordances into four different affordances i.e., modality (M), agency (A), interactivity (I) and navigability (N). These affordances contribute to a greater or lesser degree to the credibility assessment of digital media, because they are core structural elements in design. The modality affordance deals with the structural aspects of the medium through which data is presented and is more evident on the surface or in the interface. It includes text, audio and video. The modality affordance can trigger old-media heuristics, for example, if the website's layout matches that of a newspaper, this leads to good credibility assessments. If it looks similar to broadcast media, its perceived credibility would be diminished (Sundar, 2008). The agency affordance deals with the source of information on digital media and can be wide ranging, from websites to a poll of friends on SM, to a person having a profile on an online platform. Various agency heuristics related to affordance are used to identify the source, influencing the perceived credibility of the information given by the source. The agency heuristics defined by the model are an important part of the current research and therefore explained in more detail later in this section. Interactivity includes interaction and activity with digital devices. Interactivity means that the medium is sensitive to the demands of the user and that it can accommodate changes in user input. For instance, the interaction heuristic may be triggered by indications on the interface, particularly in dialog boxes that request user input, resulting in more specificity of the generated content. Navigability focuses on interface cues helping with navigation in cyberspace. For example, the mere appearance of hierarchically ordered hyperlinks on a website may elicit its own

heuristic to influence the credibility perception; well-organised, easily navigable sites are more credible.

The third study of the current research explains the factors that influence user perceptions in the light of the agency affordance. Therefore, agency related heuristics need to be explained further. Six agency related heuristics are proposed by Sundar (2008): authority, identity, bandwagon, machine, social presence, and helper. The authority heuristic is triggered when the source content is a domain expert or an official entity. The identity heuristic is likely to be triggered whenever the user is able to express themselves through manipulating content. The user interface of SM platforms can be designed to generate different verifications of identity, and potential followers may use these for their own evaluation of a profile. The bandwagon heuristic reflects a group endorsement and the popularity of the underlying content and source's reputation. The machine heuristic may be triggered if an interface looks machine-like, causing mechanical traits such as randomness and objectivity to be attributed to its performance/function. The machine heuristic is triggered, for instance, when greater quality is ascribed to a news story if it is believed that a computer, rather than a person, has selected the story, the expectation being that it was done objectively and without ideological prejudice. Social presence heuristics provoke feelings of the presence of another entity. The notion is that the user is interacting with a social entity rather than an inanimate object. The social presence heuristic can toggle with the machine heuristic, with one leading to more positive credibility judgements than the other depending on the nature of the content. Another agency-related heuristic is that of the helper, which is triggered by, for example, online chat bots, because they are considered helpers. Users have by and large responded positively to affect-support agents inhabiting a computer, even when the bad impact that has to be repaired was first produced by the computer (housing the agent). In short, all technological affordances explained in the MAIN model rely on different heuristics, which combine to produce quality attributes about a message and hence help in credibility judgement.

These six proposed theoretical frameworks for web credibility, namely the assessment judgement model (Wathen & Burkell, 2002), prominence-interpretation (P-I)

(B. Fogg, 2002), dual processing (Metzger, 2007), unifying framework (Hilligoss & Rieh, 2008), the 3S model (Lucassen et al., 2013), and the MAIN model (Sundar, 2008) have been reviewed and summarised by (Choi & Stvilia, 2015) and are shown in Table 2.1. These frameworks share common aspects; however, the six individual theoretical models have unique features depending on their focus. Most of the frameworks share the following four major aspects: context, user characteristics, operationalisation, and process. Context means determining if the framework considers contextual factors or not. Credibility is assessed differently depending on the contextual situation (e.g., topic, medium) in which information is consumed. The three frameworks emphasising the importance of the context are the judgement model (Wathen & Burkell, 2002), P-I theory (B. Fogg, 2002), and the unifying framework (Hilligoss & Rieh, 2008). Given that credibility assessment is based mainly on users' perception, most of the models (except the MAIN model (Sundar, 2008)) consider user characteristics, including demographics, user involvement, and information skills, to theorise the process of web credibility assessment. Operationalisation classifies how each model measures information credibility; namely, the type of credibility cues in terms of source, message, and structural characteristics of web resources. Process classifies whether the framework is process based or judgement based. The former illustrates the entire process, whereas the latter focuses on particular factors affecting user perception (Choi & Stvilia, 2015).

To sum up, the proposed theoretical frameworks and models for the credibility assessment of information on the web mainly aid the understanding of factors that influence user perceptions of information credibility, as well as selection processes of information in initial forms of websites including UGC (e.g., blogs), which used to have fewer content creators, unlike social platforms. Although some influential factors of credibility perceptions of web pages can apply to social platforms where there are many creators (e.g., Twitter), many are not applicable. For example, visual features (e.g., graphics and structure of information) affect users' judgement when they search for certain information types on web pages (Rieh, 2002). This may not apply to some social networks (e.g., Twitter), given the structure and public nature of the platforms (Morris et al., 2012). Also, social platforms are often overloaded with information

Table 2.1: Theoretical frameworks of web credibility assessment (Choi & Stvilia, 2015).

Model/Theory	Context	User Characteristics	Operational-ization	Process
P-I Theory (Fogg, 2003)	✓	✓	✓	✓
Judgment Model (Wathen & Burkell, 2002)	✓	✓	✓	✓
MAIN Model (Sundar, 2008)			✓	✓
Unifying Model (Hilligoss and Rieh, 2008)	✓	✓	✓	✓
Dual Model (Metzger, 2007)		✓	✓	✓
Revised-3S Model (Lucassen et al., 2013a)		✓	✓	

and lack cues for source credibility. Therefore, determining whether the information provided on social platforms is accurate might be difficult for users.

The topic of the information provided also influences the overall credibility assessment. Topics differ depending on information needs. For example, the most important information topic for caregivers of PWD is about observing behavioural changes in the patients they care for, such as forgetfulness and repeating questions (Steiner, Pierce, & Salvador, 2016). Information consumers devote their efforts to finding credible information to meet their information needs.

As a result, there is still a need for further investigation to find out if the existing proposed theoretical frameworks can be used as a lens to either explain users' credibility perceptions, and/or to complement an existing theoretical framework when it is used to investigate credibility perceptions in the context of social platforms. This research conducted an empirical study (study 3) examining users' credibility assessment of information on social platforms (Twitter) on the topic of dementia, using a qualitative approach. The results are explained in light of one of these theoretical models( Main Model (Sundar, 2008).

## Chapter Summary

This chapter reviews the main aspects of the credibility concept in the literature. It also presents how credibility is conceptualised by the current research. Two aspects constituting credibility are investigated in the current research, namely information quality and perceived credibility. A classification of the main directions of studies on credibility in the computer and information science literature is also presented. Two directions which are investigated in this research are predicting and analysing. The approaches used for these two directions are discussed in more detail in the next chapter. Finally, most of the six existing theoretical frameworks reviewed in this chapter have been established as the most influential for web information credibility assessment, however, they have been primarily developed and used for assessing early forms of web information in different contexts (Baxter, Marcella, & Walicka, 2019). Most features of these platforms are static, like on other websites, yet SM is characterised by dynamic features (e.g., number of followers). Moreover, the variations in the frameworks exist to meet the needs of credibility operationalisation in various contexts. As new technology and information systems emerge on the web (e.g., SM platforms), there will be a continuous need for understanding credibility cues and heuristics. Thus, the current research investigates whether certain aspects of the existing frameworks can explain the credibility assessments of information in a particular context on a social platform (Twitter) and topic (dementia).

## Chapter 3

# Social Information Retrieval (SIR) Credibility

This chapter starts with discussing the conceptualisation and operationalisation of credibility aspects on social platforms. Then, it reviews different types of low-quality information and their sources, and how this is defined in the health context. The remainder of this chapter considers the main approaches to credibility assessments on social platforms to address different credibility aspects in different contexts Section 3.2 and on health information in particular Section 3.3.

### 3.1 Introduction

As mentioned in Chapter 2, credibility can be assessed from an objective point of view and from a subjective one. Studies that regard credibility as a subjective concept and that investigate consumer perceptions about different credibility aspects are reviewed in Section 3.2.1 below. Studies that regard credibility as an objective concept and that propose different computational (machine) methods to measure different credibility aspects are reviewed in Section 3.2, and 3.3 below.

The current section reviews the literature on credibility in terms of conceptualisation (e.g., key aspects of credibility), and operationalisation (e.g., measures for the aspects) on social platforms. Researchers first had to conceptualise the credibility con-

cept, and then operationalise it through a specific set of metrics and/or heuristics. Through conceptualisation, the key aspects or dimensions of credibility (i.e., expertise or trustworthiness) are identified, while operationalisation indicates the measures that can be used to examine the credibility aspects or dimensions (Choi, 2015). This allows researchers to define the meaning of the concept by converting the theoretical and conceptual variable of interest into a collection of measurements (Choi & Stvilia, 2015). A web credibility measure describes a certain attribute of online-based resources with a number or symbol that can be utilised for systematic and/or objective credibility assessments (Choi & Stvilia, 2015). The next sections review the measures that have been used in the literature to operationalise and measure credibility in the context of social platforms.

The two aspects used most to conceptualise credibility in the literature are “expertise” and “trustworthiness” (Choi & Stvilia, 2015). These two aspects are mostly related to information sources on the web and on social platforms, and they are usually operationalised by utilising properties of the source and the network in the context of social platforms. Section 3.2.2.1 below covers the propagation-based approach and provides more detail about the methods applied to assess expertise and trustworthiness. For example, the expertise of a source can be measured by analysing content or links between nodes or by determining their ability to respond to questions or to express opinions on a given issue, regardless of their intent or the quality of the information they provide (Ginsca et al., 2015). This disregard of intent or quality of provided information is the focus of this research.

The emergence of social platforms poses a challenge, since these platforms display a wide disparity in content quality. As a result, filtering and ranking tasks in such systems are more difficult than in other fields (Agichtein, Castillo, Donato, Gionis, & Mishne, 2008). Another challenge is the large number of users and the amount of information generated on these platforms. However, SIR has certain advantages compared to IR; the complex structure of social platforms means that more information can be provided (Agichtein et al., 2008). Social platforms offer a wide range of user-to-document relation types and user-to-user interactions, along with document content



and link structure (Agichtein et al., 2008). Researchers take advantage of these different types of information to automate the process of assessing the quality of information, especially considering the fast growth of information on the social web. Section 3.2.2 and 3.2.3 provide more details about the methods applied in this regard.

Before discussing the different credibility assessment approaches in Section 3.2 and 3.3, it is necessary to review important terms regarding information quality on social platforms, since this is part of the investigation in the current research. Previously, information quality has been defined in the literature as fitness for use (R. Y. Wang & Strong, 1996) and purpose (Ginsca et al., 2015). The concept of information quality differentiates between low (inaccurate) and high (accurate) quality. As a result, information that does not fulfil a set quality requirements is labelled as low quality; otherwise, it is regarded as high quality (Washha, 2018). Various types of low-quality information are discussed below. These types have been identified and characterised by researchers for the purposes of assessing or creating effective early detection algorithms and tools, on social platforms and on the web.

The following two sections review the general types of low-quality information on SM and their common sources. A definition of low-quality health information for the purposes of this research is also provided and the source types that are important elements in quality analysis are discussed.

### **3.1.1 Low-Quality Information Types**

Low-quality information on social platforms is defined as ‘an umbrella concept that refers to misinformation and subcategories such as disinformation’ (Bastos, Walker, & Simeone, 2021). Low-quality information is misleading or inaccurate information (Cisneros-Velarde, Oliveira, & Chan, 2019). Generally, it refers to any form of written or audio-visual content that has the ability to deceive, confuse, or misinform online decision makers (Bastos et al., 2021).

The term “false information” is commonly used to indicate low-quality information on social platforms and the web. Scholars have classified and identified false information in various ways. It can be classified based on the intention to deceive, which is divided

into two main categories, “misinformation”, meaning false or misleading information, and “disinformation”, meaning false information intentionally spread to deceive people (S. Kumar & Shah, 2018). The intentions behind the spread of mis- or disinformation could be malicious, to gain influence, for financial gain, etc. (Zannettou, Sirivianos, Blackburn, & Kourtellis, 2019). S. Kumar and Shah (2018) classified false information on the web and SM into two main categories: opinion-based, to sway a reader’s opinion or decision (e.g., fake reviews on online products) and fact-based, involving information which goes against, falsifies, or condenses a single fact, including rumours and hoaxes.

Zannettou et al. (2019) defines eight types of false information on the web: Fabricated information: entirely made-up stories with no connection to reality. Propaganda: a type of fabricated story often used to promote a political party, cause or nation state, or to damage it, by e.g., influencing an election result. Conspiracy theories: stories (typically about governments or powerful individuals committing illegal acts) attempting to explain a situation or incident without providing evidence. They frequently ignore reality by distributing unsourced information, or completely avoid an explanation. An example of a conspiracy theory is the idea that the coronavirus is a Chinese-engineered bioweapon. This conspiracy theory has been disseminated bot-style on Twitter since January 2020 (Graham, Bruns, Zhu, & Campbell, 2020). Rumour: stories whose trustworthiness is unclear or never proven. Clickbait: intentional use of deceptive headlines and thumbnails with the purpose of luring visitors to click on a link to a webpage. The truth (or lack thereof) of the headline can only be verified by reading the full content. Satire: mimicking true news stories, using both irony and non sequitur in an effort to share humorous insight (Burfoot & Baldwin, 2009). Biased or one-sided: stories that are highly partisan and prejudiced regarding a specific person or event. Hoaxes: deliberately fabricated stories containing half-truths, masquerading as legitimate facts.

These classifications are an attempt to provide clear definitions and expand on the broad definition of the low-quality information concept to determine the specific features of each type, striving to improve the automated solution (algorithms) to detect them (Molina, Sundar, Le, & Lee, 2021). The false information concept is complicated by the fact that false information could belong to more than one type. For instance,

a rumour could employ clickbait to attract users to view the story (Zannettou et al., 2019). Several types of false information have been extensively investigated in the domain of political and mass communication.

Some types of false information (e.g., propaganda) are more applicable in a political context rather than a health one, and this research focuses on the health domain. Therefore, it is important to shed light on how scholars have previously identified false health related information. In the health context, “rumour” and the broad term “misinformation” are the most common terminologies used to refer to low-quality health information on social platforms. However, the majority of articles use “misinformation” as a collective term to describe different types of inaccurate information, without distinguishing between levels of truth and falsehood to avoid conceptual ambiguity of the term (Y.-J. Li, Cheung, Shen, & Lee, 2019). Chou, Oh, and Klein (2018) defines health misinformation as ‘a health-related claim of fact that is currently false due to a lack of scientific evidence’. Vraga and Bode (2018) adopted a similar definition for health-related misinformation, expanding it to include expert opinion as another characteristic of untrue information: ‘factual matters [that] are not supported by clear evidence and expert opinion’. However, these definitions of health misinformation can overlap with the general definition of disinformation, introduced in the beginning of this section, if the intention behind the misinformation is unknown. It is difficult even for social network administrators and researchers to determine if misinformation was intentionally created or not (L. Wu, Morstatter, Carley, & Liu, 2019). This is why the current research uses the term “myths”. This term follows Chou et al. (2018) and Vraga and Bode’s (2018) definition of “health misinformation” as ‘facts that are presently false without being supported by either scientific evidence or expert opinion’, but this research’s term “myths” specifically takes into consideration that it is unclear whether the information is intentionally or unintentionally propagated.

### **3.1.2 Types of Sources of Low-Quality Information**

As alluded to earlier, it is difficult to determine the intention of the low-quality or “false” information spreader. However, the literature covers a wide range of sources that con-

tribute to low-quality or false information dissemination online. The most common types are described by (Zannettou et al., 2019). These sources could be individuals, for example, individuals who gain personal advantage from spreading the false information (e.g., business owners, politicians); journalists, altering some stories to make their newspaper or websites more popular; true believers and conspiracy theorists, who really believe that they are sharing a truth that other individuals ought to hear about; hidden paid posters, who are paid to post false information to sway people towards certain marketing or social tendencies; and useful idiots, ordinary people misled or influenced by organisations to distribute false information, often largely unaware of the aims behind the influence. Groups can also be behind the spread of false information. Groups include criminal or terrorist organisations sharing false information to attain their goals; governments, who may aim to change public opinion about a specific theme or country; and trolls, who make deliberately offensive or provocative online posts to induce arguments. The worst actors to spread false information on a large-scale use bots, which are created and controlled by an entity (a single individual or a software program) (S. Kumar & Shah, 2018). As mentioned in section 1.3, a Twitter bot is a software program that automates the process of tweet posting, retweeting, replying to tweets or following users. Bots are employed for two major purposes: to speedily disseminate the same information to a broad audience and to inflate the "social standing" of specific users. These tactics make false information seem credible and legitimate (S. Kumar & Shah, 2018). Bots occupy crucial positions in information networks, allowing them to distribute false information (S. Kumar & Shah, 2018). Bots can be run by individuals, terrorist organisations, or any type of source mentioned earlier. Some bots belong to a bot network, which is called a "botnet".

Thus, this research considers the presence of bots as an indicator for low-quality information. The next section discusses different approaches applied in assessing information credibility on social platforms and lists the shortcomings found.

## 3.2 Approaches Applied in SIR Credibility Literature

Different approaches have been applied by researchers to investigate the credibility assessment of information available on social platforms. There are three main approaches: human-based, machine-based, and hybrid. A comprehensive review of these approaches is provided in the following sections.

### 3.2.1 The Human Approach

According to Danielson and Rieh (2007), the procedure to apply the concept of credibility to traditional IR systems and websites can be performed in three different ways. The first and most important is to train individuals to assess information so that they can obtain it from credible sources by employing checklists or the critical thinking approach. Second, web designers who seek to improve a site's credibility might utilise the same assessment criteria as guiding principles. The third focus is on developing IR systems that incorporate multiple aspects of credibility judgements with topic relevance to enhance the search performance. Assessing information credibility and designing systems and websites are two sides of the same coin to ensure useful, reliable and trustworthy data to satisfy the information needs of users (Danielson & Rieh, 2007). Similarly, researchers have adopted human-based approaches, called user studies, to examine people's perceptions regarding information credibility on the social web, mainly to explore specific factors involved in the formation of credibility judgements of information in different contexts and at different levels. This understanding can be used to improve the design of information literacy programmes, health information content and systems, and the relationship between patients and healthcare providers (Sbaffi et al., 2017). The other purpose is to collect credibility ratings from participants in order to build a ground truth dataset which can be utilised for automated solutions (supervised learning, which is discussed in detail in the next section). Crowdsourcing is often used to recruit participants for this purpose.

User studies have been used in different research fields, ranging from computer science (e.g., information science, human-computer interaction) to the social sciences

### Chapter 3. Social Information Retrieval (SIR) Credibility

(e.g., communication and journalism). In communication related research, the focus is more likely on the source and the medium, while the focus of information science research is on the message (Danielson & Rieh, 2007).

A summary of most of the previous studies on credibility using a human-based approach in different contexts is presented in Table 3.1. The credibility of information available on SM is generally assessed at three distinct levels, the post, the topic, and/or the source levels (Alrubaian et al., 2018). Most of the previous studies have designed an experiment to test a theory (Sbaffi et al., 2017), or to test one or more hypotheses quantitatively. Some studies measure a credibility construct as dependent variable after they conceptualise it differently applying sub-constructs (e.g., trustworthiness, competence, goodwill), while many others use a broad definition of credibility. Online questionnaires (surveys) are often used as a data collection instrument for showing different manipulation levels of features, including source, post, and/or topic. Below, the relevant studies are discussed in more detail.

Table 3.1: User studies in credibility literature (quantitative).

Study	No of Features /Hypothesis Used	Theory /Model Driven	Features Manipulation	Study Context	Level	Credibility Dimensions
(Morris et al., 2012)	25 features, 5 people	×	✓	News in politics, science, and entertainment	Topic /Post /Source	×
(Westerman, Spence, & Van Der Heide, 2012)	1 Hypothesis Number of followers, Ratio of followers and follows	SIPT Theory & MAIN Model	✓	H1N1 crisis	Sources	Trustworthiness Competence Goodwill
(J. Yang, Counts, Morris, & Hoff, 2013)	5 Features (gender, user name, profile image, location ,network overlap)	×	✓	Tweets General Health vs Politics	Post	×
(Edwards, Spence, Gentile, Edwards, & Edwards, 2013)	1 Hypothesis Klout Score	SIPT Theory & Main Model	✓	Sources/General-not specific	Source	competence, character, and caring
(Chorley, Colombo, Allen, & Whitaker, 2015)	8 Features (meta data of tweets. 4 quantitative, 4 qualitative)	×	✓	No content -Without revealing the text of the tweet	Post	Friendship
(Lin, Spence, & Lachlan, 2016)	2 Hypothesis (Expert vs Peer, Peer vs Stranger)	MAIN Model	✓	Drug-resistant Gonorrhoea	Source	Trustworthiness Competence Goodwill
(Jahng & Littau, 2016)	2 Hypothesis /Features (social cue and Interactivity)	SIPT & Social presence	✓	Journalist	Source	×
(Shariff, Zhang, & Sanderson, 2017)	3 Features (News type, Year, Trending)	×	×	News (breaking news, natural disaster news and politic news)	Post	×

The earliest user study of credibility perceptions in tweets was by (Morris et al., 2012). Morris et al. (2012) conducted two controlled experiments with the purpose of uncovering user or content features that affected the perception of credibility. A pilot study with five participants revealed 26 relevant features. In the survey, participants were then asked which of these features they usually considered when reading tweets and to give each feature a score. Three features were rated highly, namely message topic, username, and profile image. Follow-up experiments were designed to further assess these features. The experiment manipulated different features of tweets, showing participants the tweets under different conditions and measuring the impact on credibility perceptions. The study found that manipulating usernames had a significant impact on the credibility perception of a profile. The profile image had a bigger impact on the credibility perception of entertainment topics when compared to political or science related topics. As for topic, topics related to science were rated higher than politics and entertainment.

J. Yang et al. (2013) used a survey and experiment to compare the cultural differences in credibility perceptions of information between users from China and the United States. The study focused on Twitter in the USA and Sina-Weibo in China. The experiments included stimuli posts, in which the features of the research interest were manipulated. These features included profile features (profile gender, profile name, profile image, location, and closeness between user and participants) and topic features, such as politics or health. The data revealed that the US used the same username styles (topical usernames such as “Political\_news” vs. internet-style usernames such as “Akalala99”) to assess credibility across a variety of topics, while the Chinese displayed different perceptions depending on whether the content was about politics or health. In addition, Chinese participants rated false tweets as more credible than the U.S. participants. This disparity could be attributed to the socially oriented nature of the way they consume information, information scarcity from other sources due to censorship, or culturally specific microblog services adopted in China (J. Yang et al., 2013).

Westerman et al. (2012) designed an experiment to investigate the impact of a

source's credibility on both the number of followers and the ratio between followers and following, in terms of three source credibility aspects (competence/expertise, trustworthiness and goodwill). The definitions of the first two constructs were introduced in Chapter 2, whereas goodwill points to the degree to which a perceiver believes a source is looking out for their best interest. Mock Twitter profiles were created in which the number of followers and the ratio between followers and following were manipulated and viewed by the participants. It was observed that having either a too low or too high number of followers resulted in reduced credibility in terms of perceived trustworthiness and expertise of the profile. Participants judged profiles with a tight gap between followers and following as competent; yet this ratio had no influence on trustworthiness or goodwill constructs.

Another experiment was designed by Edwards et al. (2013) to examine the impact of system-generated cues using the Klout score, a metric provided by a third party which ranges from 1 to 100 and indicates the user's influence online, on a source's credibility. This experiment investigated three source credibility aspects (competence, character, and caring). However, the study did not have a clear definition of the three aspects, instead adopting the source credibility instrument developed by (McCroskey & Teven, 1999). The study lists examples of items used for each aspect. For example, "intelligent/unintelligent" used for competence, "trustworthy/untrustworthy" used for character, and "cares about me/does not care about me" used for caring. When compared to mock Twitter pages with high, moderate and low Klout scores at the top of Twitter profiles, profiles with a high Klout score were perceived as higher in competence and character. For the caring dimension, there is no difference in perceived credibility. The study provided a possible explanation for this, being that the idea of caring is unrelated to the Klout score. Coming to conclusions about whether someone with a high Klout score (or anybody else) has the specific user's best interests at heart may thus affect the user's attention to the message content, which was not considered in this study.

Chorley et al. (2015) ran an experiment in which participants were shown the quantitative (i.e., follower count, following count, tweets count, and number of retweets) and qualitative metadata (i.e., screen name, name, avatar, and friendship, i.e. whether the



author is followed by the authenticated user) of two tweets. Participants were asked to choose which one they would want to read. The study found that the strongest quantitative indicator of a preference for one tweet over another is the number of retweets. However, when quantitative and qualitative metadata were shown together, the most important indicators were qualitative (friendship data).

Shariff et al. (2017) also used tweet features (i.e., author, topic, auxiliary, style), but expanded to focus on how reader demographics (gender, age, education, location) and news attributes (type, trending, year) influence the reader's credibility perception. The authors found that readers' education level and geographic location correlated significantly with their credibility judgements. According to the study, even with differences in users' demographics, a tweet can be identified by users as credible based on features such as topic keyword and tweet writing style. Additionally, the study discovered that news attributes to some degree affected the reader's credibility perception. For instance, the readers perceived breaking news as very credible. The percentage of readers who depended on topic and style features for credibility judgements exceeded 26%. On the other hand, features such as the auxiliary features (additional information besides the text, such as URL links, pictures, or videos) and author features seemed to be less important to the readers. Moreover, Shariff et al. (2017) compared the reader credibility rating with TweetCred's rating. TweetCred is a public tweet credibility prediction tool based on 24 features categorised into six types: tweets meta-data, content-based features (simple lexical and other linguistic features), user-based features, external URL reputation, and network-based features (Gupta, Kumaraguru, Castillo, & Meier, 2014). Shariff et al. (2017) found that readers easily believe in a news-related tweet's credibility; this could be because they concentrated on superficial features displayed in tweet contents.

Lin et al. (2016) examined credibility perceptions of profiles that shared information regarding the increase of drug-resistant Gonorrhoea. They focused on three types of source heuristics proposed in the MAIN model (Sundar, 2008), namely authority, identity and bandwagon (as explained in Section 2.3). An experimental manipulation of six mock profiles was shown to the participants to measure the impact of the three

heuristics on three credibility constructs: competence, goodwill, and trustworthiness. The results showed all three heuristics influenced participants' perception, yet authority had the greatest influence on credibility perception (Lin & Spence, 2018).

These previous investigations examined the relationship between specific features and users' perception of the credibility of the profile or message. However, they do not consider bots. Research on the subject of interpersonal communication and computer mediated communication (CMC) has investigated how users interact with bots on Twitter. For example, scholars have examined the source credibility perceptions of Twitter Bots as part of perceptions of communication quality. An experimental approach conducted by Edwards, Edwards, Spence, and Shelton (2014) used two mock Twitter pages to represent a bot agent and a human agent which tweeted for the Centers for Disease Control and Prevention (CDC) on the subject of sexually transmitted infections. Both pages included the exact same information; the only difference being that the author of one page was described as a CDC Twitterbot and the other one as a CDC Scientist. The purpose was to examine users' perceptions about the communication quality of Twitterbots used by a credible organisation (CDC); in other words, to see if a Twitterbot was perceived as a variant of a human agent or perceived differently. Dependent variables included source credibility, task and social attraction, CMC competence and intention of interaction. The study's findings demonstrated that Twitterbots were considered credible, attractive and interactive. However, human sources were rated higher than bots in terms of social and task attraction. Yet the study focused on the CDC only as an experimental stimulus, as the CDC is an organisation well known to most people. It is possible that more general health information or different types of organisations would elicit different perceptions. Bots control not only organisations but can also control individual accounts, which can be harder to detect. Bots exhibit differences in behaviour when controlling individual accounts compared to organisation accounts, which can sometimes be negative. For this reason, another study (Spence, Edwards, Edwards, & Jin, 2019) examined the same variables as those used by Edwards et al. (2014) for weather-related information. The aim was to test variations of user perceptions of an amateur meteorologist, a professional meteorologist,

and a weather Twitterbot. Participants viewed three mock Twitter pages, all with the exact same features (e.g., posts, followers) except for the bio description and images, which were manipulated to represent each of the three types. Generally, the weather Twitterbot was perceived as credible. The results of the study showed that the bot was found to be more task oriented than the amateur meteorologist posting identical messages. This result contradicted previous findings by (Edwards et al., 2014), in the context of the CDC; the human agent was perceived to be more task attractive. This indicates that the topic is an important factor in perceptions of communication quality. In both studies, participants knew whether they were interacting with humans or bots. Yet, users may encounter bot profiles without recognising them. In addition, both bot and human profiles were manipulated to present the same features (e.g., posts, number of followers), which would not occur in realistic settings.

To sum up, many attempts have been made to analyse the perception of credibility on SM and assess whether specific features or cues affect this. Additionally, quality communication researchers attempt to understand user interaction between humans and Twitter bots, providing a glimpse into the challenging area of source credibility perceptions. However, there are some limitations in the studies discussed, as listed below:

First, most studies that examined human perceptions of Twitter information credibility employed a quantitative methodology, in which credibility was investigated as a dependent variable. The purpose of this quantitative approach is to develop or test theories (Sbaffi et al., 2017). However, since the constructs used to define credibility have no single definition and the relationship among variables is not clear, it is difficult to generate comprehensive and coherent results (Sbaffi et al., 2017). Although participants in some existing studies were confined to providing specific answers to open questions to express their opinions in the form of qualitative data, this does not include all possible real-life influences. Moreover, most experiments conducted in these studies focus on manipulating screenshots or creating mock Twitter profile pages, tweets or both, along with their features, rather than showing live feeds. These mocks were shown to participants to rate the credibility of profiles or messages to identify the in-

fluence of certain features (Shariff, 2020). However, exposure to a static view of the feed could lead to different perceptions than exposure to a live view (Edwards et al., 2014). This demonstrates the need to further examine participants' perceptions using live feeds. Therefore, qualitative studies are required to gain a better understanding of the context, procedures, and perceived credibility decisions (Sbaffi et al., 2017).

Secondly, in terms of context, most user studies in the SM credibility area have targeted topics related to social events, politics, and news. Twitter has grown in popularity as a social platform not only for events, news and political information, but also for health information. A recent review of challenges to information evaluation on SM by Keshavarz (2020) found that only a few studies have focused on people's assessments of health-related information on SM, as opposed to the many studies on users' assessment of databases and static websites, such as (Liao & Fu, 2014) and (Klawitter & Hargittai, 2018).

This emphasises the existing gap in the research on information credibility on social platforms in terms of context and methodological approaches. Few researchers have looked at the credibility assessment of information related to specific health conditions from the information consumers' perspective. This research fills that gap by adopting qualitative methods for a deeper understanding of health information judgements on SM with a fully "live" feed in a real setting.

### **3.2.2 The Machine Approach**

A wide variety of studies have adopted social graph and/or ML techniques to address the issue of credibility assessment on social platforms. This approach is often used to develop predictive models to measure one or more credibility aspects. Two common methods are reported in the literature, the propagation-based method and the classification-based method (S. Kumar & Shah, 2018; Alrubaian et al., 2018; Pasi & Viviani, 2020). These are discussed below.

### 3.2.2.1 Propagation-Based Methods

Propagation-based methods, also called graph-based methods, utilise network structure or social graph representation to assess the credibility of the source. Most research applying graph-based methods focuses on credibility aspects related to the source, because false news is three times more likely to be shared on SM than verified news (Vosoughi, Roy, & Aral, 2018). False tweets propagate much faster and wider, while verified tweets are retweeted less, have a lower overall reach, and take roughly six times longer to reach 1,500 users (Sommariva, Vamos, Mantzarlis, ào, & Martinez Tyson, 2018).

The social graph is constructed using relations between entities, such as retweets or following/follower. Nodes in the graph represent the entities (e.g., users, tweets, or topics), and the (directed or undirected) edges represent the relationships between them. Propagation-based features focus on the characteristics of the social (propagation) graphs (e.g., the depth of the retweet tree). The graph-based approach can be used to rank the reputation of information sources based on followers and shares, derived from a retweet-based network, where users are the nodes and retweets are the edges. Weights are calculated by the proportion of the number of retweets gained by a source from another user in relation to the total retweets gained by the source, with a ‘discount rate’ which indicates whether there are any relationships (e.g., a following relationship) between users (Weitzel, de Oliveira, & Quaresma, 2014).

However, a major problem with graph-based methods is that they assume the post is credible if it originated or is propagated through a highly influential or central user (Pasi & Viviani, 2020), but the quality of the post itself is ignored and, as mentioned above, false news is far more likely to be shared than verified news. Another problem is that the source credibility is determined by the user’s influence in a graph formed by links indicating relationships (e.g., follower, following, retweets), which could be manipulated by bots or malicious users.

Before moving to the next section, it is important to note that there is another use of propagation-based methods, which takes direct advantage of propagation paths for dense block detection, or detection of hidden groups with similar characteristics. How-

ever, these studies do not address credibility but rather other domains such as fraud detection or security. For example, Jiang, Cui, Beutel, Faloutsos, and Yang (2014) studied who-follows-who graphs on Weibo and found groups of followers functioning together, continuously following the same set of followees, often with no additional activity. Overbey, Ek, Pinzhoffer, and Williams (2019) investigated common enemy graphs to discover groups of accounts demonstrating shared activity, particularly those with potential coordination or automation. Edges in the enemy graph of a retweet network represent relations between accounts that have retweeted one or more posts from the same users. The authors developed edge weight variations of fuzzy competition graphs. Typically, the majority of graph-based attempts strive to find a dense block of users, information or activity in an underlying adjacency matrix occurring in brief periods of time. In this case, it is doubtful that small-scale activity will be detected, because the algorithms focus primarily on large-scale correlated activities and the densest blocks (S. Kumar & Shah, 2018). Overall, these attempts are mainly focused on the propagation of false information and malicious user behaviour. This is different from information credibility assessments, which concern more feature-based approaches (Pasi & Viviani, 2020). These are employed by this research and discussed next.

### 3.2.2.2 Classification-Based Methods

Classification-based methods, also called feature-based algorithms, are commonly used to transform observations into features (attributes related to the entities, including users, the information items that are created and shared, and the virtual relationships between entities). Credibility features are derived from one or more of the following categories (S. Kumar & Shah, 2018; Pasi & Viviani, 2020). The first category is content based: features focusing on the information of posts that users generate. This could be features extracted from the text of the post itself (e.g., semantic, lexicon, sentiment, etc.) or popularity features that express post engagement attributes (e.g., number of likes, number of retweets, replies etc). The second category is user based and focuses on the properties of the user account generating and disseminating content

(e.g., physical location, image used, registration time). The third category is temporal. It focuses on patterns such as posting time (e.g., the average amount of time between two messages). The last category focuses on domain-specific features, that is, features that are specific to the platform (e.g., hashtags on Twitter). Features can be used individually or in combination, and they can differentiate between credible and non-credible content, to be fed into ML algorithms. Usually, supervised learning algorithms (explained in Chapter 4) are utilised for classifying or predicting the credibility of an entity. It is worth mentioning that the aim of this classification or prediction is to assess the quality of the information, although the that is not always the term used to denote it. At the post level, studies typically collect posts and their labels to train a classifier based on the collected content. Other contextual information, such as user or temporal information, can be incorporated. However, there is a heavy focus on content features, fuelled by the underlying assumption that false information may be contain specific keywords and/or a combination of keywords, allowing an individual post with adequate misinformation cues to be classified (L. Wu et al., 2019).

Early attempts in assessing information credibility on social platforms adapted classification-based methods in the context of events, politics and news. For example, Gupta and Kumaraguru (2012) trained supervised ML classifiers on user and content features to predict the credibility of events related to 14 different topics (e.g., UK riots, the Libya crisis, an earthquake in Virginia). Qazvinian, Rosengren, Radev, and Mei (2011) used content-, network-, and Twitter-specific meme features (hashtags and URLs) to identify rumours in general datasets (e.g., rumours about Barack Obama being a Muslim or Sarah Palin getting divorced). Hamidian and Diab (2016) used the same datasets as (Qazvinian et al., 2011) and the same features, together with newly proposed features related to the content, popularity and Twitter specific features. J. Ma, Gao, Wei, Lu, and Wong (2015) argued that content, user and propagation features vary over time in a rumour context, therefore temporal features or the changes in these features over the rumour’s lifecycle should be studied. An example of these variations is that, in the final stages of the rumour’s diffusion, the non-rumour uses less question marks than the rumour. These differences reflect the features of

rumours and non-rumours as they spread over time. The results indicated improvement in the classification performance over methods that do not consider temporal features. Alrubaiyan, Al-qurishi, Al-rakhami, Rahman, and Alamri (2015) presented a multi-staged credibility framework for content that related to political topics (the Islamic State of Iraq and the Levant). The framework consisted of four components: feature extractions, features' relative importance, classification, and an opinion-mining component. Features used were user-based features, tweet popularity features (e.g., number of replies) and Twitter features (e.g., hashtags). Alrubaiyan et al. (2015) used a naïve Bayes classifier to classify tweets based on the result of ranked features and the opinion-mining component that analyses the sentiment of people who engage with the tweets. Similarly, to identify the credibility of Arabic news, Sabbeh and Baatwah (2018) employed a supervised learning algorithm using user-based features, tweet popularity features and Twitter features, plus an additional feature, the polarity of users' comments, derived by analysing the sentiment of user replies. In contrast to supervised algorithms, Gupta et al. (2014) proposed a semi-supervised algorithm that scores tweets in real time, based on their credibility. Crowdsourcing was used to find a ground truth for training their model. The training datasets were related to six different events (e.g., the Boston Marathon blasts in the US). Features included the tweet content, users, and information about external URLs. Different ranking schemes, such as AdaRank (J. Xu & Li, 2007) and Coordinate Ascent (Metzler & Bruce Croft, 2007) were evaluated, with similar performance.

With the advancement of neural networks, studies have harnessed deep learning methods, which could help in representation learning, in place of traditional ML algorithms. For example, Volkova, Shaffer, Jang, and Hodas (2017) trained a neural network model on content and graph features to evaluate the credibility of news and classify it into satire, hoaxes, clickbait, and propaganda. Their findings show that recurrent and convolutional neural networks are effective at differentiating news in the four categories indicated. Similarly, J. Ma, Gao, and Wong (2018) used deep learning by forming recursive neural network algorithms to represent sequential posts for rumour detection on Twitter. They utilised bottom-up and top-down designs to characterise



the propagation process, which may capture the indicative features of the propagation path efficiently. In other words, they aimed to bridge the content semantics of posts and propagation clues through a recursive feature learning process along the tree structure. However, a key drawback of deep learning models is that they require a large amount of training data and training time, as well as parameter tuning, and their performance is sometimes difficult to assess (Su, Wan, Liu, Huang, et al., 2020).

One of the major shortcomings of the classification-based methods is that most studies utilise profile features (e.g., followers,) and/or post popularity features (e.g., retweets) that are frequently manipulated by actors such as bots, resulting in a wholly misleading impression of credibility (Qureshi et al., 2021). Moreover, research using classification-based methods has largely focused on politics, news and events and concerns have been raised about the applicability of such methods to other datasets (Zannettou et al., 2019), such as health. It is thus vital to develop automatic or semi-automatic approaches that assist people in avoiding the potentially detrimental implications of what Pasi and Viviani (2020, p.7) calls ‘social word of mouth’, that is, word of mouth on SM, particularly in sensitive contexts such as healthcare, such as dementia related information. Only a few studies have proposed solutions on this, and with some limitations; these are discussed in detail in Section 3.3.

### 3.2.3 Hybrid Approach

Hybrid methods utilise the advantages of both graph-based and classification-based (feature-based) methods. Hybrid studies can take two forms, either starting from the feature-based model or starting from the graph-based model. Hybrid based studies typically begin with utilising the feature-based model to obtain seed scores for entities (e.g., tweets, users) which then become nodes in a network where links between entities are made and weights are assigned. Later, graph-based optimisation algorithms are applied for score convergence, and various credibility prediction thresholds are used (Qureshi et al., 2021). For example, Karagöz (2016) proposed a hybrid solution combining feature-based and graph-based methods for credibility analysis of Turkish news and discussion programmes on TV. A total of 22 features of tweet texts were used to

classify Turkish tweets related to television programmes as credible or not credible, by employing supervised feature-based methods. Afterwards, a graph was constructed to represent tweets and users as nodes. Users were linked by their following/follower relationship, and tweets linked by predefined cosine similarity. By using graph-based algorithms, the study determined whether credible tweets originated from a closer group in the graph and were similar in context.

The other form of hybrid methods is to obtain propagation information initially and then use it in combination with other features. In fact, the first study on credibility automation by C. Castillo, Mendoza, and Poblete (2011) used a hybrid approach. A propagation tree was built from the re-tweets graph. A combination of message-based, user-based, Twitter-specific, and propagation-based features (e.g., depth of the retweet tree) was used to predict the credibility of event-related tweets by applying supervised algorithms. Ferrara, Varol, Menczer, and Flammini (2016) generated propagation features by constructing three types of graphs, namely retweet-, mention-, and hashtag co-occurrence networks, to identify promoted information campaigns. Propagation features were then used with a group of features including user, timing, content and sentiment.

As another example, in a study to identify rumours on Sina Weibo, (K. Wu, Yang, & Zhu, 2015) represented the information thread as a tree, with each post forming the root, and the replies as children. Each post is associated with the details of the user who posted it, the timestamp, and the client information from which the post was sent (e.g., mobile). Features from post propagation trees, including temporal behaviour, sentiment of re-posts, and user details, are extracted. A feature-based method, support vector machine (SVM), is then used to classify different propagation trees. The SVM algorithm uses a hybrid novel random walk graph kernel (Gärtner, Flach, & Wrobel, 2003) and normal radial basis function (RBF) (Buhmann, 2000). The random walk kernel used to calculate similarity between propagation trees and RBF kernel calculate the high dimensional distance between two vectors of traditional and semantic features.

Despite the advantages of the hybrid approach, it is subject to the same drawbacks as the propagation-based and classification-based methods. As described in the two

previous sections, the main flaw of propagation-based systems is that they believe a post is credible if it came from or was spread by influential or central users, but the quality of the post is neglected. Moreover, the propagation features (e.g., retweets) needed to construct the graph are sometimes not available at the early stage of information circulation. Another issue is that the source's credibility is judged by the user's influence in a graph formed by links signalling relationships (e.g., follower, following, retweets), which bots or malicious users might manipulate. Likewise, the drawback of classification-based methods is that most studies employ features that are regularly influenced by actors such as bots, for example user attributes (e.g., followers) and/or post popularity features (e.g., retweets), resulting in a completely false sense of credibility (Qureshi et al., 2021). Importantly, the hybrid approach would not perform well in some scenarios, such as at the early stages of the information propagation where there are only a few posts available or a post has not yet been reposted. Consequently, creating a graph could be infeasible (Kwon, Cha, & Jung, 2017). Kwon et al. (2017) used user, linguistic, propagation, and temporal features over varying time spans. The study compared the prediction power of each feature category in distinguishing between rumour and non-rumour events using classification methods, assuming that propagation features would change during observation periods. The findings showed that the network features perform poorly during the early circulation period of rumour, and they need a longer time period to become predictive. It is also observed that the combination of user and linguistic features proved to be powerful and that it performed consistently over short and long time period windows when compared to the other predictive features of rumour (e.g., propagation features). Propagation features impact the prediction performance depending on the observation windows. For these reasons, the current research does not use propagation features, but rather rely on users' and posts' linguistic features using the classification-based approach.

### 3.3 The Health SIR Credibility Literature

The previous section presented different machine approaches applied in the literature on credibility in SIR. This section elaborates on these approaches, with specific reference to health information quality. In terms of the research context, a large number of solutions has been proposed in the literature on credibility assessment of social platform information. However, the investigation on health-related information in this context is still in its early stages (Y.-J. Li et al., 2019). Research on health information credibility on SM has been published in different disciplines, including information systems, healthcare, communication (Y.-J. Li et al., 2019), psychology, computational science, and epidemiology (Y. Wang et al., 2019) and it has potential to be a promising research field (Y.-J. Li et al., 2019).

At the topic level, prior research on health credibility has mainly focused on vaccines and infectious diseases (Y. Wang et al., 2019). Since the Zika outbreak was declared a Public Health Emergency of International Concern in 2016, research has focused mainly on the analysis of Zika misinformation on SM (Y. Zhao, Da, & Yan, 2021). Similarly, the ongoing COVID-19 pandemic necessitates continuous research on the topic.

A vast number of prior studies on health information quality on SM applied content analysis (Y.-J. Li et al., 2019; Y. Wang et al., 2019). Content analysis is ‘the intellectual process of categorising qualitative textual data into clusters of similar entities, or conceptual categories, to identify consistent patterns and relationships between variables or themes’ (Julien, 2008, p. 120). The process of content analysis includes both quantitative and qualitative techniques. More about the two methods is explained in Chapter4. For instance, quantitative content analysis was utilised in the study by Waszak, Kasprzycka-Waszak, and Kubanek (2018) to quantify the amount of health misinformation among the top links shared in Polish on BuzzSumo (a SM analytics tool). Forty percent of the most frequently shared links was classified as fake news. News about strokes and heart attacks was found to be correctly reported and of high quality; most of the inaccurate content concerned vaccines.

Sommariva et al. (2018) applied content analysis to Zika-related news stories in

order to verify their accuracy. They also analysed the volume of shares and carried out a thematic analysis of headlines. Half of the top 10 news stories about the Zika virus in 2016 were classified as rumours. The analysis also found a close correlation between the popularity of a topic, measured by the number of times a story's link was shared, and the likelihood of fake news on the topic. The most common rumours included stories with headlines about pesticides' role in the epidemic and blame about a person or organisation. Chen, Wang, Peng, et al. (2018) performed content analysis of gynaecological cancer-related tweets on Weibo (the Chinese microblog equivalent of Twitter) to distinguish between accurate and misinformation and to identify sources of misinformation. The findings revealed that more than 30% of gynaecological cancer-related tweets contained misinformation, but the majority of tweets contained medically appropriate information. Content analyses are often done by two or more coders who are able to assess the quality and classify the information as false or true based on their knowledge or according to evidence provided by official health authorities. Some studies have adopted a theoretical framework or model as basis for the content analysis. For example, Y. Li, Zhang, and Wang (2017) performed open coding using the Credibility, Accuracy, Reasonableness, Support (CARS) Checklist as a framework (Harris, 1997) to guide their code to identify low quality information on 428 posts collected from WeChat, a social media platform in China. Four main categories of features were identified, namely, lack of credibility (e.g., negative information, or business promotion), lack of accuracy (e.g., grammatical errors, typo), unreasonableness (e.g., overblown importance), and lack of support (e.g., no source). The goal of these studies was to apply traditional content analysis to identify the most relevant themes or topics of misinformation about health information and their popularity on social platforms. The most obvious drawbacks of traditional content analysis are the time and effort required to carry it out. Furthermore, traditional content analysis is usually performed on a small amount of text-based data. On the other hand, automated features-based analysis allows for the analysis of large-scale data and reduces the costs of manual annotation. Therefore, automatic assessment of the quality of online health information is necessary, particularly given the massive increase of online content (Al-jefri, Evans,

Ghezzi, & Uchyigit, 2017).

Another line of study on the credibility assessment of health information featured on SM has adopted a feature-based and hybrid approach. Articles that apply a classification (feature-based) approach to assess whether the post contains false or true information use terms like “detection”, “classification”, and “tracking” to describe the study’s aims, however, most studies targeted the content quality aspect of credibility even when this was not explicitly referred to.

At the beginning of the current research project (the end of 2017), there were only a few studies that proposed ML solutions to assess the quality of health-related information circulating on SM, despite the high level of interest in this field. A recent survey of articles reported the importance of and called for the development of automated or semi-automatic methods (Pasi & Viviani, 2020; Gupta & Katarya, 2020) to help people to avoid the potentially harmful implications of social word of mouth, particularly in such a sensitive context as health. Most of the extant research has used methods that have proven their success in detecting fake news on SM. Studies that have developed automatic solutions (ML approach) for assessing health information quality on Twitter only are discussed below, because of the differences in structure between SM platforms. Platforms such as Facebook, YouTube, and Instagram exhibit different information features than Twitter. As mentioned earlier, the term “quality” is not always used in these studies, although the aim of these automatic solutions is to assess the quality of the information. Terms such as “misinformation” and “rumours” are most likely to appear in the relevant studies.

To find the appropriate studies for review, a unified query was used in four different databases: Institute of Electrical and Electronics Engineers (IEEE), Springer, Science Direct and Association for Computing Machinery (ACM). These are the most common databases, and they cover the major articles in the areas of Computer and Information Science. The query contained keywords often used in the literature, such as “health misinformation”, “health rumours”, “information quality”, “false information”, “detection”, “machine learning”, “classifier” and “twitter”. The query was (“Health”) AND (“Misinformation” OR “rumours” OR “Information quality” OR “false information”)

Table 3.2: Relevant studies on quality of health related information on SM (2017-2020).

Database	Total	Selected	2017	2018	2019	2020
Science Direct	3	1	0	1	0	0
ACM	2	1	0	1	0	0
IEEE	6	3	1	1	1	0
Springer	124	0	0	0	0	0
<b>Total</b>	135	5	1	3	1	0

AND (“Detection” OR “Machine learning” OR “Classifier”) AND (“Twitter”). The keyword searches were restricted to abstracts only, focused on articles in the English language. Articles were collected for four years from 2017 to 2020. This timeframe was chosen because the first early study conducted was in 2017 (as per the study by Ghenai and Mejova (2017), this study ‘is a first application of the state-of-the-art SM analytic tools to the problem of health rumor tracking’), and 2020 was the time of finalising writing the literature review of the current research. A total of 142 studies were identified, of which only five are relevant (see Table 3.2).

A large number of articles were found in Springer database, because it does not allow for restricting the query to the abstract, but only provides a full text search. These results were filtered to include articles and conference papers only. Having read the abstracts of all retrieved articles, it was determined that most were irrelevant to the topic in question although they have all the search keywords used in the full text of the article.

The irrelevant and duplicated articles were disregarded. Irrelevant articles include studies where the quality of health information on Twitter is not the main focus. Two articles (Sicilia, Giudice, Pei, Pechenizkiy, & Soda, 2018; Rath, Gao, & Srivastava, 2019) appeared in two database results (Science Direct and IEEE, and ACM and IEEE, respectively) and were counted only once. After removing duplicates and irrelevant articles, there were only five articles left, which are discussed in detail below.

Research carried out by (Ghenai & Mejova, 2017) is very early and forms the foundation of research in the area of health information quality on SM using ML. The research proposed a supervised ML model to study the quality features on the Zika

virus on Twitter using different feature categories: profile features (e.g., number of followers, number of following), content features (linguistic features, sentiment features, medical features, readability), post popularity features (e.g., if it is retweeted), and Twitter features (e.g., number of hashtags). These feature categories had been utilised in prior news credibility studies, however, Ghenai and Mejova (2017) explored new content feature categories, including readability features and medical features. The readability features refer to predefined readability scores (e.g., the proportion of complex terms, the average number of syllables per word) and the number of words not in word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Word2vec is a model trained on Google news data and provided by Google. If a word is not in word2vec, this could indicate slang language. Word2vec is one of the most popular techniques to learn word embeddings, that is, vector representations of a particular word (Mikolov, Chen, Corrado, & Dean, 2013). The medical features category focuses on two types of features: tweets' medical lexicon and source reliability based on shared URLs. The tweet's medical lexicon counts the number of medical words in a tweet. For this purpose, the study created a specialised medical lexicon by downloading 113 Wikipedia articles focusing on infectious disease. This yielded 22,123 terms forming a corpus named M. Another 22,123 terms of the most frequent words on all of Wikipedia, forming a general corpus named G, were also downloaded. This was to rank the terms in corpus M based on their popularity in the corpus G. As a result, only the top 13,300 meaningful terms were kept (e.g., syphilis, bronchitis, tetanus, diarrhoea, epidemiology, treatment, life). Second, from the Wikipedia pages downloaded to create the medical lexicon, the study pulled a list of referenced URLs that refer to 441 distinct domains, based on the assumption that references found on Wikipedia are reliable sources. Then, the domains were categorised into four groups, advocacy (advocating for certain activities or policies or claiming to be the best at giving information without having any official affiliations), SM, news, and informative (medical information: medical companies, government sites, Snopes, etc.) vs. non-informative (URLs without a specific domain type). Thus, four features for each tweet represent the number of URLs belonging to one of the four predefined domain types. Three different supervised algorithms,



namely naïve Bayes (H. Zhang & Su, 2004), random forest (RF) (Breiman, 2001) and decision tree (DT) (Du & Zhan, 2002) were trained on the labelled tweets to classify them into rumour or non-rumour. The best result was achieved using the random tree classifier using only the top 10 features. These features were selected using two techniques, information gain (Cord & Cunningham, 2008) and the greedy backward elimination technique (Cord & Cunningham, 2008). The 10 most significant features were medical features related to the URL types (number of advocacy domains, number of Wikipedia domains), linguistic features (question marks/exclamation marks, verb count, adverb count), sentiment features (sentiment score, number of negative words), Twitter features (number of mentions), popularity (retweet count), and profile features (age). Notably, none of the profile features contributed to the classification, except for account age. Although the proposed system achieved a very good F1 of about 94%, the authors noted that the results may be overfitted because of the high imbalance in the number of tweets between the two classes (a ratio of 32% rumours to 68% non-rumours). In addition, this study did not consider whether the sources of the tweets in the sample were legitimate users or bots.

The same authors proposed another classification solution (Ghenai & Mejova, 2018) to identify users prone to propagate misinformation about cancer treatment. The main idea behind this research was to identify suspect accounts to enable early detection of new, potentially questionable content prior to possibly spreading across the network. The study used the same feature categories (profile features, content features (linguistic features, sentiment features, medical features, readability), post popularity features, and Twitter features) as in their previous work (Ghenai & Mejova, 2017), with the addition of a new temporal feature. The temporal feature includes the entropy of the intervals between a user's posts. The linguistic features were extended to include psycholinguistic features from Linguistic Inquiry and Word Count (LIWC; see Chapter 4 for more details about LIWC), and the medical features limited to the URL domains category. The dataset was balanced (4,212 users) for each two classes (cancer rumour users vs. control users). User accounts that were not controlled by individuals, such as organisations or bots, were eliminated. The Humanizer tool (McCorriston, Jurgens, &

Ruths, 2015) was used to identify organisations. To identify human accounts, the authors relied on either comparing usernames to dictionaries of published baby names by Social Security databases or heuristics (such as having “Mrs.” or “Mr.”) and excluded users with a high average number of tweets per day. Logistic regression (LR) with a least absolute shrinkage and selection operator (LASSO) regularisation was trained on the labelled tweets. It was found that message-based features such as readability, tentative language and avoidance of personal pronouns are important cues for the likelihood of cancer treatment misinformation. However, the research was limited by its focus on individual sources only, since misinformation can also originate from organisations or companies, for example for advertising purposes. Also, bots can mimic humans and select human names for their screen names. Improvement in the detection of bots will allow for a more accurate account selection (Ghenai & Mejova, 2018) than relying on dictionaries of names.

Sicilia, Giudice, et al. (2018) proposed a detection system using a data set containing 709 samples about the Zika virus (pulled using the #Zika hashtag), including 54% rumours, 30% non-rumours and 16% unknown. The main aim was to detect rumours related to Zika using a hybrid approach. User, post, and popularity features and propagation features of a graph formed by both retweets and replies were used to train classifiers. A set of 24 features was grouped into three categories and each category captured features from user and/or network levels. The first category, influence potential, refers to the features that have the power or ability to cause effect. Features in this category were derived from user and network levels. For example, influence potential features on the user level included the number of followers and followings, whether they were followers of another user involved in a conversation and the age of the account. The network level features include the average number of followers. The second category, propagation (network characteristics), refers to the properties of propagation graphs created via retweets and replies. Features in this category were derived from the network level (e.g., page rank, closeness, betweenness, centrality and conversation size, as well as scores assigned to all tweets belonging to each conversation). The third category, personal interest, refers to features that express people’s reactions to specific news

in terms of opinions and sentiment. Features in this category were derived from the user and network levels. All features were evaluated using the wrapper feature selection method (Huang, 2015), specifically RF, resulting in 20 top features and discarding four features (number of followers, number of statuses, whether the user was a follower of another user involved in a conversation (“if follower”), and the presence of a question mark in a specific tweet). Notably, the most informative features belong to the network level, whereas only a few features belong to the user level: two content related features (the presence of URLs in a tweet and sentiment scores), and one popularity feature (the probability that a tweet is retweeted). The study noted that profile features were eliminated (e.g., number of followers and status) because they provided little information. Different ML algorithms, including multi-layer perceptron, a nearest neighbour, SVM, DT, multiclass Adaboost, and RF, were trained on labelled tweets to classify tweet posts into rumour, non-rumour and unknown. The RF algorithm achieved the best result with an overall accuracy of 73.6% using the best 20 features.

Sicilia, Merone, et al. (2018) examined the same feature categories on the same dataset as (Sicilia, Giudice, et al., 2018) and on a larger dataset on vaccines (pulled using the #Vaccine hashtag), containing 1,409 tweets, with 28% rumours, 30% unknowns and 42% non-rumours. In the previous work (Sicilia, Giudice, et al., 2018), the wrapper method (RF) was used for feature selection, but in this work (Sicilia, Merone, et al., 2018), ranking feature selection (Relieff) (Huang, 2015) was performed. Ranking feature is independent from classification algorithms, whereas the wrapper compares the performances of classification algorithms to various candidate sets of variables. This selection was justified by the study’s focus on the representative power of features among two different datasets. The RF algorithm was trained on both datasets using the best features identified. It achieved a higher overall accuracy of 96% on the #Vaccine dataset compared to the 82.3% on the #Zika dataset (Sicilia, Merone, et al., 2018). The recall values indicated the percentage of correctly classified rumours was 95.2% in the vaccine datasets and 88.4% for Zikavirus dataset. Notably, the accuracy of 82.3% in the Zika dataset is higher than the 73.6% accuracy in (Sicilia, Giudice, et al., 2018). This increase could be caused by the application of different feature selection methods;

ranking features rather than wrapped features for feature evaluation. The study reports the recall for the rumour class and the overall precision and recall, but it does not provide the recall for non-rumour and unknown. Since the data is unbalanced, it would have been better if the author had provided recall for both classes and/or other better metrics, such as Matthews' correlation coefficient (MCC) (defined in Section 4.2.1.3). Although accuracy and the F1 score derived using confusion matrices have been (and continue to be) among the most frequently used measures in binary classification tasks, these statistical techniques might produce potentially overoptimistic inflated outcomes, particularly for unbalanced datasets (Chicco & Jurman, 2020).

The prorogation feature is frequently manipulated by bad actors such as bots, resulting in a wholly misleading impression of credibility (Qureshi et al., 2021). For example, user influence in a social graph could be infected with bots who have fake followers. Therefore, a potential limitation of both (Sicilia, Giudice, et al., 2018) and (Sicilia, Merone, et al., 2018) is that these studies only took into account tweets that received a retweet or a reply and discarded tweets whose propagation graphs could not be built. However, most SM users interact with content in a passive manner, without replying to or liking it, so this might lead to a high number of rumour tweets that were not included in the detection. Also, in both (Sicilia, Giudice, et al., 2018) and (Sicilia, Merone, et al., 2018), the datasets contained rumours on health-related news only. News without reference is considered a rumour (Sicilia, Giudice, et al., 2018). According to Sicilia, Giudice, et al.'s (2018) definition, rumours are news in circulation without a reference, which renders it unverifiable. Non-rumour is news containing at least one reference to a verified and official link to for example hospitals or universities. "Unknown" refers to indeterminable news; news with the potential to be true but lacking a reference or containing a link to an empty page or a page not connected to the main topic.

Bhattacharjee, Srijith, and Desarkar (2019) classified false information regarding anti-vaccination tweets on a small dataset consisting of 895 tweets related to vaccination and 78 to anti-vaccination. The authors rely only on term frequency inverse document frequency (TF-IDF) (Soucy & Mineau, 2005) to create feature vectors; representing

Table 3.3: Summary of health-related studies in the SIR credibility literature (machine-based).

Study	Level	Disease/ Outbreak	Dataset Size	Features					Performance Measure/Classifier
				C	U	T	P	Prop	
(Ghenai & Mejova, 2017)	Post	Zika outbreak	N = 26,728 tweets (32% rumour 68% non-rumour)	✓	✓		✓		F1 (94%) DT
(Ghenai & Mejova, 2018)	User	Cancer	N = 4,212 rumour users, 4,212 non-rumour users	✓	✓	✓	✓		R2 McFadden (0.906) LR
(Sicilia, Giudice, et al., 2018)	Post	Zika outbreak	N = 709 (54% rumours, 30% non-rumour, 6% unknown)	✓	✓	✓	✓	✓	Accuracy (73.6%) RF
(Sicilia, Merone, et al., 2018)	Post	Zika Vaccine	N = 1409 (28% rumours, 42% non-rumours, 30% unknown)	✓	✓	✓	✓	✓	Accuracy (82.3%) #Zika (96%) #Vaccine RF
(Bhattacharjee et al., 2019)	Post	Anti-vaccine	N = 895 (78 anti-vaccination 817 pro-vaccinatio)	✓					FP (20.6%) SVM

\* C = content, U = user, T = temporal, Pop = popularity, Prop = propagation

each document or short text as a vector reflecting the importance of a word to the corpus. Three algorithms, namely LR, SVM (using linear kernel and RBV kernel), and gradient boosting (Friedman, 2002) were tested to classify the tweets. The results showed very high false positive (FP) rates (see FP definition in Section 4.2.1.3) of 93% for LR, 86% for SVM (linear kernel), 100%, for SVM (RBV kernel), and 86% for gradient boosting. This indicates that all the algorithms perform poorly in terms of classifying anti-vaccination tweets. The authors claim that the reason for the high FP is that the data size is very small and involves class imbalance. Therefore, the study employed under sampling technique (Zheng, Cai, & Li, 2015), a technique to reduce the number of observations (tweets) of the majority class (positive class) in order to balance the number of observations in the minority class (negative class). This was done to examine if it the FP rate using the same three algorithms (LR, SVM, and gradient boosting) would be reduced. The results showed that SVM (using RBF) improved significantly with the FP rate decreased to 20%. The limitation of this study is that using under-sampling techniques could cause loss on some amount of information about

the majority class. Another issue is that the study did not reveal the most distinctive features of both classes.

As a result of the COVID-19 pandemic, a great number of studies emerged related to identifying COVID-19 misinformation on Twitter in particular, due to the platform's popularity and the large amount of content in many languages. In addition, false posts related to COVID-19 on Twitter was found to be more prevalent than on other platforms: 59% of all tweets compared to 27% on YouTube and 24% on Facebook (Brennen, Simon, Howard, & Nielsen, 2020). Most studies relied on fact-checking websites as ground truth data. For example, Al-rakhami and Al-amri (2020) used ensemble techniques based on user, tweet, and Twitter features to identify misinformation related to COVID-19 and found SVM and RF to be the best models. Ng and Carley (2021) proposed a multi-class classification of coronavirus-related stories from three fact-checking websites. They classified these stories into six different classes using the BERT embeddings algorithm. This classification system was expanded to classify tweets with misinformation. An average of 59% of stories and 43% of tweets were correctly identified. Although most COVID-19 studies focused on English datasets, research was conducted using datasets in other languages such as Arabic (Alsudias & Rayson, 2020; Alqurashi, Hamoui, Alashaikh, Alhindi, & Alanazi, 2021; Haouari, Hasanain, Suwaileh, & Elsayed, 2020) and even multilingual datasets (Qazi, Imran, & Ofi, 2020; Elhadad, Li, & Gebali, 2020).

In conclusion, this section has covered the limited number of studies on the assessment of health information quality on SM using the machine learning approach, as illustrated in Table 3.3. Health topics investigated are generally related to outbreaks (e.g., Zika virus). The definitions and terms of low-quality information vary between studies; "rumour" and "misinformation" are the most common terms used. The reviewed studies showed that most of the existing information quality assessments of health information on SM employed a supervised classification-based approach or hybrid approach (graph-based and classification-based) with some limitations. An essential aspect which is largely ignored in credibility assessment employing the graph approach is the assessment of bot profiles (Qureshi et al., 2021). Bot features identifi-

### Chapter 3. Social Information Retrieval (SIR) Credibility

cation and/or removal should be undertaken in order to initiate credibility assessment (Qureshi et al., 2021). For example, to compute true user influence or expertise score without bot manipulations, bot profiles present in a friend network must be eliminated before examining the user's rank/influence. On the other hand, studies employing the classification-based approach to assess health information quality, usually use hand-crafted features that were successful in past works (e.g., credibility of news- or politics related information) and/or are incorporated with newly designed features. Generally, these features are usually categorised into post, user, popularity, and propagation. However, some features, like profile features, are frequently manipulated by bad actors such as bots, which could result in a misleading credibility assessment (Qureshi et al., 2021). Yet, there is a lack of understanding whether bot features can be used as indicators for health content quality on SM, even though there is a large amount of research focused on detecting bots on SM in separate studies. Therefore, this research examines bot features and see how they could inform automatic assessment of health related information quality on SM. To achieve this, the current research began by looking into bot features in certain health contexts (dementia) and then used these features together with features defined in the previous work to assess dementia information quality. More details about the methods used for the current research are described in the next Chapters (Chapter 5 and 6).

#### **Chapter Summary**

The review of the existing research in this chapter shows different approaches and methods used for SIR credibility assessment. Credibility studies focused on only one, or at most two, aspects of credibility (e.g., trustworthiness, quality) as the main subject of research. Different credibility aspects have been investigated in different contexts. However, the number of studies addressing the aspects of information quality and consumer perceptions of health information on social platforms is very small. The perceived credibility of health-related information obtained on social platforms and how this information is used by patients, their caregivers, and other lay health consumers has been

raised as a concern among healthcare professionals and policy makers, especially given that false, ambiguous, or too technical information can have health-related consequences for many types of users (Dalmer, 2017). In perceived credibility assessment research employing a human-based approach, different theoretical and experimental studies conceptualise credibility or assume a number of credibility cues, and then operationalise them. Operationalisation has mostly been conducted by examining the quantitative relationship between credibility cues and people's perceived credibility assessments in different contexts. However, there is a lack of studies focusing on how consumers assess the reliability of health information on SM, which can be addressed by employing the qualitative approaches. There are several studies investigating the quality of information aspect using the machine-based approach, and this is a prominent aspect of credibility of SIR, however, the quality of health information in particular has received less attention (Pasi & Viviani, 2020) and had some limitations. In conclusion, there is a lack of comprehensive investigation into credibility assessments for health information on SM. The current research fills this gap by investigating the information quality aspect by using the ML approach and perceived credibility aspect by using a human approach, as well as the link between both, in the health context. The research also proposes a framework of the two credibility aspects that can complement each other. .



## Chapter 4

# Methodology

This chapter introduces the methodology and tools that were used to address the objectives of this research. It starts with describing the research design, with a brief overview of mixed methods and the rationale for using it. Then, explanatory sequential mixed methods are discussed in detail. The rest of the chapter describes the main methods and tools employed for data collection and feature extraction. A research ethics statement is also presented.

### 4.1 Research Design

The current research provides a comprehensive understanding of information credibility related to dementia on Twitter by investigating two credibility aspects, namely information quality (using the machine approach) and perceived credibility (using the human approach). The research proposes a credibility assessment framework of dementia information that combines both aspects.

A sequential explanatory mixed methods design was employed to answer the research questions. This methodological design encompasses quantitative methods in two parts (Phase1a and Phase1b), followed by a qualitative approach (Phase 2). Three studies form the basis of this research: the first two using quantitative study (statistical analysis and ML experiments performed on SM data), and the third is a follow-up qualitative study (a think-aloud interviews). The results from each phase affected the design

of the next, resulting in a better understanding of the research problem. Each of the quantitative and qualitative phases answered different aspects of the overall research question, through which a more comprehensive picture was obtained. The quantitative and qualitative data and analysis were reported separately in each phase. The last stage of the research connects the outcomes of these two phases.

In Phase 1a (Study 1), a quantitative approach was used to establish a general understanding of the research problem (Chapter 5). Data was collected from Twitter from 8,400 users and 16,691 tweets. The study applied inductive coding to users' profile descriptions to categorise them into groups. It also evaluated the bot presence in identified groups using descriptive statistical analysis, since bots are an indicating element of information quality. The study analysed the principal features indicating bot-like behaviour using different descriptive and inferential analyses. From this first phase, a number of features were selected to be tested in the second ML phase of the study.

The results from Phase 1a influenced the design of Phase 1b (Study 2) (Chapter 6). This second study evaluated how much bots contribute to the dissemination of low-quality dementia information through a variety of statistical tests. Combinations of features developed in the first study and features discussed in the literature were used to train and test independent supervised ML algorithms to assess the quality of dementia related tweets.

In Phase 2, the qualitative study (Study 3), which was informed and guided by the findings of the two quantitative phases, think-aloud sessions and semi-structured interviews were conducted with thirteen dementia caregivers (Chapter 7). The data collected in this study was used to complement the quantitative results and provide a detailed understanding of users' assessment of credibility of information on SM. The findings of the quantitative analysis informed the criterion-based procedures (e.g., profiles shown to the users during the think-aloud session) and criterion-based participant sampling. The results of the first phase showed that the individual category represents the largest number of users (3,899 out of 8,400 users). The individual category includes sub-categories like dementia caregiver, health activist, artist, marketer, author, and

others. A description of all categories is provided in Appendix K. However, dementia caregivers are the main information consumers; Twitter is used by them in particular (Danilovich et al., 2018; Al-bahrani et al., 2017). Dementia caregivers displayed an increased need for information about the disease (Martínez-Pérez, de la Torre-Díez, Bargiela-Flórez, López-Coronado, & Rodrigues, 2015), a desire to share the caregiving experience (Al-bahrani et al., 2017; Danilovich et al., 2018), a need for support (Martínez-Pérez et al., 2015), and a desire to learn about support services (Danilovich et al., 2018). Based on these facts, the decision was made to use a sample of caregivers. The flow of the research is summarised in Figure 4.1. The stages of data collection, pre-processing, and analysis of each phase are described in detail in the chapters dealing with the individual studies.

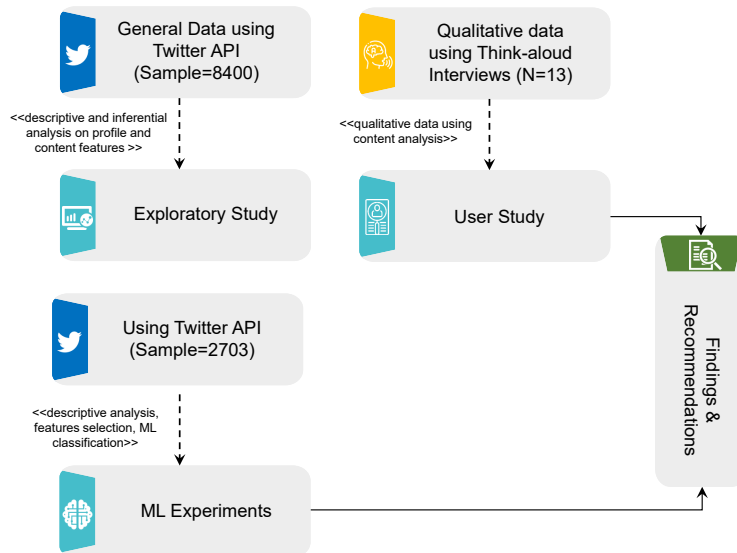


Figure 4.1: Research phases.

#### 4.1.1 Mixed Methods Approach Overview

Two general methodological approaches are followed in the research literature: quantitative and qualitative. According to Vanderstoep and Johnson (2008) quantitative research assigns numerical values to the phenomena under study, whereas qualitative

research ‘produces narrative or textual descriptions of the phenomena under study’ (Vanderstoep & Johnson, 2008, p.7). Both approaches have different characteristics, justifications for use, as well as benefits and drawbacks. A comparison between these two approaches is described by (Vanderstoep & Johnson, 2008) and shown in Table 4.1.

Table 4.1: Quantitative and qualitative research design characteristics.

<b>Characteristic</b>	<b>Quantitative research</b>	<b>Qualitative Research</b>
Type of data	Phenomena are described numerically	Phenomena are described in a narrative fashion
Analysis	Descriptive and inferential statistics	Identification of major themes
Scope of inquiry	Specific questions or hypotheses	Broad, thematic concerns
Primary advantage	Large sample, statistical validity, accurately reflects the population	Rich, in-depth, narrative description of sample
Primary disadvantage	Superficial understanding of participants’ thoughts and feelings	Small sample, not generalisable to the population at large

Source: (Vanderstoep & Johnson, 2008, p.7)

An alternative approach that embraces the traditional qualitative and quantitative approaches is the mixed methods approach. The mixed methods approach is defined as ‘the type of research in which a researcher or a research team combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the purposes of depth and breadth of understanding and corroboration’ (Onwuegbuzie, Johnson, & Turner, 2007, p.132). Creswell (2003, p.21) describes the mixed method approach as ‘employing strategies of inquiry that involve collecting data either simultaneously or sequentially to best understand research problems. The data collection involves gathering both numeric information (e.g., by questionnaire) as well as text information (e.g., by interviews) so that the final database represents both quantitative and qualitative information’.

### 4.1.2 Rationale for Mixed Methods Design

In general, the rationale for employing mixed methods research is when an integrated understanding of a phenomenon is not reached via either quantitative or qualitative methods. Mixed methods research allows researchers to broaden the scope of their investigation and strengthen their analytical capability in order to address issues that cannot be answered using only one method (Sandelowski, 2000). For example, when results often need to be explained, multiple cases have to be contrasted, participants must be included in the study, or a fundamental experimental design must be expanded (Creswell & Clark, 2011).

The justification for opting for the mixed methods approach in a single research study is determined by the general purposes of that research. According to Greene, Caracelli, and Graham (1989), five general purposes for using this approach are seeking convergence (triangulation), complementarity (examining different facets of a phenomenon to elaborate understanding of the phenomenon), development (using the first method to develop the second one), initiation (discovering and learning new perspectives), and expansion (adding breadth to the scope of the study). Researchers of mixed methods identify one or more reasons for using mixed methods designs, and more may emerge as the research progresses (Bryman, 2006).

Although the mixed methods approach is beneficial to gain a full comprehension of the research problem, it does present challenges in terms of skills, time, resources, and justifications for employing them (Creswell & Clark, 2011). Researchers have to be skilled in data collecting and analysis procedures for the two approaches required, which involves a significant amount of time, resources, and effort on their part.

The general purpose of adopting mixed methodology in the current research was based on the notions of complementarity as a motivation for combining qualitative and quantitative research methods to produce results that examine both facets of the problem (machine and human). This design permitted the data from the different research phases to be combined and to ensure a level of methodological complementarity and a comprehensive view of the research problem.

### 4.1.3 Mixed Methods Design Types

There are several major types of mixed methods research. Four factors are attributed to each type and are considered when selecting a mixed technique design type (Creswell & Clark, 2011). These factors include the level of interaction between quantitative and qualitative approaches, the relative priority, the timing, and the points of interface for mixing both approaches. The degree to which the quantitative and qualitative strands are kept separate or interact with one another is referred to as the level of interaction (Creswell & Clark, 2011). A strand is a part of a study that includes the basics of quantitative or qualitative research, such as formulating a question, collecting data, evaluating data, and interpreting results. Priority relates to the relative importance of the quantitative and qualitative strands inside the design: equal priority for both, or priority of one over the other (Creswell & Clark, 2011). Timing relates to the temporal aspects of the two strands in terms of data gathering and usage of the results. Timing might involve quantitative and qualitative approaches simultaneously, or sequentially (in different phases), as collaboration timing (Creswell & Clark, 2011). Points of interface refers to when the mixing process takes place in the research process (i.e., interpretation, data analysis, data collection). Mixing strategy forms can be either integrating two data sets or linking the analysis from the first dataset to the next dataset, immersing one or both data sets within the overall design, and employing a framework to tie the data sets together (Creswell & Clark, 2011).

The most common classification of mixed methods is into three basic designs and includes a convergent parallel mixed methods design, a sequential explanatory mixed methods design, and an exploratory sequential mixed methods design (Creswell, 2014). In a convergent parallel design, a researcher gathers both quantitative and qualitative data during the same phase, analyses the results of each study separately, and then combines the results for comprehensive interpretation. The same priority is given to each data set, because the researcher uses both types of methods equally and collects and analyses the data at the same time. When the outcomes of two data sets are integrated for interpretation, the point of interface occurs: the researcher may immediately compare, contrast, integrate, or change the individual findings for further analysis

(Creswell & Clark, 2011).

The exploratory sequential design approach is conducted in two different phases. It entails determining qualitative results based on a small number of people in the first phase, then designing an instrument and testing it on a bigger sample in the second phase. The researcher focuses on qualitative data exploration before collecting quantitative data to test or generalise the initial qualitative results. If a researcher wants to build an instrument but does not know what measurements or variables to employ, this design is appropriate (Creswell & Clark, 2011).

In an explanatory sequential design, quantitative data is collected first, followed by qualitative data to explain or expound on the quantitative results. The justification for this type is that quantitative data and findings give a broad picture of the study topic, while qualitative data analyses provide further clarification, addition, descriptions, or elaboration (Creswell & Clark, 2011). This is the approach used in the current research and it is therefore discussed in more detail in the next section.

#### **4.1.4 Explanatory Sequential Mixed Methods Design**

An investigation into trends in mixed methods designs for SM research by Snelson (2016) revealed that SM studies following an explanatory sequential mixed methods design primarily aim to acquire data from people through surveys and follow-up interviews or focus groups. Yet, some research entailed the integration of data from SM content in an explanatory sequential design as well (Snelson, 2016). For instance, a study on user-created videos about Islam on YouTube, conducted by Mosemghvdlishvili and Jansz (2013), began by analysing 120 videos through content analysis, involving coding of both quantitative and qualitative variables (e.g., video characteristics, video creator demographics, and valence framing (i.e., positive or negative expression)). The content analysis was followed by interviews with 15 users who created the videos in order to obtain a better understanding of the various aspects influencing their production and sharing videos on YouTube (Mosemghvdlishvili & Jansz, 2013).

A study by J. S. Lee (2016) utilised an explanatory sequential design to investigate Twitter use and political information behaviour by residents in Korea at the time

of Seoul’s mayoral election 2014. The explanatory sequential design permitted the researcher to first gather quantitative information to develop an overall picture of the research issue by conducting social network analysis and tweet content analysis at the time of the election to find out what kind of information was shared, as well as the collaborative information sharing behaviours of users. The qualitative data was then collected by conducting semi-structured interviews with 13 opinion leaders in the second stage to add greater exploration, augmentation, and clarification of the overall picture (J. S. Lee, 2016). The main reason for using mixed methods in J. S. Lee ’s (2016) study is the complexity of the massive amount of data available on SM sites like Twitter. Furthermore, while social network analysis is valuable for determining the position and relationships of Twitter users in terms of information behaviour, social network analysis could not uncover, in more depth, users’ goals, perceptions, and judgements of their political information behaviour in their interactions with other people, which is why utilising social network analysis only to analyse political information behaviour was insufficient (J. S. Lee, 2016).

Overall, the initial step of quantitative data gathering and analysis is given top priority by the researcher using explanatory sequential design. The researcher selects specific quantitative outcomes that require more explanation, and then creates a second phase follow-up procedure of qualitative data collection and analysis of the first quantitative phase. The quantitative outcomes for extra clarification incorporate a couple of regular or outlier cases, for example, patterns or exceptions (Caracelli & Greene, 1993). Figure 4.2 shows the typical flow for the sequential explanatory design process.



Figure 4.2: The typical flow of the sequential explanatory design process.



A mixed methods design is an effective approach for achieving the aim of the current research, which is to develop an understanding of dementia information credibility on Twitter. The primary rationale for employing the mixed methods sequential explanatory design, in particular, is that obtaining qualitative data in a second phase is important to complement the quantitative results.

The quantitative analysis established a general understanding of the quality of information related to dementia on Twitter, addressing the features that predict the information quality of posts in a large dataset using a machine approach. Yet, as explained in Section 1.1, the single approach (machine approach) used in previous studies does not provide an understanding of human perception. It is important to understand the factors that influence users' assessment of digital health information. These features are usually used to build ML algorithms to predict what a user will perceive as credible or not (Ginsca et al., 2015). Data gathering using the machine method is frequently reliant on human annotation. Since people do not consider all elements as having the same weight when judging the credibility of information, criteria for the differences in the way people assess credibility are not adequate, which has been less studied or ignored in many computational research studies (Jo, Kim, & Han, 2019). There is a gap between the computational and human-centred methods and these two approaches need to be connected (Jo et al., 2019). Moreover, as shown in the literature review Section 3.2.1, there are limited qualitative studies regarding factors affecting human credibility perceptions in the health domain particularly.

Therefore, in the current study, after quantitative data had been collected, qualitative data was gathered to complement the outcomes of the quantitative phase. This way, the quantitative data supplied understanding of the research problem from the machine perspective, while participants' views were explored in more depth via qualitative data analysis, which added insight. The quantitative analysis also contributed to the design of the procedures in the second phase.

Another advantage of employing mixed methodology is that there is a lack of mixed methods research on SM. Sayed, Dafoulas, and Saleeb (2018) conducted a descriptive analysis of the methodological approaches used in literature from 2006 to 2016 ad-

designing social network sites or SM. The results revealed that the most common design methodology is quantitative methods, while only 9% out of a total of 112 studies used mixed methods.

## 4.2 Methods

### 4.2.1 Machine Learning Experiment (Phase 1b)

This section introduces the experimental setting that is used in Phase 1b (Study 2). It depicts a short description of the ML algorithms, feature selection techniques, and evaluation measures that are referred to in the remaining chapters.

#### 4.2.1.1 Machine Learning Algorithms

The study of ML considers the theories, algorithms, and applications of systems that learn like humans (Sugiyama, 2015). There are different ML types (e.g., supervised and unsupervised). The most fundamental type of ML is supervised ML. To explain this type, Sugiyama (2015) gives the example of a student learning from a supervisor by questioning and replying. In ML, the student is the computer and the supervisor is the computer user. The computer learns by mapping a question to its response by comparing samples of questions and replies. The purpose of supervised learning is to develop generalisation ability, which refers to the ability to correctly predict responses for problems that have not been taught. As a consequence, the user is not required to train the computer on every single thing, rather the computer has to deal with unknown scenarios on its own by learning only a portion of knowledge in advance (Sugiyama, 2015). A large number of real-world classification research problems, such as spam filtering and IR, have been addressed successfully via supervised learning.

In this thesis, supervised ML classification algorithms are mainly used in the second study to classify whether a given tweet is true or false based on its quality. Apart from assigning the tweet to a certain class, the main purpose of ML is to test different subsets of features with different algorithms which could indicate the quality.

Five common supervised classification algorithms with grid search using 10-fold

cross validation (see Section 4.2.1.3) are utilised in the experimental work in Phase 1b (Study 2) to identify the influence of different combinations of features on the classification task. Grid search is used for tuning ML algorithm hyperparameters. This process involves tuning the algorithm to perform at an optimal level (Idris, 2016). For instance, the RF algorithm's hyperparameter is the number of trees and the k-nearest neighbour (KNN) algorithm's hyperparameter is K. Grid search generates a grid of all the possible parameters and every grid combination model is built. Grid search may take a long time and use more processing resources, but it analyses all parameter options.

The ML classification algorithms used in this thesis and their advantages and drawbacks are described below:

1. Decision tree is a mathematical model for making decisions. In this case, the input is an object described by a set of features, and the output is a decision for the matching input.

The decision is made by recursively selecting features and splitting the dataset on those features. A sequential decision stream-based model, based on the dataset's actual values of features, is created by this model. The choices are organised in a tree-like layout. A decision is made at each node of this tree, unless a predicate is produced for a specific input data item or if it reaches the maximum depth. The advantage of this algorithm is that it is a straightforward strategy that is simple to comprehend and visualise, as well as being quick and requiring minimal data pre-processing (Sen, Hajra, & Ghosh, 2020). However, this approach can sometimes result in a complex tree structure that is not sufficiently generalised, as well as a model that can be unstable (Sen et al., 2020), since even minor differences in the data might result in an entirely new tree being created.

2. Random forest is an ensemble learning algorithm for classification, regression or search tasks. It operates by building a forest consisting of a number of DTs during the training phase, which are then used for class prediction. The algorithm selects the best prediction through voting. In other words, selection is based on

the classes determined by all the individual DTs; the class that is chosen most often is considered to be the output (Breiman, 2001). Random forest is termed random, because it repeatedly selects a random sample of the training data, and forest, because it uses various DTs. The primary benefit of RF is that it does not require tree pruning throughout the creation process and is resistant to overfitting: when a classifier model learns the data very well and is unable to generalise. The greatest downside of this strategy is that it uses a lot of memory, determined by the size of the dataset.

3. Support vector machine is an algorithm widely used for classification tasks. It converts the original data training set to a higher dimension using nonlinear mapping (Han, Pei, & Kamber, 2011). The goal of the SVM algorithm is to find a linear optimal hyperplane in an N-dimensional space (N represents the number of features) that distinguishes between data points and separates the dataset into two classes. If the hyperplane is well built, SVM performs well, and this is also memory efficient because of the use of subset training points in the decision function. One problem is that the training period is rather long in comparison to other methods, therefore if the dataset is very large, the prediction task would be considerably slower. When target classes overlap, the dataset's performance suffers as a result of the increased noise.
4. The KNN algorithm stores all available training instances and predicts the class of new instances based on the likelihood of similarity measurements (the majority class) with the closest k neighbours (Han et al., 2011). K-nearest neighbour is effective with noisy training data, works well with very large training data, and is simple to deploy. The disadvantage of this algorithm is that in order to predict every new instance, the distance between each of the k neighbours must be calculated repeatedly, resulting in a significant increase in computational time. The k value needs to be determined efficiently (Sen et al., 2020).
5. Logistic regression is a statistical method that uses a sigmoid function to model the data. It measures statistical significance by measuring the relationship among

the categorical dependent variable and each independent variable with respect to probability (Han et al., 2011). Logistic regression utilises an s-shape “sigmoid capacity” rather than fitting a line utilising the samples. The s-shape bend goes from zero to one, which means that LR determines the likelihood of the instance as 1 (true) or 0 (false). Logistic regression has many advantages, including ease of implementation, computing and training-based efficacy, and regularisation ease. The main limitation of using LR is that the dependent variable or target is a dichotomous variable (in which there are only two possible outcomes). Yet, predictors do not have to be normally distributed or have equal variance in each group. The capacity to solve a nonlinear issue is prone to overfitting.

All empirical experiments executed on the dataset have been conducted on Python version 3.8 using Scikit-learn library.

#### 4.2.1.2 Feature Selection Techniques

The purpose of feature selection techniques is to choose a subgroup of the original features while keeping helpful and required information in separate classes. Three common feature selection categories used in Phase 1b (Study 2) are described below, with their advantages and drawbacks.

1. **Filter-based feature selection methods** use statistical techniques to evaluate the correlation or relationship between each feature and the target variable to find the most relevant features. The filter method is also called a ranking method, because it ranks all features in the input feature set. ANOVA (analysis of variance), chi square, and Pearson’s correlation are examples of these techniques. The advantage of filter-based feature selection is that it does not rely on learning algorithms as other feature selections aim to tune features to fit for/by a given learning algorithm. Rather, a filter-based selection method gives a generic list of variables. Filter based is fast in terms of computing and avoids overfitting (Guyon & Elisseeff, 2003; Chandrashekar & Sahin, 2014). The disadvantage of this method is that it ignores the features that are less individually informative

but are informative once coupled with other features. Also, because the underlying learning method is neglected, finding an appropriate learning algorithm can be difficult (Guyon & Elisseeff, 2003; Chandrashekar & Sahin, 2014).

2. **Wrapper-based features** create different models by using a different subset of features to find the best subset according to the model's performance metrics. In other words, the feature subset is evaluated using the predictor as a black box and the predictor performance as an objective function (Chandrashekar & Sahin, 2014). Several search algorithms may be used to discover a subset of variables that optimises the objective function, which is the classification performance (Guyon & Elisseeff, 2003; Chandrashekar & Sahin, 2014). Wrapper methods can either start with an empty set or a full set and add or remove features prior to the greatest objective function being found, or the best objective function can be found by assessing alternative subsets. Recursive feature elimination (REF) and backward elimination are examples of wrapper methods. The disadvantage of this method is the high computational level needed to get the feature subset. The predictor builds a new model for each subset evaluation, meaning that the predictor is trained for each subset and evaluated to acquire the classifier accuracy. As a result, the majority of the algorithm execution time is spent training the predictor, especially if there is a considerable number of samples. Also, when the performance of the classifier is used as the goal function, the classifiers are in the habit of overfitting. As a result, the classifier is more likely to be biased and expand errors in classification. A second holdout test set can be utilised to influence the prediction accuracy to avoid this.
3. **Embedded-based feature methods** use ML algorithms with built-in feature selection methods as part of the learning process, selecting features automatically. The advantage of the embedded model is that it is constructed once to determine the feature scores; thus it has lower computational costs and it takes less time to reclassify different subsets than wrapper methods, which require the construction of multiple models due to their iterative process. Examples of this category are

DT, RF, and LASSO (Rani, Gill, & Gulia, 2021).

Overall, each category has advantages and disadvantages. For example, the filter-based category, which examines each element separately, is simple to comprehend. However, it has the drawback of being unable to reduce duplicate features in highly correlated feature subsets, and it disregards the modelling's importance. The wrapper-based and embedded-based categories utilise ML models for feature ranking. For example, RF is a popular ensemble model; its key benefit is that there is essentially little need to manually alter the features, and it is difficult to overfit. Non-linear, collinearity, and interaction data can all be appropriately analysed with RF (Breiman, 2001). The feature selection process in embedded methods is an integral part of the classification, whereas the wrapper and filter methods are separate processes, in terms of choosing or discarding features.

While the majority of extant research has compared the outcomes of different types of classification algorithms using features generated from different feature selection methods, Salih and Abdulrazaq (2019) suggest combining feature selection methods (correlation feature selection, Information Gain, Gain Ratio) then combining the best features from each type of selection method, and then testing them on different classification algorithms in order to achieve a high accuracy performance in the classification. Therefore, to determine the best subset of features and reduce the classification error of the text quality analysis in the current research, one of each of the three feature selection categories, namely filter, wrapper and embedding methods were used. ANOVA tests were applied as filter-based, RFE as wrapper-based, and RF as embedded methods. This was accomplished by using the Scikit-learn Python package (Kramer, 2016), which provides an implementation of all these selection methods.

- SelectKBest in Scikit-learn selects the best features ranked using ANOVA. ANOVA is a statistical univariate method which analyses the differences among group means in a sample and selects the k best features (Fisher, 1992).
- The feature\_selection function gradually eliminates features that have been assigned a low weight by a classifier. This procedure is repeated until the desired

result is obtained.

- The RF algorithm has an embedded feature selection method called `features_importance` that can provide important metrics for each input feature while analysing data.

The top 30 informative features for each selection method are ranked according to their score or rank, which is considered imperative in classifying the two classes of tweets (true/myths) when they are listed in the output of two or three of the feature selection methods.

#### **4.2.1.3 Model Evaluations**

This section presents the validation used to build the ML models used in Phase 1b (Study 2) and the performance measures used to evaluate these models.

##### **1- Cross-Validation**

One of the most important rules to follow when developing ML models is to avoid testing against the datasets that were used to train them (James, Witten, Hastie, & Tibshirani, 2013) to avoid overfitting. A validation set (test data), which has never been seen by the model, is used to offer an unbiased evaluation of the final model. Cross-validation can be accomplished by dividing the training set into  $k$  equal folds (groups), which is known as  $k$ -fold cross-validation. The model is trained on the other  $k-1$  folds combined and then tested on that fold. This procedure is repeated  $k$  times to test different folds, after which the results of the iterations are averaged. The main advantage of  $k$ -fold cross-validation is that it reduces the variance by averaging the validation accuracy for all  $k$  partitions. Also, each observation in the dataset presents at least once in both the training and the test set. Therefore, Study 2 performed stratified 10-fold cross-validation to evaluate the performance of the trained model. Stratified sampling maintains that the class-ratio are the same across training and test sets while generating the 10 subsets. Figure 4.3 illustrates its implementation.

##### **2- Evaluation Metrics**

The performance of a ML classification algorithm is commonly evaluated by different



## Chapter 4. Methodology

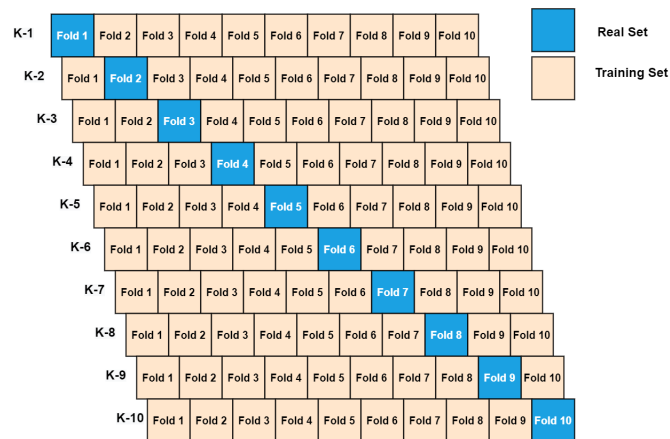


Figure 4.3: Ten-fold cross-validation in ML.

measures. These measures are generally used based on how well data is balanced. It is possible to have an equal chance of being present in both the training and testing samples, for instance, if the dataset is evenly distributed among the classes. Thus, accuracy can be an acceptable metric to assess the performance of a classifier. On the other hand, accuracy may be a misleading measure of the exact performance of the classifier if the dataset is not evenly distributed among the classes. Below, several evaluation metrics are described.

The simplest way to show the summary of prediction results is by using a  $2 \times 2$  confusion matrix as shown in Figure 4.4.

Where,

- True Positive (TP) is the number of true positives: class instances that are correctly predicted as true.
- True Negative (TN) is the number of true negatives: class instances that are correctly predicted as false.
- False Positive (FP) is the number of false positives: class instances that are false but predicted as true.
- False Negative (FN) is the number of false negatives: class instances that are true but predicted as false.

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)

Figure 4.4: Confusion matrix in ML.

Accuracy was used in this research, as it was the most intuitive performance measure. Precision, recall, F1-score, and MCC were used for further assessment. Metrics are defined as follows:

1. Accuracy measures the ratio of only correctly predicted instances to the total number of instances. Accuracy does not imply incorrect predictions. Therefore, it is an ineffective measure especially in the case of unbalanced datasets. For example, if the ratio of two classes in a dataset is 9:1, a classifier that predicts 100 samples will always predict the majority class, and because it lacks enough data to train on the minority class, it will eventually acquire an accuracy of 90%.

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

2. Precision measures the number of true positive instances in identified positive instances. It is defined by the following equation:

$$Precision = \frac{TP}{TP + FP}$$

3. Recall computes the number of true positive class instances that are classified

correctly; therefore, it is a better metric of classifier performance than accuracy. It is defined by the following equation:

$$Recall = \frac{TP}{TP + FN}$$

4. F1-score combines precision and recall as an overall assessment of the performance. It does not compute the arithmetic mean, but it is the harmonic mean of both measurements. The harmonic mean is used when one measure is very low and the other is very high, such as high recall and poor precision. It will skew toward the lower number to reflect the classifier's real performance. It is defined by the following equation:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

5. Matthews' correlation coefficient is a more reliable statistical measure that works on confusion matrices to evaluate binary classification performance. The classifier must make valid predictions on both the majority of negative and the majority of positive cases, regardless of their ratios in the overall dataset. Matthews' correlation coefficient yields a high score only if the prediction performed well in all four confusion matrix categories (TP, TN, FN, and FP), proportional to the amount of positive and negative items in the dataset (Chicco & Jurman, 2020). Matthews' correlation coefficient is calculated as follows:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Matthews' correlation coefficient values can range from -1 to 1, with 1 denoting perfect classification and -1 complete misclassification (Boughorbel, Jarray, & El-Anbari, 2017). An interpretation of MCC results is provided in Table 4.2.

Table 4.2: MCC interpretations.

<b>MCC</b>	<b>Interpretation</b>
< 0.3	Negligible
0.3-0.5	Weak
0.5-0.7	Moderate
0.7-0.9	Strong
0.9-1.0	Very good

### 4.2.2 Think-Aloud Interviews (Phase 2)

Data collection instruments in qualitative research include participant observations, interviews, and documented material (Creswell, 2014). The choice of an appropriate instrument for data collection is a critical step in ensuring that the data needed to answer the research questions will be gathered (Vanderstoep & Johnson, 2008). The think-aloud interview is the most effective method for understanding consumer judgment and decision-making processes (Kuusela & Paul, 2000). This section gives a brief explanation of the think-aloud interview, one of the methods used in Phase 2 (Study 3). Think-aloud interview method was originally developed by cognitive psychologists Ericsson and Simon (Ericsson & Simon, 1980). The think-aloud method has been used across various research fields, including psychology, human-computer interaction and IR. The think-aloud is a popular protocol for exploring the thoughts and cognitive processes of participants during a task (Ericsson & Simon, 1984). Think-aloud interviews is used to uncover potential factors impacting the credibility judgement of health-related content on websites and during internet searches (Ghenai et al., 2020; Kattenbeck & Elsweiler, 2019; Klawitter & Hargittai, 2018; Muntinga & Taylor, 2018). There are two types of the think-aloud interviews: the concurrent think-aloud (CTA) and the retrospective think-aloud (RTA) method (Ericsson & Simon, 1984). The CTA asks participants to verbalise their thought processes as they perform the task. This is considered to be a valid reflection of the participant's short-term memory at the time. The RTA requires participants to recall their thoughts as they performed a task during an immediate post-interview. This entails accessing both short-term and long-term

memory. The goal of CTA is to capture real-time thoughts of participants, allowing for richer descriptions of user experience. Participants are encouraged to express themselves in the moment rather than trying to recall their feelings during the experience (Van Someren, Barnard, & Sandberg, 1994). The one challenge in using CTA is that participants may find it challenging to express their thoughts concurrently while performing the task. On the other hand, RTA may elicit a more reasonable and reflective report on task information processing behaviour, as the computer-supported task has already been accomplished and cannot be changed (Peute, de Keizer, & Jaspers, 2015). The RTA method is helpful when participants do not articulate enough of their thoughts while they are performing the task. The purpose of the RTA interview is to capture the participants' decision processes and their thoughts after completing the task. Ericsson and Simon (1984) recommend combining CTA and RTA, into what is called the hybrid method (Alhadreti & Mayhew, 2018). Through the triangulation of concurrent and retrospective data, they claim, a way is provided to enrich the verbal data acquired and to strengthen the validity and reliability of the method. The typical way to use both types, CTA and RTA, are in two sessions: the concurrent session and the retrospective session as described by (Padilla & Leighton, 2017). The most significant part of the interview is the concurrent session, when the participant verbalizes their thoughts aloud in response to a problem-solving assignment. The participant has to do the task while verbally describing their thought processes as they go along, in real time. The interviewer should refrain from interjecting throughout this part of the interview with queries that would impede the verbalization of problem-solving processes. Only non-directed reminders to the participant to express thoughts while solving a problem should be used by the interviewer during this session (Padilla & Leighton, 2017). The retrospective session is the second part of the think-aloud interview and plays a secondary role. Here the participant is asked to recall how the solution of the problem-solving task was reached. The concurrent session is directly followed by the retrospective session (Padilla & Leighton, 2017).

Therefore, Phase 2 in the current research used a combination of two types of the think-aloud interviews to collect data. Concurrent session followed by retrospective

interviewing which in this case was in the form of a semi-structured interview as it is applied in (Tawfik, Gill, Hogan, S York, & Keene, 2019), it is also called (post-task interviews). In the semi-structured interview, the researcher will have a list of questions to be addressed, however the researcher may add inquiries to address the research questions (Heigham & Croker, 2009).

### **4.2.3 Content Analysis (Phase 1a and Phase 2)**

Content analysis is a research method for subjectively interpreting textual data via a systematic coding and pattern identification process (Hsieh & Shannon, 2005). Content analysis is thought to be a qualitative research method by many people, but quantitative and qualitative techniques can be included in the process of content analysis (Neuendorf, 2017). This section gives a brief explanation of content analysis approaches (quantitative and qualitative) and coding approaches (inductive and deductive). Content analysis is applied in Study 1 and Study 3.

Quantitative content analysis is ‘the process of establishing categories and then counting the number of instances under each of them’ (Silverman, 2015, p. 116). The purpose of quantitative content analysis is to develop a standardised codebook to code content systematically and then to describe it using statistics (Metag, 2016; Morgan, 1993). Data are usually categorised using an algorithmic search procedure rather than by reading the data, and are only analyzed quantitatively (Morgan, 1993).

Beyond just counting words, qualitative content analysis closely examines language with the goal of organizing massive volumes of text into a manageable number of categories that correspond to similar meanings. Data are classified using categories that are at least partly created inductively, and are typically applied to the data through careful reading (Morgan, 1993)

Both quantitative and qualitative content analysis have characteristics in common, inclusive of data sampling and collection (describing the origin and content volume to be gathered for analysis), coding process (describing the units of analysis, training coders, and constructing the coding scheme), and results validation (judging reliability and validity of the findings). These characteristics can fluctuate in accordance with the

purpose of the study (Hamad, Savundranayagam, Holmes, Kinsella, & Johnson, 2016).

For coding, there are two main approaches of content analysis that can be either deductive or inductive (Forman & Damschroder, 2007). Deductive code categories are determined and identified prior to data analysis. Categories are derived from existing theories or previous literature (Forman & Damschroder, 2007). Inductive code categories are generated from the data itself (Forman & Damschroder, 2007). The decision to follow a deductive or inductive approach of data analysis is determined by the goal for the study and previous knowledge of the issue being examined. Hsieh and Shannon (2005) divide qualitative content analysis into three distinct approaches: conventional (inductive), directed (deductive), and summative content analysis. The key difference between these approaches is in the way the initial codes are developed. The conventional (inductive) coding categories are obtained straight from the text data; this approach is appropriate when there is limited information on existing theories. Categories and names for categories are obtained by the emergence of the data. Related theories or other findings of the study are addressed in the discussion section of the study (Hsieh & Shannon, 2005). For the directed (deductive) approach, a theory or related research findings are used as a guide for the initial codes. This approach is appropriate for validating or extending a theory or theoretical framework. The summative approach aims to understand the use of words for their content, and in context, and this approach begins by identifying and quantifying certain words (Hsieh & Shannon, 2005).

In the current research, quantitative inductive content analysis has been applied in Phase 1a (Study 1) to analyse Twitter bio profiles that participate in dementia, and to divide them into categories to be statically described. This is because quantitative analyses provide numerical descriptions of generated codes. Qualitative inductive (conventional) content analysis has been applied as the main method in Phase 2 (Study 3) to analyse data generated from think-aloud interviews; this is because the study aims to generate the subjective interpretation of the content of text to understand the participants' reasoning behind their credibility assessment decisions.

D. R. Thomas (2006) summarised the following procedure for the inductive analysis of qualitative data: 1) data cleaning and construction of raw data files; 2) close reading

of text to become familiar with the content and comprehend the themes and events covered in it; 3) creation of categories or themes. The aim is to determine the upper level or more general categories, while multiple readings of the raw data illuminate the lower-level or specific categories. This is sometimes referred to as in-vivo coding. In inductive coding, categories are derived from actual phrases or meanings in specific text fragments. When there are large portions of text data, the coding process can be accelerated by using qualitative analysis software. 4) Continued revision and refinement of the category system: Subtopics, contradictory points of view and new insights might be found within each category. The core theme or essence of a category can be conveyed by appropriate quotations selected from the text and linked to superordinate categories with similar meanings.

### **4.3 Tools**

Three tools were used for data feature extraction, including Botometer as a bot detection tool and two text analysis tools, LIWC and Posit. A description of these tools is presented in the following subsections.

#### **4.3.1 Linguistic Inquiry and Word Count (LIWC)**

Linguistic Inquiry and Word Count is a text analysis application that provides an efficient and effective method of analysing a person's speech, both written and verbal, to identify potential variances in emotional, cognitive, and structural patterns in any form of language used in blogs, books and science articles (Pennebaker, Boyd, Jordan, & Blackburn, 2015). Linguistic Inquiry and Word Count was first developed in the early 1990s (Pennebaker et al., 2015). The latest version of LIWC was released in 2015, with significant changes from previous versions, including a new dictionary and software options. This is the version used in this research. The software is meant to swiftly and efficiently examine individual or numerous language files. It also tends to be transparent and adaptable in its operation, allowing the user to experiment with word usage in a variety of options. Linguistic Inquiry and Word Count contains a



multi-thousand-word dictionary. The LIWC2015 default dictionary has 6,400 words, word stems, and select emoticons. It also has the ability to read “netspeak”, that is, the language commonly found on SM such as Twitter and Facebook (Pennebaker et al., 2015).

In LIWC, words are sorted into two broad categories: content and style words. Content words include nouns, most verbs, adjectives and adverbs exhibiting content material of communication. Style words, also known as function words, are pronouns, prepositions, articles, conjunctions, auxiliary verbs, and some other esoteric categories. Style words, from a psychological standpoint, represent how individuals communicate, while content words express what they are saying. It’s no surprise, therefore, that style words are far more strongly tied to indicators of people’s social and psychological lives (Tausczik & Pennebaker, 2010).

Every target word passed to this tool is scaled across seven predefined main categories: summary language variables, general descriptor categories, standard linguistic dimensions, word categories tapping psychological constructs, personal concern categories, informal language categories, and punctuation categories. There are 90 categories and subcategories, therefore, each text file passed into LIWC2015 will by default generate approximately 90 output features. A detailed list of categories and subcategories can be found in Appendix A.

### **LIWC in Research**

Existing studies have widely used LIWC and confirmed its effectiveness for analysing text and extracting features to be utilised for credibility cues. For example, Pérez-Rosas, Kleinberg, Lefevre, and Mihalcea (2017) proposed a detection system for fake news websites using LIWC in combination with readability and syntactic features such as N-grams. They found that classifiers that rely on the semantic information provided in the LIWC lexicon perform consistently well across datasets from different domains (e.g., entertainment, business, politics). Del Pilar Salas-Zárate et al. (2017) trained different ML algorithms by only analysing word usage in LIWC to classify fake news into satirical and non-satirical types on Twitter. The model achieved a good accuracy F-measure yielding up to 85.5%. The work by Patro et al. (2019) used LIWC features to

identify exaggerated health news content on Twitter and found that there is significant difference in LIWC linguistic features of tweets sharing exaggerated news compared to tweets that share non-exaggerated news.

Therefore, the motivation for using LIWC in this research is its successful application in different studies with regards to identifying non-legitimate language. Hence, LIWC software<sup>1</sup> licensed for academic purposes is used in combination with the Posit toolset (described in the next section).

### 4.3.2 Posit Toolset

Posit is a text profiling toolset developed by George Weir of the Department of Computer and Information Sciences at the University of Strathclyde (Weir, 2007). The Posit system is designed to generate quantitative features at the word, sentence, and part-of-speech (POS) level of texts. The features target various related aspects of textual analysis. The Posit toolset uses Unix-based script. An example of posit run command is shown in Appendix B. Posit scripts are easily modularised and integrated with executable and off-the-shelf POS-taggers to make updating and maintenance reasonably simple. Posit is made up of a number of software modules that work together to give a wide range of textual analysis tools. The core module is POS Profiler, which focuses on POS and analyses a given text corpus to provide statistics on the POS features of that corpus (Weir, 2007; Weir, Dos Santos, Cartwright, & Frank, 2016).

The summary output of Posit ranges from values for total words (tokens), total unique words (types), and type/token ratio, to number and average length of sentences, number of characters, average word length, to POS: nouns, verbs, adjectives, adverbs, prepositions, personal pronouns, determiners, possessives, interjections, particles, and POS types totalling 27 features. Part-of-speech types indicate the total number of component types found in the analysed corpus for each major POS. For example, with nouns, Posit calculates common nouns (singular and plural), proper nouns (singular and plural). For verbs, Posit aggregates the total number of base form verbs, gerunds, past forms, past participles, present (3rd person), present (not 3rd person),

---

<sup>1</sup><http://liwc.wpengine.com/>

and modal auxiliary verbs. Similarly, aggregation is produced for adjectives, adverbs and pronouns.

An example of Posit output file is shown in Appendix C. A detailed list of Posit features can be found in Appendix D.

### **Posit in Research**

The features generated by Posit provides a basis for comparing samples of different text datasets. Posit’s comprehensive test analysis, for example, allows for a contrasting comparison of two or more text documents. This approach had been used, for example, to compare the grammatical approaches of many generations of Japanese English textbooks (Weir & Ozasa, 2010). Previous research has proven Posit analysis to be effective as a feature set for use in text credibility classification tasks of both websites and SM datasets. For example, Weir et al. (2016) used Posit to extract features from webpages with content that could be regarded as terrorist or extremist, in order to train a ML algorithm to detect terrorist webpages, which resulted in overall accuracy of 95.3%. A project by the International CyberCrime Research Centre (Cartwright, Weir, & Frank, 2019) used Posit and other tools to analyse a large sample of SM posts containing ‘fake news’ disseminated by Russia’s Internet Research agency to develop automated classification to identify hostile disinformation in real time . Using different types of features generated by Posit achieved an accuracy of up to 90.12%.

Based on Posit’s effectiveness in generating textual features and detecting fake information, Posit is used in the current research.

### **4.3.3 Botometer as Bot Detection Tool**

Botometer estimates the bot likelihood of a given Twitter account by using a supervised ML classifier, which learns from training datasets consisting of examples of bot and human accounts. Botometer extracts more than a thousand features from public profile data fetched by the Twitter application programming interface (API). Features are mainly grouped into six feature categories: network, friends, user, temporal, content and sentiment. Network features capture multiple dimensions of information diffusion patterns. Networks are built based on retweets, mentions, and hashtag co-occurrence,

and their statistical characteristics, e.g., degree distribution and centrality measures, are extracted. Friend features are computed features such as median and entropy of distribution of their number of posts and followers. Profile features are based on Twitter metadata related to an account, such as language, geographical location and account creation date. Temporal features are based on timing patterns of content creation and consumption, including inter-tweet time distribution and tweet rate. Content features are based on linguistic cues derived from natural language processing, particularly POS tagging. Sentiment features utilise generic and Twitter-specific sentiment analysis algorithms, such as happiness, emoticons, arousal, dominance, and valence. Finally, all features are used by ML algorithms to compute the bot scores, with high scores indicating likely bot accounts and low scores indicating likely human accounts (K.-C. Yang et al., 2019).

Bot detection methods are easily evaded, due to the evolution of Twitter bots. Therefore, Botometer releases new versions incorporating different training datasets for each new version, based on new research efforts and updated criteria (see Appendix E). The current research used Botometer v3 (K.-C. Yang et al., 2019) for Phase 1a and v4 (Sayyadiharikandeh et al., 2020) for Phase 1b.

The main reason for upgrading to Botometer v4 was that it periodically reequips the model with new annotated datasets. Botometer v4 implements new architecture, Ensemble of Specialized Classifiers (ESC), which aims to train specific classifiers for various classes of bots and collate their decisions based on their confidence. This is motivated by the observation that bot accounts have a variety of behaviours, each with definite characteristics, whereas human accounts exhibit similar behaviours. Accessing the Botometer API<sup>2</sup> is facilitated by Python script. It submits a Twitter screen name/user\_id to the API and returns a bot score as output. The variables in the Botometer API v4 response are given in (see Appendix F).

There are two types of bot scores computed, one utilises the English language, the other is Universal. The English score uses all six categories of features, whereas the Universal score provides a language-independent score that does not include the

---

<sup>2</sup><https://github.com/IUNetSci/botometer-python>

linguistic features such as sentiment and content features. Botometer is based on linguistic features if the language in the majority of recent tweets by the account is English. The main scores in all Botometer released versions are:

1. Overall scores can be in [0-5] scale, called the display score, or in the [0-1] scale, called the raw score.
2. Complete Automation Probability (CAP) score, which ranges from 0 to 1: an account with a score close to zero indicates the highest classifier confidence for a human, whereas a score close to one has a high probability of being a bot.
3. Botometer v4 additionally reports six sub-scores for each bot type which come from the specialised bot classifiers and estimate how similar an account is to different types of bots. The six bot type scores are as follows:
  - Astroturf score: political bots and accounts involved in following or systematically deleting political content in high volumes (the dataset used to train this class consist of political bots plus a subset of astroturf, see Appendix E)
  - Fake follower score: bot accounts purchased to increase follower counts (the dataset used to train this class are “cresci-17” and “vendor-purchased” see Appendix E)
  - Financial score: bots which post using cashtags (the dataset used to train this class are “cresci-stock”, see Appendix E). A cashtag is a dollar symbol in front of the abbreviation of a company’s name (e.g., the cashtag for Apple, Inc. is \$AAPL).
  - Self-declared score: self-identified bot accounts from botwiki.org., a database where interesting and creative Twitter bots are kept.
  - Spammer score: accounts labelled as spam bots from several datasets (the dataset used to train this class are “pron-bots” and a subset of “cresci-17”, see Appendix E)

- Other score: miscellaneous other bots obtained from manual annotation and user feedback, etc. are aggregated into the others category.

### **Botometer in Research**

There are few publicly accessible bot detection systems available at the time of this research. The first publicly available system was DeBot, developed by (Chavoshi et al., 2016), which employs an unsupervised approach to detect bot accounts on Twitter. This system calculates tweeting activity correlations among different accounts to reveal their coordination based on temporal patterns in their post timelines. The system's method is to find two or more accounts posting the same content around the same time, through constant monitoring by the Debot team, and then compiling a database of correlating accounts.

The second publicly available system, Botometer (Davis, Varol, Ferrara, Flammini, & Menczer, 2016), formerly known as BotOrNot, is a system that was made available to the public in 2014. The system utilises a supervised classification approach using more than a thousand features to evaluate how closely a Twitter account resembles the known characteristics of bots. Botometer scores are only applied to active accounts (i.e., not suspended, private, or otherwise shutdown accounts).

Both DeBot and Botometer provide open-source access to their hosted detection platforms to researchers via an API. However, Botometer was selected for the current research over DeBot for the following three reasons: In the first place, many pairs of accounts need to be considered by DeBot, which slows down the process of detecting coordination. Secondly, Bot-hunter, a multi-tiered supervised classification ML bot detection tool developed by researchers at Carnegie Mellon University (Beskow & Carley, 2020), was evaluated against the two state-of-art Debot (Chavoshi et al., 2016) and Botometer (Davis et al., 2016). In the evaluation, Botometer had a solid, steady performance when predicting new bots across all classification metrics, compared to the low accuracy achieved by the Debot model. However, Bot-hunter is not accessible via a public API (Beskow & Carley, 2018). Thirdly, Botometer is arguably the most popular bot detection method and much research has been done using this system. Botometer has been utilised, for example, to investigate the spread of misinformation in Twitter

posts during and following the 2016 US presidential campaign (Shao et al., 2018). Similarly, Botometer has been used to investigate the impact of SM bots on the 2016 US presidential election (Bessi & Ferrara, 2016). In that situation, using the proposed detection algorithm, it was discovered that a high number (about one-fifth) of the sources was non-human, yet they produced enormous quantities of information. In this specific instance, research showed the effective use of Botometer as a benchmark for topics not only related to politics (Bessi & Ferrara, 2016) but also to a few related health topics. Botometer was also utilised in the context of controversies around vaccination (Broniatowski et al., 2018). A random set of tweets were collected, using keywords such as “vax” or “vacc”. The tweets were tagged for their relevance to vaccination using a ML classifier. The credibility of these tweets was evaluated using Botometer. A score for each tweet was assigned, reflecting the likelihood of the source being a bot. This shows that Botometer is the standard of bot detection in many research projects in computational social science. Therefore, Botometer was regarded as the most suitable tool for this research project for bot analysis. However, because Botometer may result in false positives. To address this issue, two types of additional validation were carried out in this research: setting high conservative scores for determining bots and manually inspecting a sample of the data.

#### **4.4 Research Ethics Statement**

As the research involves human participation in think-aloud interviews in Phase 2 (Study 3), it requires ethical approval to maintain the participants’ dignity and safety while participating in the research. Therefore, ethics approval from the Ethics Committee in the Computer and Information Science Department at the University Of Strathclyde was granted (see Appendix G). Adhering to University’s Code of Practice on Investigations on Human Beings, the participants must sign a consent form to indicate their agreement to participate in the study (see Appendix H). The goal and procedures of Study 3 were explained and provided to the participants using the information sheet (see Appendix I).

## **Chapter Summary**

This chapter has presented the research design overview. As mentioned in Chapter 1, various methods were used to study dementia information credibility on Twitter. In this research, three studies were conducted. The first and second study followed a quantitative approach in which the Twitter API was mainly used for the data collection. A qualitative approach was applied in the third study, consisting of think-aloud interviews with dementia caregivers. Data collection, analysis methods, findings, and detailed procedures of the three studies will be presented in Chapter 5, 6, and 7.



## Chapter 5

# Study 1: Exploratory Study

### 5.1 Introduction

This chapter addresses research questions RQ1-RQ3 established in the first chapter. As stated in Chapter 1, the objective of Study 1 was to gain a broad overview of the research topic. The study assessed the role of bots in spreading information regarding dementia and identifies features that distinguish bots from humans. Features were then utilised as quality cues to classify dementia related tweets in Chapter 6. These are the research questions for Study 1.

- **RQ1:** What profile types participate in dementia-related discussions on Twitter?
- **RQ2:** Are there bot activities in the context of dementia information dissemination on Twitter? If so, what is the relationship between bot patterns and different profile types?
- **RQ3:** What profile features and content features contribute most to demonstrating bot-like behaviour?

At the beginning of the current research project, a considerable number of studies had investigated the role of bots in the political domain, yet rather less attention had been paid to health-related topics. In the political domain, Bessi and Ferrara (2016) discovered that bots generated a substantial amount of information on Twitter during

the 2016 US presidential election, potentially affecting online dialogues. The fact that bots consistently produced more favourable content in support of a specific candidate may have swayed the perceptions of those who were exposed to it, giving the impression that a candidate had genuine, grassroots support, while it was all artificially produced. Similarly, Shao et al. (2018) found that bots played a crucial part in the propagation of low-credibility content on Twitter during and after the 2016 US presidential election. During the 2017 French Presidential Election, Ferrara (2017) found that about 18,000 bots were present on Twitter and participated in disinformation campaigns.

In the health domain, some studies have investigated the role of bots in topics such as public-health-related behaviours, attitudes around vaccination policies, and issues about smoking. Broniatowski et al. (2018) investigated how bots and trolls contributed to #VaccinateUS related posts on Twitter. These were Russian bots which were identified by NBC News, documenting Russian interference in the US political system. It was found that Russian trolls and sophisticated Twitter bots were substantially more likely than the average user to publish posts concerning vaccination. It is possible that this legitimised vaccine hesitancy and thus contributed to reduced vaccination rates and increased numbers of vaccine-preventable diseases. Allem et al. (2017) compared topics regarding smoking discussed by bots and human users. When compared to human users, bots were significantly more likely to utilise hashtags that mentioned smoking cessation and new products.

Overall, long-term health conditions, such as dementia, have received less attention. Botometer (K.-C. Yang et al., 2019) in particular has helped many researchers to evaluate bot presence and discover proof of their influence in swaying public opinion and, for example, jeopardising the presidential election's legitimacy (Bessi & Ferrara, 2016) by spreading articles from low-credibility sources (Shao et al., 2018), promoting the vaccine debate (Broniatowski et al., 2018), or engaging in the discourse on e-cigarettes (Allem et al., 2017). As described in Section 4.3.3, Botometer's algorithm uses a set of more than 1,000 features, covering six broad categories of tweet content and sentiment, network patterns, temporal activity, and user and friends meta-data (K.-C. Yang et al., 2019). These features are aggregated and analysed to ascertain the likelihood that

the account in question is a bot. It is important to note that this algorithm does not employ features capturing the quality of a tweet (Varol, Ferrara, Davis, Menczer, & Flammini, 2017). Two categories, user metadata and content features, are proven to be most valuable sources of features to detect bots by Botometer (Varol et al., 2017). However, information about the individual features used by this algorithm is not publicly available. This research assumes that understanding the principal features used by the algorithm will considerably aid not only understanding the prime features selected to detect bots, but also using these features in analysing information quality. Principal features can be identified by analysing manifest features. Manifest features are a user's explicit attributes or statistical data which can be directly obtained through the Twitter API (Mei, Zhong, & Yang, 2015). In other words, this study aims to analyse the principal features indicating bot-like behaviour using different statistical techniques on Twitter data features (e.g., profile metadata and tweet content).

The current study sought to first understand and evaluate the role of bots in the dementia context using Botometer (K.-C. Yang et al., 2019). This involved quantifying bot involvement in the dissemination of dementia related information on Twitter and examining the relationship between groups of profile types, which were classified based on their profile descriptions and bot patterns. Then, a comprehensive analysis of the principal (manifest) features for evaluating bots is performed to address the issue of feature selection in analysing information quality. Identified bot features are used for evaluating information quality in Study 2.

An earlier version of this chapter was published in (Alhayan & Pennington, 2020).

## 5.2 Methods

Figure 5.1 illustrates the data collection and data analysis process in this study. Details are provided in the following sections.

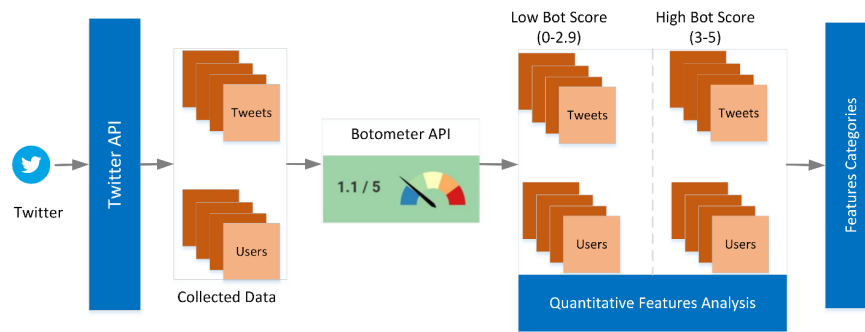


Figure 5.1: Study 1 method overview.

### 5.2.1 Data Collection

Publicly available tweets were collected through Tweepy library (Roesslein, 2009) in Python 3.7.5 and the Twitter Streaming API was used for real-time information extraction spanning eight weeks, from 16 December 2018 to 6 February 2019. The search query string consisted of keywords such as “dementia” OR “Alzheimer” OR “Alz” OR “Alzheimers”. The Python script ran on the University of Strathclyde server to guarantee the script kept running and the internet connection remained stable. Only original tweets in English were used and all retweets were discarded, resulting in a total number of 16,691 tweets. Each tweet’s author information was collected, along with bot scores.

The author information included the following profile features: bio description, verified, followers\_count, friends\_count, listed\_count, favourites\_count, statuses\_count, and geo\_enabled. Bio description is a short public summary about the profile owner. Verified indicates that the user has a verified account. Followers\_count specifies the number of followers of the account. Friends\_count is the number of users the account follows (in other words, their “followings”). Listed\_count shows the number of public lists of which this user is a member. Favourites\_count is the number of Tweets this user has liked in the account’s lifetime. Statuses\_count indicates the number of tweets (including retweets) issued by the user. Geo\_enabled refers to the possibility to attach a location (such as a city or neighborhood) when posting a tweet.

Bot scores for all tweet authors were obtained by using Botometer v3 (K.-C. Yang et al., 2019) and a Python script (3.7.5) that fed all user IDs to the Botometer API

Pro to compute bot scores. Since Botometer API Pro had no subscription fee and provided a rate limit of 2,000 requests per day, the Python script ran on the university server to keep the script running and ensure a stable internet connection. The output resulted in a CSV file with all user IDs and their scores. Scores for seven users were not generated, because these accounts were either suspended or deleted while the script ran. Duplicated users were removed, bringing the total to 16,691 tweets by 8,400 unique authors.

For the selection of Botometer cut-off values, as described in Section 4.3.3, the Python Botometer API queries the Twitter API to extract thousands of account features, classified into six categories, and feeds these features to ML classifiers, which generates a score for each category and an overall bot score. Overall scores can be in [0-5] scale, called the display score. These scores are used to determine whether the given account is a bot. A value of 0 indicates more human-like behaviour and 5 indicates high bot-like behaviour.

A great number of studies assume that accounts with a score equal or higher than the middle of the scale (2.5) have a higher probability to be bots (Shao et al., 2018; Sandim, Azevedo, da Silva, & Moro, 2018; Santini, Salles, Tucci, Ferreira, & Grael, 2020; Giachanou, Ghanem, & Rosso, 2021; Tshimula, Chikhaoui, & Wang, 2022). For the current study, the overall score from the 0-5 scale was used to evaluate the bot presence and feature analysis. Initially, the threshold of 2.5 was used, but after manual assessment of a sample of profiles consisting of 1,000 profiles, it was found that a threshold of 3 gave more realistic results. This threshold is also used by (Al-rawi, 2019). Therefore, scores ranging from 3-5 were considered to indicate the profile is a bot.

### 5.2.2 Profile Categorisation

Profile categorisation differentiates tweets' authors based on their relevant groups (types), such as organisations or individuals. Quantitative content analysis using inductive coding (described in Section 4.2.3) has been applied to systematically analyse a sample of dementia Twitter data produced by different profile types. The purpose of quantitative

content analysis is to develop a standardised codebook to code content systematically then to describe it using statistics (Metag, 2016).

The profile bio was chosen as the unit of analysis, because profile bios are the sole source of information to determine and validate profile type. Inductive coding is recommended when little prior information is available. Various studies have incorporated a similar coding approach and used profile bios for classifying Twitter users who discuss health information (Addawood, Balakumar, & Diesner, 2019; Addawood et al., 2019; Y. Liu, 2016; Park et al., 2016). All 8,400 profile bios were analysed using inductive coding to examine profile types participating in posting tweets related to dementia. Following that, bot patterns in different profile types were also analysed and visualised.

The coding procedures in the current study are as follows: once the text from profile descriptions had been collected, a sample of user profiles was read closely, and three broad categories (professional, non-professional and entities (organisations)) were defined to form an initial codebook. For general categorisation, the RegEx package<sup>1</sup> in Python, a character sequence that creates a search pattern, was used to split the three general categories based on high-level keywords or phrases defined in the query. The list of high-level keywords used for *professionals* and *organisation* type users is provided in Appendix J. These keywords or phrases were determined based on initial observations. For example, account descriptions containing keywords about areas of expertise such as professional titles (e.g., neurologist, therapist) were categorised as *professionals*. Similarly, certain words or phrases (e.g., organisation, association) were identified as *organisations*. The rest of the users were classified as *individuals*. The profiles gathered through the search queries were further evaluated by multiple readings of the raw data. Different text segments were manually marked and copied into emerging categories. As a result, more low-level categories from organisations (entities) were derived. An example of the list of low-level keywords used for the organisation type users is provided in Appendix J. The query in the Python script was modified with more segregation keywords for five additional categories: *general organisation*, *care providers*, *promoters*, *media*, *books/apps*.

---

<sup>1</sup><https://docs.python.org/3/library/re.html>

A category *empty/unknown* was created for profile descriptions that had been left empty or profiles that did not belong to any identified category (e.g., a conference profile). A second low-level categorisation was developed to identify six subcategories in the individuals category, namely *health activists, caregivers, artists, marketers, authors and others*. Finally, a total of eight main categories were created (*professionals, individuals, general organisations, care providers, promoters, media, books/apps, and empty/unknown*). For each category, two files were created: one for users and one for their tweets. The final codebook, comprising all possible categories of users is shown in Appendix K.

It is possible that automated categorisation with the help of Python scripts results in outliers or incorrectly categorised accounts. Therefore, in the process of continued revision and refinement, the researcher reviewed all entries manually to ensure that these instances were restricted to an acceptable maximum to ensure reliable final labelling. The researcher used the codebook as a guide for reviewing the user categorisation manually. For data sample validation, the data of 2,000 users from the entire dataset of 8,400 users were sampled by an annotator (a PhD student in computer and information science). The codebook was provided to this annotator and they assigned a category to each user in the sample, based on the codebook. Cohen's Kappa (Cohen, 1960) was applied to the resulting data to determine the level of agreement between two annotators in assigning categories to profiles. The annotators' categorisation results were intersected with the categories that were pinpointed, for common identifiers using an R Studio script<sup>2</sup>. The results indicated a 76% agreement between the annotators, which is a substantial level of agreement and aligns with the established acceptable percentage for Twitter data categorisation (Y. Liu, 2016; Park et al., 2016).

### 5.2.3 Profile Feature Analysis

Three datasets with different overall bot-score ranges were used for profile feature analysis: one for all profiles, one for profiles with low scores (0-2.9) and one for profiles with high scores (3-5). The objective of this analysis was to find the most important principal

---

<sup>2</sup><https://www.rstudio.com/>

features for evaluating bots by examining the relationship between profile features and overall bot score, using stepwise multiple linear regression (SMLR). Stepwise multiple linear regression is applied in different studies that analyse Twitter features in relation to scores generated by different tools, for instance, tools that generate scores to quantify users' influence on Twitter (Mei et al., 2015). Since there is no publicly available information on exact features employed by these tools (Mei et al., 2015), various statistical tests have been employed to analyse the principal features used by different tools that generated the influence score. For example, Mei et al. (2015) used SMLR to identify the most important predictors (i.e., followers, total tweets) for an influence score generated from four popular influence services (Klout, PeerIndex, Kred, Followerwonk) and to learn more about components that have strong links to the dependent variable (user influence score). Similarly, Lahuerta-Otero and Cordero-Gutiérrez (2016) used SMLR to analyse the different independent variables of user tweets (i.e., lexical diversity and average number of words, to find which variables have impacted on the dependent variable (user influence score). User influence scores generated by the well-known tool PIAR, which determines user influence and popularity on Twitter. PIAR measures the influence on a scale from 0 to 100. The study aimed to analyse which tweet-related features are shared by influential users, using SMLR. Influence scores were used as dependent variables to analyse the different independent variables of users' tweets, such as lexical diversity and average number of words. The current study followed a similar approach to test which principal profile features could predict profiles that were likely to be bots by applying SMLR analysis through the IBM Statistical Package for the Social Sciences (SPSS)<sup>3</sup>. This analysis was performed due to the unavailability of information on the exact features employed by the algorithms used by the Botometer's bot detection tool (described in Section 4.3.3) and in order to examine which features can be used as a bot indicator in the context of dementia information. The bot score generated by the Botometer was used as the dependent variable and profile features were used as the independent variable. Seven publicly available profile features provided in raw user metadata were selected as candidate features for the evaluation, namely  $X_1 = \text{verified}$ ,

---

<sup>3</sup><https://www.ibm.com/analytics/spss-statistics-software>



X2 = followers\_count, X3 = friends\_count, X4 = listed\_count, X5 = favorites\_count, X6 = statuses\_count, and X7 = geo\_enabled.

In SMLR, variables are added to the regression equation one by one, utilising the statistical criterion of maximising the R2 (coefficient of determination) of the added variables (Montgomery, Peck, & Vining, 2013). The process of adding additional variables comes to an end when all variables have been added or when no other variables can be used to achieve a statistically significant improvement in R2. The result is the variable with the highest R2, which indicates most of the target variables can be explained. Measures of regression and a description are provided in Table 5.1.

Table 5.1: Descriptions of SMLR measures.

Variable	Description
Standardized Coefficient Beta ( $\beta$ )	when all other independent variables are maintained constant, a standardized beta coefficient show how much the dependent variable fluctuates due to a change of independent variable in terms of standardized or standard deviation units. The larger the value, the greater the reliance. A beta of -.9, for instance, has a greater influence than a beta of +.8.
Unstandardized Coefficient (B)	when all other independent variables are maintained constant, unstandardized coefficients show how much the dependent variable fluctuates due to a change of independent variable in terms of e in terms of one units of changes.
R-Square	the percentage of variance in the outcome that is explained for by the predictor variables.
T Value	t-test score to determine the confidence of b or ( $\beta$ ) to predict the weight of influence
Sig.	shows the statistical significance of b or ( $\beta$ ) related to each predictor variable. If a p-value is less than .01, then that variable has a significant relation with the depended variable.
Standard Error	these are standard errors for the coefficients. It contains the error values associated with the unstandardized beta coefficients.

#### 5.2.4 Content Feature Analysis

Two equal-length tweet sets were selected randomly from the larger collected tweet set, one belonging to bot-likely authors with a bot score ranging between 0-2.9 and the other belonging to human-likely authors with bot score ranging between 3-5 for content feature analysis. The objective of the content analysis was to determine how

variations in linguistic features and domain-specific features (such as URLs) indicate bot-like behaviour compared to profiles controlled by humans.

The choice of these features followed K. Shu, Sliva, Wang, Tang, and Liu (2017), who noted that linguistic features and domain-specific features were the most common features used to capture low-quality news, which is likely to be made and disseminated by bots (K. Shu et al., 2017). Two typical linguistic feature types contribute to various classification tasks in NLP in a given text, namely syntactic features and lexical features (K. Shu et al., 2017). Syntactic features are sentence-level features and include features that describe the connection between words and their role in a sentence. The process of annotating syntactic categories for per word in a corpus is known as POS tagging (Dhanalakshmi, Kumar, Shivapratap, Soman, & Rajendran, 2009). Part-of-speech tagging is mapping a word based on its the location in a sentence, as well as its definition. For example, in the phrase “John likes Apple”, “John” and “apple” will be tagged as nouns (NN) and “likes” as a verb (VB). Lexical features make up character-level and word-level features such as total number of unique words, vocabulary richness and number of characters per word. Domain-specific features are aligned to a particular domain, such as external links. In the Twitter context, domain-specific features (also called Twitter-specific features) are features unique to the Twitter platform that are related to tweets, including URLs, hashtags, and mentions.

To assess linguistic features in both tweet sets, bot-likely and human-likely tweets, the Posit toolset, which is introduced in Section 4.3.3, was used. Posit performs quantitative textual analysis on very large data sets to generate 27 features (see Appendix D). Posit targets various aspects of textual analysis, it applies a POS tagger to text and generates quantitative information about the textual content in terms of individual words (e.g., total words and total unique words) and overall sentence (e.g., average sentence length), POS, and POS types (the total number of subcategories of each main POS classification).

All tweets were pre-processed; URLs and mentions (@) were removed, since they could affect the number of tokens in Posit.

In some cases, Twitter’s limitation of 280 characters per tweet incentivises a user

to provide a URL with more detail relating to their tweet. Thus, the frequency of a particular domain in a URL appearing within tweets could be a good indicator of tweets' quality. The objective of the URL analysis was to compare the frequency and types of URLs disseminated by profiles with high bot scores compared to the profiles with low bot scores. The URLs were counted in the same two tweet sets that were used for the linguistic analysis. In order to find the frequency of a particular URL type used within tweets, all URLs embedded in the tweet dataset were extracted ( $N = 16,691$ ). These URLs were cleaned, and only the domain name was shortlisted for analysis. Also, if the same user included the same URL more than once, it was counted as one.

## 5.3 Results

### 5.3.1 Descriptive Analysis

The process of inductive coding of the 8,400 profile descriptions resulted in eight distinct profile types (*professionals, individuals, organisations, providers, promoters, media, apps/books, and empty and unknowns*). The *individuals* category contained the highest number of users and the *apps/books* category had the lowest number of users. Descriptive statistical analysis was then performed to assess the bot scores among categorised users through the *overall bot-score* mean, the percentage and means of users with a high bot score in the range 3-5, and bot scores of profiles of all categories that were identified through inductive coding Table 5.2.

The distribution of profiles with bot-like scores in different categories is presented as bar graphs for better understanding in Figure 5.2. All categories show similar bot-score patterns, negatively correlating with the percentage of profiles in that category, except for *care providers* and *apps/books*. These two categories also had a comparatively high average bot score, *care providers* (38.5%) and *apps/books* (39.83%). The *promoters* category also showed a high percentage of average bot score (%25.76). In contrast, the *individuals* and *professional* categories exhibited a very low percentage of profiles with high bot scores, (%7.51), and (%9.05), respectively. Overall, the results show the mean

Table 5.2: User categories and bot scores.

Category	Total users	Overall bot score mean	% (3-5) Bot score	(3-5) Bot score mean
All	8,400	1.32	13.85	3.73
Individuals	3,899	0.91	7.51	3.71
Organisations	1,223	1.62	13.39	3.65
Care providers	831	2.48	39.83	3.73
Media	474	1.63	15.3	3.75
Professionals	784	1.07	9.05	3.83
Promoters	520	1.95	25.76	3.75
Empty/unknown	593	1.18	10.79	3.70
Apps/books	76	2.35	38.15	4.00

overall bot score of the 8,400 profiles was 1.32, indicating a general tendency towards human profiles.

### 5.3.2 Profile Feature Analysis Result

The objective of performing empirical analysis using SMLR was to find the most useful set of predictors (principal profile features) for the dependent variable (*overall bot score*) and removing the least significant predictors. Stepwise multiple linear regression analysis was conducted on three different datasets. The first dataset contained all the profiles with bot scores ranging from 0-5 (see Table 5.3). This provided a general understanding of the variables that impact overall bot score. After eliminating non-contributing features, the results of SMLR analysis, as shown in Table 5.3, reveal the model consists of four contributing predictors,  $X_1 = \text{verified}$ ,  $X_3 = \text{friends\_count}$ ,  $X_5 = \text{favorites\_count}$ , and  $X_7 = \text{geo\_enabled}$ , and reaches adjusted R2 square of .053. This means the model could predict 5.3% of the variances of users' bot scores. The standard coefficient ( $\beta$ ) in Table 5.4 reveals that the variables  $\text{geo\_enabled}$ ,  $\text{favorites\_count}$  and  $\text{verified}$  indicate a negative relationship with the bot score ( $\beta = -0.187$ ,  $\beta = -0.109$ , and  $\beta = -0.048$ , respectively), while the  $\text{friends\_count}$  ( $\beta = 0.074$ ) had a positive relationship with bot score. The F-value (the overall statistical significance of the model as a whole) was  $p < 0.01$ , indicating the statistical significance for factors  $X_7$ ,  $X_5$ ,  $X_3$ , and  $X_1$  that contributed to the prediction of dependent variable Z (bot score). Thus, the

Chapter 5. Study 1: Exploratory Study

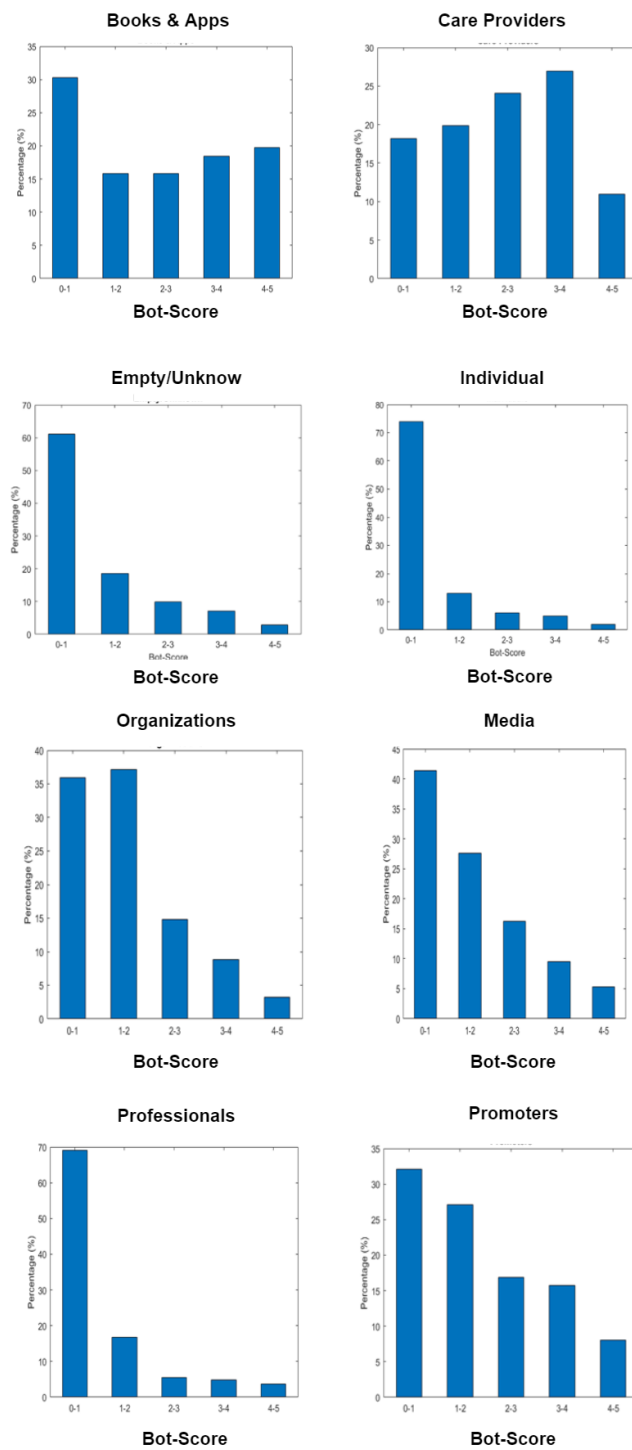


Figure 5.2: Bot score-wise distribution in each category.

Chapter 5. Study 1: Exploratory Study

regression equation is defined as follows:

$$\text{Estimated } Z = A1 \times X7 + A2 \times X5 + A3 \times X3 + A4 \times X1 + A5 \times X2 + C \quad (1)$$

$A1 = -0.448$ ,  $A2 = -0.00000452$ ,  $A3 = 0.00000934$ ,  $A4 = -0.281$  and  $C$  is the value of the constant of regression = 1.563.

Table 5.3: Results of SMLR analysis of total dataset

No	Variable Entered	R	R-Square	Adjusted R-Square
1	X7	.192	.037	.037
2	X7 + X5	.215	.046	.046
3	X7 + X5 + X3	.226	.051	.051
4	X7 + X5 + X3 + X1	.231	.053	.053

Table 5.4: Coefficient and constant values for regression equation

Variables	Unstandardized Coefficients B	Standard Error	Standardized Coefficient Beta ( $\beta$ )	T Value	Significance Constant
Constant (C)	1.563	0.18		85.685	.000
X7	-.448	0.025	-0.187	-17.578	.000
X5	-0.00000452	0.000	-0.109	-10.068	.000
X3	0.00000934	0.000	0.074	6.815	.000
X1	-.281	0.062	-0.048	-4.522	.000

Furthermore, the dataset was split into two subsets: one ranging from 0-2.9 and the second from 3-5. The purpose of conducting SMLR on split data was to highlight the profile features that contribute to human and bot profile behaviour (see Table 5.5 and 5.6). The SMLR for profiles with a bot score of 0-2.9 shows the adjusted R2 value as 0.018 (see Table 5.5), which indicates the total variance for the bot-score variable is 1.8% for all variables. The bot-score analysis in the 3-5 range provides an adjusted R2 value of 0.019 (see Table 5.6), which indicates the total variance for the bot-score variable is 1.9% for all variables. X4 (listed\_count) affected high bot-score ranges,

whereas it did not affect low bot-score ranges and the whole dataset.

Table 5.5: Results of SMLR analysis for the profiles with low bot scores

No	Variable Entered	R	R-Square	Adjusted R-Square
1	X7	.101	.010	.010
2	X7+X5	.132	.017	.017
3	X7+X5+X3	.137	.019	.018

Table 5.6: Results of SMLR analysis for the profiles with bot-likely profiles

No	Variable Entered	R	R-Square	Adjusted R-Square
1	X7	.104	.011	.010
2	X7+X4	.132	.017	.016
3	X7+X4+X5	.146	.021	.019

Another important finding was the negative coefficient values of X1 and X5 (geo\_enabled and favorites\_count), for all three sets: the whole dataset, bot score 0-2.9, and bot score 3-5 datasets. Table 5.7 summarises the SMLR analysis using the three samples showing variables with and without effect on the depended variable (bot score). It is important to note that the features X5 (favorites\_count) and X7 (geo\_enabled) affected all bot-score ranges. This suggests that these two features can be more effective in terms of determining bot- and human-likely profiles.

Table 5.7: Summary of SMLR analysis results of the all three datasets.

	Variables without effect	Variables with effect
All profiles	X4, X6, X2	X7, X5, X3, X1
Profiles with extreme scores	X1, X2, X3, X6	X7, X4, X5
Profiles with low scores	X1, X2, X4, X6	X7, X5, X3

### 5.3.3 Content Features Analysis Result

Quantitative analysis on the content was performed to find the variations in linguistic features of tweets and frequency and types of different URLs embedded in tweets posted by bot profiles compared to human profiles.

#### 5.3.3.1 Linguistic Feature Analysis

The features generated from Posit provide a basis for comparing samples of tweets with different bot score ranges: those belonging to bot-likely authors with a bot score ranging between 0-2.9 and those belonging to human-likely authors with a bot score ranging between 3-5. The summary output of Posit included 27 different features in total and is shown in Table 5.8 below. To validate the results, another two equal-length tweet sets, this time those with extreme scores, one with bot scores ranging from 0-1 and the other with scores ranging from 4-5, were randomly selected and analysed in Table 5.9 below.

Results showed clear variations of the linguistic feature count (e.g, word-level, sentence-level, POS, and POS type features) between bot and human tweets. Looking at both tables, it is clear that the number of characters in likely-bot tweets (145,993 characters in the non-extreme set and 230,234 in the extreme set) was higher than those in human profile tweets (116,110 characters in the non-extreme set and 216,889 in the extreme set). In contrast, the total number of words (tokens) was higher for tweets posted by likely-human profiles (24,859 tokens in the non-extreme set and 23,498 in the extreme set) than for tweets posted by the likely-bot profiles (22,006 tokens in the non-extreme set and 20,397 in the extreme set). It was also demonstrated in the average word length (AWL) in human tweets which appeared lower (4.67 in the non-extreme set and 9.23 in the extreme set) compared to bot tweets (6.63 in the non-extreme set and 11.28 in the extreme set). The reason could be that bot profiles used longer words compared to regular profiles. Humans tend to use a higher number of short words. However, bot profile tweets utilised fewer unique words in overall tweets (18.18% in the non-extreme set and 22.81% in the extreme set) compared to regular profiles (19.94% in the non-extreme set and 28.58% in the extreme set). This indicates that bots utilise



Chapter 5. Study 1: Exploratory Study

Table 5.8: Linguistic analysis of tweets from profiles with non-extreme high/low bot scores.

Feature	Bot score 0-2.9		Bot score 3-5	
	Count	Percentage	Count	Percentage
Total words (tokens)	<b>24,859</b>		<b>22,006</b>	
Total unique words	<b>4,959</b>	<b>19.94</b>	<b>4,001</b>	<b>18.18</b>
Type/token ratio (TTR)	5.01291		5.50012	
Number of sentences	<b>1,431</b>	<b>5.75</b>	<b>1,994</b>	<b>9.06</b>
Average sentence length (ASL)	17.37176		11.03610	
Number of characters	<b>116,110</b>		<b>145,993</b>	
Average word length (AWL)	<b>4.67074</b>		<b>6.63424</b>	
Nouns	<b>9,584</b>	<b>38.55</b>	<b>9,521</b>	<b>43.26</b>
Verbs	<b>4,160</b>	<b>16.73</b>	<b>3,410</b>	<b>15.49</b>
Prepositions	2,680	10.78	2,330	10.58
Determiners	1,672	6.72	1,327	6.03
Adjectives	1,506	6.05	1,318	5.98
Adverbs	995	3.99	664	3.01
Personal pronouns	<b>968</b>	<b>3.89</b>	<b>618</b>	<b>2.80</b>
Possessive pronouns	476	1.91	345	1.56
Particles	91	0.36	56	0.25
Interjections	10	0.04	2	0.01
Noun types	<b>3,683</b>	<b>14.81</b>	<b>2,966</b>	<b>13.47</b>
Verb types	1,293	5.20	1,010	4.58
Adjective types	646	2.59	528	2.39
Adverb types	176	0.70	135	0.61
Preposition types	65	0.26	62	0.28
Personal pronoun types	23	0.09	22	0.01
Determiner types	22	0.08	17	0.07
Possessive pronoun types	13	0.05	9	0.04
Particle types	9	0.03	8	0.03
Interjection types	7	0.02	2	0.01

specific words in their tweets, whereas human profiles utilise diverse words in their tweets. Furthermore, both tables showed that human tweets use fewer numbers of sentences (5.75% in the non-extreme set and 14.88 % in the extreme set) compared to bot tweets ( 9.06% in the non-extreme set and 16.89% in the extreme set). Yet, the average sentence length (ASL) of human tweets were longer (17.37 in the non-extreme set and 6.71755 in the extreme set) contrary to bot tweets (11.03 in the non-extreme set and 5.91 in the extreme set).

In terms of POS, the percentage of most forms of POS (e.g, verbs, prepositions, determiners, adjectives, adverbs, personal pronouns, possessive pronouns, particles, interjections) is higher in human tweets compared to bot tweets with the exception of

Table 5.9: Linguistic analysis of tweets from profiles with extreme high/low bot scores.

Feature	Bot score 0-1		Bot score 4-5	
	Count	Percentage	Count	Percentage
Total words (tokens)	<b>23,498</b>		<b>20,397</b>	
Total unique words (types)	<b>6,716</b>	<b>28.58</b>	4,653	<b>22.81</b>
Type/token ratio (TTR)	3.49881		4.38362	
Number of sentences	<b>3.498</b>	<b>14.88</b>	<b>3.446</b>	<b>16.89</b>
Average sentence length (ASL)	6.71755		5.91904	
Number of characters	<b>216,889</b>		<b>230,234</b>	
Average word length (AWL)	<b>9.2301</b>		<b>11.2876</b>	
Nouns	<b>11,674</b>	<b>49.68</b>	<b>12,238</b>	<b>59.99</b>
Verbs	<b>4,587</b>	<b>19.52</b>	<b>3,641</b>	<b>17.85</b>
Prepositions	2,296	9.77	1,762	8.63
Determiners	1,848	7.86	1,527	7.48
Adjectives	1,489	6.33	1,223	5.99
Adverbs	1,073	4.56	615	3.01
Personal pronouns	<b>1,011</b>	<b>4.30</b>	<b>481</b>	<b>2.35</b>
Possessive pronouns	469	1.99	221	1.08
Particles	82	0.34	53	0.25
Interjections	12	0.05	3	0.01
Noun types	<b>4.342</b>	<b>18.47</b>	<b>2.782</b>	<b>13.63</b>
Verb types	1.327	5.64	802	3.93
Adjective types	764	3.25	602	2.95
Adverb types	228	0.97	135	0.66
Preposition types	77	0.32	55	0.26
Personal pronoun types	31	0.31	20	0.09
Determiner types	21	0.08	18	0.08
Possessive pronoun types	16	0.06	9	0.04
Particle types	9	0.03	5	0.02
Interjection types	6	0.02	2	0.01

\*Numbers in bold indicate important features where feature values between human vs bot in both tables (both score ranges) are either larger/smaller, and the difference between percentages exceed 1% in both tables.

nouns. POS forms, including verbs and personal pronouns showed about 1-2% differences between both human and bot tweet datasets (extreme and non-extreme). For example, the total number of personal pronouns used in likely-human tweets (3.89% in the non-extreme set and 4.30% in the extreme set) was greater than the number used in likely-bot tweets (2.80% in the non-extreme set and 2.35% in the extreme set). Likewise, the number of verbs in likely-human tweets (16.73% in the non-extreme set and 19.52% in the extreme set) was found to be higher than that in bot tweets (15.49% in the non-extreme set and 17.85% in the extreme set). Other forms (prepositions, determiners, adjectives, adverbs, personal pronouns, possessive pronouns, particles, in-

terjections) showed small differences (less than 1%) between both human and bot tweet datasets. However, only one POS form, nouns, was used more in tweets from likely-bot profiles (43.26% in the non-extreme set and 59.99% in the extreme set) compared to the use in regular profile tweets (38.55% in the non-extreme set and 49.68% in the extreme set).

In terms of POS types, the percentage of most POS types (verb types, adjective types, adverb types, personal pronoun types, possessive pronoun types) was higher in human tweets compared to bot tweets. However, the percentage difference was small (less than 1%) but only one POS type, noun types, was higher (more than 1%) in human tweets compared to bot tweets. Contrarily, a few POS types (preposition types, determiner types, particle types), were used differently in both tweet sets (extreme and non-extreme), therefore they can be considered as less important features. The results indicated the percentage of noun types (14.81% in the non-extreme set and 18.47% in the extreme set) to be higher (more than 1%) compared to noun types, which were less diverse in bot tweets (13.63% in the non-extreme set and 13.47% in the extreme set). The diversity in noun types in human tweets was higher compared to bot tweets.

Overall, linguistic features, including number of words (tokens), unique word, POS, and POS types were higher in likely-human tweets than those in likely-bot tweets. Only a few features, including number of characters, number of sentences, AWL and one POS form (noun), displayed the opposite in likely-human tweets. More explicitly, ten linguistic features showed clear differences between both human and bot types: number of characters, number of unique words, number of words, number of sentences, AWL, ASL, POS (including nouns, verbs, personal pronouns) and POS type (noun-types).

### **5.3.3.2 Domain-Specific Feature (URLs) Analysis**

The objective of the URL analysis was to compare the frequency and type of URLs disseminated within dementia tweets by profiles with high bot scores compared to the profiles with low bot scores. In the first two sets, tweets by profiles with scores between 0-2.9 had 417 URLs and tweets by profiles with scores between 3-5 had 503 URLs. In tweets from profiles with extreme scores, 0-1 and 4-5, the number of URLs was 329 and

371, respectively. This suggests that bot profiles include more URLs in their tweets.

Table 5.10 shows the most common URLs along with their frequency in extreme datasets only because there was no clear difference in terms of URLs types within extreme and non-extreme datasets. The most common URLs along with their frequency in extreme dataset are shown in Table 5.10. Most URLs in both tweet datasets referred to social networks and online communities such as Instagram, Facebook and LinkedIn, and to arts and entertainment sites such as YouTube. A possible explanation for these two website domain types being found more frequently is that profiles reposting posts also shared on other platforms such as Facebook and LinkedIn and Twitter, either manually or using automatic tools. Likewise, news and media websites such as The New York Times, the BBC, Medical News Today, and BioPortfolio are commonly used in both tweets sets. However, these news websites traditionally are more authoritative information sources.

Table 5.10: URL analysis of tweets.

Bot score	URL	Frequency	Domain type
0-2.9	myalzheimersstory.com	371	Personal blog **
	www.amazon.com	270	E-commerce and shopping **
	www.youtube.com	254	Arts and entertainment **
	www.instagram.com	205	Social networks and online communities **
	www.bioportfolio.com	171	News and media**
	www.facebook.com	155	Social networks and online communities**
	www.bbc.co.uk	136	News and media**
	www.nytimes.com	134	News and media**
	memorycafedirectory.com	122	A place for individuals with Alzheimer's*
3-5	www.linkedin.com	101	Social networks and online communities**
	www.youtube.com	192	Arts and entertainment**
	www.alzheimers.net	105	Health/geriatric and aging care*
	www.amazon.com	75	E-commerce and shopping**
	www.medicalnewstoday.com	71	News and media**
	dailyaring.com	50	Health/geriatric and aging care*
	cynthiakraack.com	44	Personal blog**
	www.brightstarcare.com	44	Health/geriatric and aging care*
	www.facebook.com	42	Social networks and online communities**
www.gofundme.com	40	Social fund-raising platform*	
www.nytimes.com	31	News and media**	

\*\* domain appeared in both datasets \* domain appeared in one dataset.

Profiles from both tweet datasets also share on e-commerce websites (Amazon.com).

## Chapter 5. Study 1: Exploratory Study

The researcher investigated Amazon URLs manually to check what kind of Amazon products are shared in tweets. It was found that in both tweet sets, various Amazon products were shared. The products promoted mostly in tweets are e-books and audio-books, including novels about dementia, personal stories about dementia, or practical guides for homecare like books on nutrition, hydration and food enjoyment for PWD. Besides books, Amazon products that are commonly shared are dementia support tools, such as specialised toys and digital alarm clocks designed specifically for PWD.

Interestingly, websites with dementia-specific information such as [alzheimers.net](http://alzheimers.net), [dailycaring.com](http://dailycaring.com), and [brightstarcare.com](http://brightstarcare.com) and social fund-raising sites such as [gofundme.com](http://gofundme.com) were found only in profiles with a high bot score (3-5). There is a possibility that these types of websites use automated tools to publish the content of their sites simultaneously to their Twitter feeds. It is likely that these automated mechanisms are responsible for the proliferation of their own site-specific URLs.

Lastly, two personal blog websites were found as most frequently shared URLs in both dementia tweet sets, namely [myalzheimersstory.com](http://myalzheimersstory.com) (a blogger advocate for dementia care) and [cynthiakraack.com](http://cynthiakraack.com) (an author). Most websites in tweets are well known and thus easy to assess for credibility, but some websites, such as personal blogs, do not indicate if these websites are either credible or scam sites. The reputation of these two sites was investigated using Nibbler<sup>4</sup>, a tool to assess the technological quality of the website on different criteria: accessibility, experience, marketing, and technology. Each category analyses several criteria, such as code quality, headings, internal links, mobile accessibility (how the website appears when accessed via mobile), page titles, URL format, amount of content (amount of text and images when compared to an equalised distribution, a higher ratio of images to text or vice versa appears to be less effective), freshness, printability, meta-tags, server behaviour, Twitter presence (whether or not the website is associated with a Twitter account), analytics, domain age, incoming links (other links or other sites pointing towards these sites), and popularity. Popularity is determined by the number of new visitors who visit the site on a daily basis, as well as its Google Search Index ranking; a site that uses meta-tags has a higher Google Search

---

<sup>4</sup><https://nibbler.silktide.com>

Index ranking. A score of 0 to 10 is assigned to each criterion, with higher numbers signifying higher quality. The tool also provides an overall score on a 10-point scale, with 0 being the least credible and 10 being the most credible. The tool has been used for the same purpose in different studies (Abuqaddom, Alazzam, Hudaib, & Al-zaghoul, 2019; Wagle, Kaur, Kamat, Patil, & Kotecha, 2021).

Overall scores for the personal blogs (cynthiakraack.com and myalzheimersstory.com) are 7 and 7.9, respectively, indicating that these websites are most likely of high technical quality. Individual scores for incoming links are 144 and 6,365 for cynthiakraack.com and myalzheimersstory.com, respectively. Popularity scores (indicating how popular this website is compared to other websites, and whether popularity is rising or falling) are 3.1 out of 10 for cynthiakraack.com and 4.3 out of 10 for myalzheimersstory.com with a slight increase in popularity. This suggests that both websites are popular in terms of number of visitors. The (myalzheimersstory.com) pointed by many websites compared to (cynthiakraack.com). That being said, the quality of information provided by websites, including well-known sites such as Amazon.com, is not guaranteed. Specifically, low quality dementia related products are constantly being promoted online. For example, Block, Albanese, and Hume (2021) found dementia supplements promoted online through websites such as Amazon have insufficient evidence of efficacy and are expensive. It can be concluded that domain type is not an enough indicator of information quality.

## 5.4 Discussion

The objective of the first and second research question (RQ1) and (RQ2) in this study is to provide a broad picture of profile types who participate in the dissemination of dementia information and to evaluate the bot presence in different groups. A quantitative content analysis using inductive coding allowed the researcher to understand user types and to present a total and bot pattern within each category. Eight distinct profile types were identified based on a codebook developed using inductive coding. The *individual* category contained the highest number of users, and the *apps/books* category had the

least number of users. This suggests that most users engaged in the dementia community on Twitter are *individual* users, answering RQ1. As defined in the user categorisation book code, the individual category consists of the sub-categories *health activists, caregivers, artists, authors, and others*. Understanding profile types and evaluating the presence of bots among these different groups in specific dementia communities on Twitter can help to explain the large-scale disparity of information quality. Because bots are indicators of information quality, it follows that non-human accounts such as bots are the most likely originators and disseminators of low-quality information in news contexts (K. Shu et al., 2017).

The bot evaluation results showed that some categories contained fewer bots than others. For example, categories such as *care providers* and *apps/books* had a comparatively high average bot score. A possible explanation is that these two categories tend to use automated tools for updating their profiles and posting their tweets. However, the quality of automatically posted information is not guaranteed, as is discussed further in Chapter 3. In general, out of 8,400 users, 13.85% were likely-bot profiles. The average bot score of all profiles was 1.32. This could indicate a general tendency towards human profiles, however, even a small number of bots in the network can amplify information on a wider scale compared to humans.

The second objective of this study (RQ3) is to analyse which principal features are important for bot evaluation. Three types, namely profile, content, and domain features, were selected based on the literature and analysed using SMLR analysis, linguistic feature analysis and URL analysis. In terms of profile features, SMLR analysis shows that four features out of seven, X1 = verified, X3 = friends\_count, X5 = favorites\_count, and X7 = geo\_enabled, are relevant in explaining the bot score of Twitter users participating in disseminating dementia information in the full datasets. However, only two, X5 = favorites\_count and X7 = geo\_enabled, are relevant in the three datasets, all profiles, human profiles with score 0-2.5 and bot profiles with score 3-5, with negative coefficients indicating reverse direction between these two features and bot scores. In the case of X7 = geo\_enabled, the statistical test shows ( $\beta = -0.187$ ,  $p < 0.01$ ). As described earlier, geo-enabled means a user can include geographic information in their

## Chapter 5. Study 1: Exploratory Study

account or tweets. Users can enable or disable this feature. If it is enabled, the user can attach their physical location to their details. Enabling geographical information in user details indicates that the profile is likely to be human, whereas disabling this information indicates that the profile is likely to be a bot. In the case of  $X_5 = \text{favorites\_count}$ , the statistical test shows ( $\beta = -0.109$ ,  $p < 0.01$ ). `Favorite\_count` is the number of tweets that the user liked during the active period of the account. This suggests that humans like/favourite tweets more than bots.

On the content level, the Posit toolset was used to generate a quantitative analysis of the linguistic features of tweets, including POS tagging from datasets consisting of the same number of tweets posted by likely-bot profiles and likely-human profiles. Linguistic analysis (syntactic and lexical) of tweets showed that, generally, linguistic features of bot-like profiles are fewer in number compared to human profiles, apart from a few features including number of characters, number of sentences, AWL and one POS form: noun, which showed the opposite. For example, likely-bot profiles tend to use more nouns in their tweets compared to likely-human tweets. In addition, the results showed that there is diversity in using most POS forms and types in human tweets compared to bot tweets. URL frequency analysis shows that both human and bot-like profiles tend to use URLs extensively in tweets, however, bots use them more. The URL domain types analysis does not indicate a clear difference between both types of profiles. Limitations of this study are discussed together with the limitations of Study 2 in Chapter 6.

### Chapter Summary

This study aimed to answer questions regarding the nature and the extent of the relationship between the principal features of bots on SM in order to provide a solution to utilise these principal features to identify low-quality information. The study investigates dementia information sources and evaluates the bot presence in the dementia community on Twitter. It is the first to empirically examine bot and human characteristics in the context of spreading dementia information. The results indicate a bot



involvement in dementia information dissemination on Twitter. To distinguish between bot and human users, an analysis of principal features of profile and tweets posted by both types of users provided insight on distinctive features. Four features of user profiles, verified, friends\_count, geo\_data and favorites\_count, are revealed to be the most important features in evaluating bots. Linguistic analysis shows that some word-level and sentence-level features (e.g., number of characters), POS (e.g., nouns) and POS types (e.g., noun types) have different counts in bot profile content than in content of human profiles. The URL analysis of tweets from profiles with high bot scores shows the number of URLs in bot tweets is higher than that in human tweets. The next logical question is then, does the bot spread low-quality information? And if so, what is the contribution of bots in spreading that information? How can the features discovered in Study 1 be used to determine the quality of information? What other linguistic (syntactic and lexical), domain, and psycholinguistics features can be derived to provide better accuracy in discovering the low-quality information? These questions are the focus of the next study presented in the following chapter.

## Chapter 6

# Study 2: Machine Learning Experiments

### 6.1 Introduction

This chapter addresses the research questions RQ4-RQ5. These research questions are concerned with content quality as one of the credibility components. As defined in Chapter 2, content quality concerns two aspects: text quality analysis and the presence of spam e.g., malicious bots as an indicator of bad quality (Ginsca et al., 2015). Text quality analysis includes, for example, quantifying syntactic and lexical features (Ginsca et al., 2015). The two research questions to be answered in this chapter are:

- **RQ4:** To what extent, if at all, do the most active bots contribute to spreading low-quality dementia-related information on Twitter? Which bot types have the greatest involvement in the spread of low-quality dementia-related information on Twitter?
- **RQ5:** What are the most effective features to improve the automated assessment of dementia information quality?

The results of Study 1 provide a general picture of the research problem. Importantly, Study 1 proves the presence of bots in dementia information dissemination on Twitter. The study also revealed the principal features of likely-bot and likely-human profiles.

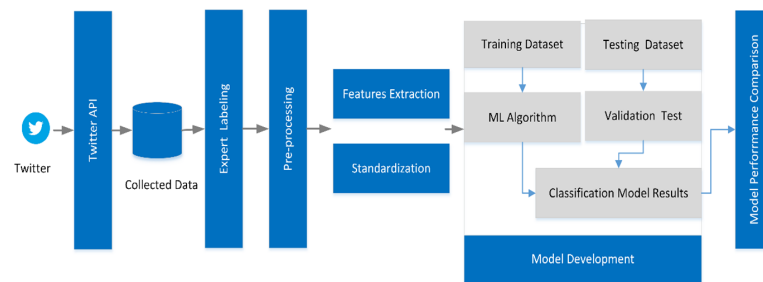


Figure 6.1: Study 2 method overview

Therefore, Study 2 explores the notion that bots are more likely to spread false information related to dementia than humans are, and studies what bot types contribute to the information spread. The study also examines the potential of using the features identified in the first study for information quality analysis using ML algorithms.

## 6.2 Methods

Figure 6.1 illustrates the methods for this study used for the dataset collection and labelling, and the steps taken to extract features from the dataset, which are explained in the following sections.

### 6.2.1 Dementia Myths Selection

One aspect of the thesis is to use automatic procedures to distinguish between two classes of dementia-related tweets, namely fact (true) and myth (false; see definition in Section 3.1.1). ‘Fact’ indicates high-quality information, and ‘myth’ indicates low-quality information. In order to develop a ground truth for information quality assessment, four dementia myths, discussed by trusted sources such as the Alzheimer’s Association<sup>1</sup> and the Alzheimer Society of Canada<sup>2</sup>, as well as in the literature, are used in this study Table 6.1. A large part of the public is unwilling to accept dementia information supported by empirical studies, due to stigma and misconceptions associated with dementia, and a high percentage believe in dementia myths (Nagel, Loetscher,

<sup>1</sup><https://www.alz.org/alzheimers-dementia/what-is-alzheimers/myths>

<sup>2</sup><https://alzheimer.ca/en/about-dementia/stigma-against-dementia/myths-realities-dementia>

Smith, & Keage, 2021). Myths are widely held false beliefs, frequently passed down through generations by retelling (Viehbeck, Petticrew, & Cummins, 2015). Belief in myths or lack of understanding of dementia can result in delays in seeking help, as well as resistance to accepting dementia information and treatment (Nagel et al., 2021). Therefore, many health studies have been conducted to evaluate the public's knowledge and understanding of dementia and its risk factors, in an attempt to enable the public to make better decisions related to health and medical matters. These studies have some concerning findings and show that there are issues that need to be resolved. Four dementia myths are frequently discussed in literature regarding both the increasing risk of dementia and curing or preventing dementia. First, there is a persistent myth that aluminium is a likely risk factor contributing to dementia (Low & Anstey, 2007; Nagel et al., 2021). A second myth is that flu vaccination can cause dementia, even though some research studies have revealed that flu shots and other vaccines actually reduce the risk of dementia and lead to better health overall (J.-C. Liu et al., 2016; Veronese et al., 2022). There are myths not only regarding factors increasing the risk of dementia, but also regarding the treatment and halting of the progress of dementia, despite strong evidence that, so far, there is no cure. Thirdly, the public supports empirically unfounded or poorly supported dementia risk reduction strategies like the use of vitamins (Cations, Radisic, Crotty, & Laver, 2018). Up to 75% of participants in different studies (Ayalon, 2013; Roberts, McLaughlin, & Connell, 2014; Cations et al., 2018) believed vitamins could decrease the risk of dementia or prevent it. While certain vitamins may help to manage and reduce some dementia symptoms, ultimately they will not be able to stop or reverse the cognitive decline caused by dementia<sup>3</sup>. The fourth myth is that marijuana or cannabis can prevent or treat dementia, for which no conclusive evidence has been found, although some research has shown that a few behavioural symptoms, such as agitation and aggression, can be managed through the use of cannabinoids (Charernboon, Lerthattasilp, & Supasitthumrong, 2021). These four myths were selected to create the ground truth dataset for the current study.

---

<sup>3</sup><https://alzheimer.ca/en/about-dementia/stigma-against-dementia/myths-realities-dementia>

### 6.2.2 Data Collection

Publicly available tweets were collected through Tweepy library in Python 3.7.5. The collected tweets were only in English and posted over a period of three years, January 2018 to December 2020. For this data collection, different search query strategies relating to the four myths, identified by trusted sources like the Alzheimer’s Association and the Alzheimer Society of Canada, were applied ( Table 6.1). These procedures were similar to those used in (Ghenai & Mejova, 2017). First, a set of terms was devised to best describe each myth, for example, “marijuana”, “Alzheimer” or “dementia”, and so forth. “Alzheimer’s disease” and “dementia” are both used because they are frequently used interchangeably although dementia is an umbrella term which describes a diverse range of brain diseases, whereas Alzheimer’s disease is the most common form of dementia (Jelavić, Klemar, & Sušić, 2018). Then, tweets related to myths terms were retrieved using different keywords linked with the AND operator (e.g., “dementia AND flu shots”, “Alzheimer AND flu shots”, “dementia AND marijuana”, “Alzheimer AND marijuana”, “cannabis AND dementia”, “cannabis AND Alzheimer”). Unwanted and irrelevant terms (e.g., dog, doggie) were observed in the initial collected data, so they were eliminated from the query using the NOT operator.

Table 6.1: Dementia myths.

	Myths	Reported by (alz.org)	Reported by (alzheimer.ca)
1	Drinking out of aluminum cans or cooking in aluminum pots and pans can lead to Alzheimer’s disease.	✓	✓
2	Flu shots increase risk of Alzheimer’s disease.	✓	
3	Certain vitamins, supplements and memory boosters can prevent and treat dementia.		✓
4	Marijuana can prevent and treat dementia. Currently there is no evidence to show that cannabis or cannabis oil (cannabidiol oil) can stop, reverse or prevent dementia.		✓

The tweets related to one myth were then combined into one data subset, resulting in four data subsets. Duplicated tweets from the same author with different dates and

times were removed to avoid bias in the analysis. Each tweet author’s information, along with bot scores, was also collected. The information includes the seven profile features, verified, followers\_count, friends\_count, listed\_count, favourites\_count, statuses\_count, and geo\_enabled. A description of all profile features is provided in Chapter 5. Bot scores for all tweet authors were obtained by using Botometer v4. The final dataset contained 2,920 tweets by 1,400 unique authors.

### 6.2.3 Data Labelling Guidelines

To verify the relevancy of tweets, and whether they could be categorised under fact (true) or myth (false), an annotator was requested to label each tweet’s relevance to the corresponding myth. The annotator was recruited via an invitation sent to different dementia and neuroscience research groups in Scotland. The elected annotator is MSc neuroscience researcher who is interested in degenerative nerve diseases including dementia.

To ensure the high quality of the manual labelling process, clear instructions were provided to the annotator for labelling the tweets. The tweets were labelled 0, 1, or none, where 0 indicates false, 1 indicates true, and none indicates irrelevant. The instructions provided were as follows:

1. Read the text of each tweet, including the hashtags.
2. Decide whether it is a relevant or irrelevant tweet. An irrelevant tweet is either not related to dementia, contains questions like “Could you link any studies that show aluminium causes Alzheimer?” or it is not possible to open a link to get enough information.
3. If deemed an irrelevant tweet, label it as none.
  - 3.1 If it is a relevant tweet, check if it is true or false:
    - 3.1.1 Click on the link to visit the page and check
      - 3.1.1.1 If the link page heading is clear, skim the article content to ensure that it matches the title.

3.1.1.2 If the link page heading does not reveal anything, read the content carefully.

The tweet should be labelled as 0 if it conveys false information, otherwise it should be labelled as 1.

In addition, tweets were divided and shared with the researcher in five separate batches to ensure high quality revision. Each batch contained one specific myth. After labelling each batch, the researcher and annotator reviewed and discussed the labels before starting with the new batch. To encourage a thorough examination of the tweets and high-quality results, the annotator was paid £1,211, using the SDRC grant.

#### **6.2.4 Bot Evaluation**

In the first study Chapter 5, bot evaluation was conducted on general tweets related to dementia. The results showed that bots make up 13.85% of overall users (sample size: 8,400). Firstly, this study aimed to determine to what extent, if at all, bots contribute to spreading low-quality dementia information (dementia myths) on Twitter. To address this question, basic descriptive statistical analyses were carried out on labelled data. The percentage of bot-likely authors was determined by the overall score and CAP. As described in Section 4.3.3, the Python Botometer API v4 queries the Twitter API to extract thousands of account features, classified into six categories, and feeds these features to an ensemble of ML classifiers, which generates a score for each category, an overall bot score and CAP. Overall scores can be in the range of 0-5 or 0-1 and are used to determine whether the given account is a bot. The value of 0 indicates more human-like behaviour and 5 indicates high bot-like behaviour.

In its May 2018 update, Botometer introduced the CAP as a more principled way to decide if an account is a bot or not. As introduced earlier in Section 4.3.3, CAP provides a probabilistic interpretation of a bot score and estimates the probability that an account with a certain score or above is indeed automated. The CAP scores are Bayesian posteriors that balance false positives and false negatives by reflecting both the Botometer classifier's findings and prior information about the presence of bots on Twitter. For example, if the CAP score is 0.90 for users with an overall score of 4.8,

there is a 90% chance that this profile is a bot and possibly operated by automated accounts, and a 10% chance that it is wrongly classified as a bot. In other words, 10% of accounts with an overall score of 4.8 or above are actually human <sup>4</sup> (K.-C. Yang, Ferrara, & Menczer, 2022). Prior studies have set different arbitrary CAP thresholds to determine whether an account is automated. For example, Gruzd and Mai (2020) and Y. Zhang et al. (2019) considered accounts with a CAP score higher than 0.25 as potentially a bot and lower than that as human. The PEW Research Center <sup>5</sup> and Alsmadi and O'Brien (2020) used a threshold of 0.37. A more conservative threshold was set by Keller and Klinger (2019) and Keller (2020) at 0.76. Keller (2020) looked at the intersection between the overall bot score and CAP. For example, accounts were considered as bots, due to their high Botometer scores with a probability greater than 76%.

On the other hand, some scholars have relied on the overall score on a 0-5 scale. Here, 0 indicates more human-like behaviour and 5 indicates bot-like behaviour. Following Keller and Klinger (2019), Keller (2020), and Gallwitz and Kreil (2021), a very high CAP threshold (0.76) was set for this study and in order to confirm the robustness of the findings, both an overall score equal to or greater than 3 with a CAP equal or greater than 0.76 was set as thresholds to identify bots. Secondly, to identify the most influential bot types contributing to the spread of false information on Twitter in the sub dataset and overall data, bot score types computed by Botometer API were used. Pearson correlation coefficients analysis was performed to find the association between the overall scores of bot-likely profiles who write false tweets and different types of bot subscores by using the SciPy package in Python.

### 6.2.5 Feature Extraction, Standardisation, and Selection

The first study Chapter 5 of this research revealed some principal bot features, including three feature types: profile features, linguistic features (syntactic and lexical features) and domain feature (URLs count). Therefore, these features are selected as candidate

---

<sup>4</sup><https://botometer.osome.iu.edu/faq>

<sup>5</sup>[https://www.pewresearch.org/internet/2018/04/09/bots-in-the-tweetsphere/pi.2018-04-09\\_twitter-bots\\_m-05](https://www.pewresearch.org/internet/2018/04/09/bots-in-the-tweetsphere/pi.2018-04-09_twitter-bots_m-05)



features to assess their suitability for automatically analysing the quality of individual tweets using ML in this study. As indicated earlier, the assumption was reached based on literature showing that fake news likely originates and is propagated by non-human accounts (K. Shu et al., 2017), such as bots, which have the ability to spread a large amount of low-quality content that negatively impacts user experience (Resende et al., 2020).

Below is a list of the features that were studied when performing the ML experiments in this study.

- Profile level features: As shown in the first study, four profile features have a relationship with bot score, namely X1 = verified, X3 = friends\_count, X5 = favorites\_count, and X7 = geo\_enabled.
- Content level features: Posit’s linguistic feature analysis in the first study showed the variations between tweets produced by bot and human profiles. Therefore, these features were utilised in this study as well. However, to improve the accuracy of classifiers and because linguistic processing using automated classification should be constructed on numerous layers of word or lexical analysis (Conroy, Rubin, & Chen, 2015), the linguistic features expanded here to include a greater number of linguistic features and psycholinguistic features generated from LIWC (Pennebaker et al., 2015).
- Domain-specific features (also called Twitter-specific features or hypertextual features (X. Zhang & Zhu, 2021)) are the elements in a post linking the text to another entity. The number of URLs in the analysis of the first study revealed a slight difference between tweets produced by bot and human profiles. The domain type is not considered since the first study showed that similar types of domains in the URLs are used in both bot tweets and human tweets, and because knowledge of a domain type does not guarantee the quality of information in the URLs. Similar to linguistic features, domain-specific features were also expanded in this study to include two other primary hypertextual features in tweets which are the number of mentions and the number of hashtags besides the URLs.

- A bot-flag feature, which is a score for the bot likelihood of a tweet author. It was used in this study to test whether adding this feature in combination with profile and/or tweet features would improve the automated classification of individual tweets.

First, all features were extracted. Profile features were extracted directly from the user profile who posted the tweet. For domain features, the number of URLs, mentions, and hashtags for each tweet were counted using Python scripts (3.8), using the RegEx Package `re`, a package that offers functions that enable searching a string for a match <sup>6</sup>. For content features, the text was pre-processed prior to extracting the features. First, all user mentions, URLs and `#` symbols were removed using the RegEx Package `re`. All contractions within the text were resolved using the contraction library <sup>7</sup>. For example, “you’re” was converted to “you are”. Next, all duplicate tweets were eliminated, resulting in a total of 1,771 tweets. After pre-processing, tweets were converted into the feature vectors using two text analysis tools, Posit and LIWC (both tools are introduced in Chapter 4). Posit focuses on quantitative linguistic features (syntactic and lexical) present in the tweet, 27 features in all, whereas LIWC includes psycholinguistic features (indicators of people’s social and psychological lives), divided into different categories like psychological, emotional, cognitive, and others, as well as linguistic features (lexical and syntactic features), 90 features in all. A complete list of the standard LIWC2015 and Posit output is included in Appendix A and D. Finally, the bot feature was derived from Botometer v4. In total, 124 features were derived from both tweets and users. These were categorised into three main categories: profile features, linguistic features, and bot feature.

For variables that are different in scale, standardisation is usually recommended. This was the case in this dataset, where profile features such as follower numbers and `status_count` ranged from one to several 1,000, whereas the numerical values generated for linguistic features from Posit and LIWC ranged from 0 to several 100. Thus, features were standardised with scale ranges from zero to one. `StandardScaler` functions in the

---

<sup>6</sup><https://docs.python.org/3/library/re.html>

<sup>7</sup><https://pypi.org/project/pycontractions/>

Scikit-learn package <sup>8</sup> in Python 3.7 were used for this standardisation.

Feature selection is important in ML classification, especially when there are huge feature sets. Large feature sets confuse the model and increase machine resource allocation without notably improving its effectiveness (Rani et al., 2021). Features selection is used mostly to reduce the original size of the feature set by omitting irrelevant and redundant information for text classification without influencing the efficiency of the model (de Moraes & Gradvohl, 2021). Thus, it can diminish both the model execution time and overfitting, and so it can improve the accuracy of the model. Importantly, feature selection methods yield easy interpretability. Therefore, different feature selection techniques are employed in this study, since the goal is to analysis the quality of text and understand the most important features. Three different common feature selection techniques categories, namely filter, wrapper and embedding, were used. These are described in Section 4.2.1.2. ANOVA was applied as filter-based, RFE was applied as wrapper-based, and RF was used as embedded methods.

## 6.3 Analysis and Results

This section details the results of the statistical and ML analyses performed on the dataset to answer the two research questions RQ4 and RQ5.

### 6.3.1 Descriptive Analysis

A descriptive analysis on the total of 2,920 annotated tweets was conducted to find out if active bots contribute to spreading low-quality dementia-related information (RQ4). The results provided in Table 6.2 show the number of overall tweets, subdivided into true, false and irrelevant tweets, in each collected myth sub-dataset. The table also provides the number and percentage of bot-likely authors when the CAP is  $\geq .76$  and overall score  $\geq 3$ .

Note that 118 (21%) tweets in the Myth 2 dataset were labelled as irrelevant. This number is high compared to the other sub-datasets. Myth 2 was about linking flu

---

<sup>8</sup><https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>

Table 6.2: Descriptive statistical analysis of myths and bots.

Dataset	All datasets				False Tweets
	N	False	True	Irrelevant data	Bot %
Dataset 1	901	746 (85%)	127 (15%)	28	315 (42%)
Dataset 2	565	357 (80%)	90 (20%)	118	59 (17%)
Dataset 3	615	359 (59%)	254 (41%)	2	133 (37%)
Dataset 4	839	118 (15%)	652 (85%)	69	37 (31%)
<b>Total</b>	<b>2920</b>	<b>1580(58%)</b>	<b>1123 (42%)</b>	<b>217</b>	<b>592 (37%)</b>

\* Dataset 1= Myth 1 (aluminium), Dataset 2= Myth 2(flu vaccination), Dataset 3= Myth 3 (vitamins), Dataset 4= Myth 4 (cannabis/marijuana)

vaccination with dementia. One reason for the high number of irrelevant tweets could be that flu shot related tweets appeared in COVID-19 discussions; the pandemic was emerging when that dataset was collected. For example, one of the tweets was: ‘I have to get the flu jab to visit my grandmother with dementia in the hospital’. Although the flu shot and dementia are mentioned in the same tweet, they are not otherwise related here. Another example of an irrelevant tweet is: ‘Go ahead and point out all the minor things that the president has wrong with his health but remember that Joe Biden has dementia and that is incurable and he will be going downhill very fast. Also, if you get flu shots, you might want to check the ingredients’. These tweets are labelled as irrelevant. In the Myth 4 dataset, about cannabis, more than half of the tweets mentioned that cannabis helped dementia patients, not that it could prevent dementia or slow down progression. This is supported by evidence, so these cases were labelled as true. However, tweets about certain marijuana or cannabidiol oils (cannabidiol is a prime component of cannabis) having the ability to act as a neuroprotectant and prevent the onset of dementia were labelled as false, as there is not enough evidence supporting this claim.

Overall, the descriptive analysis results show that bots make up 592 (37%) of the 1,580 false information authors. This suggests more than a third of false information in the dementia community is spread by bots. On the sub-dataset level, the percentage of bots for Myths 1, 3 and 4 (about aluminium, vitamins, and cannabis/marijuana, respectively) was high, at 42%, 37%, and 31%, respectively. The dataset for Myth 2 (flu vaccine) had lower bot participation, at 16% compared to the aluminium, vitamin

and cannabis/marijuana datasets, indicating that most of the false information spread about the flu vaccination was by presumed human accounts.

Table 6.3: Results of the Pearson correlation coefficient between bot score and bot-type score.

Bot type/Dataset	All		Dataset 1		Dataset 2		Dataset 3		Dataset 4	
	r	p	r	p	r	p	r	p	r	p
<b>Astroturf</b>	-0.15	.00	-0.16	.01	0.34	0.02	-0.26	.00	-0.19	0.33
<b>Fake follower</b>	0.34	.00	0.31	.00	-0.01	0.95	0.40	.00	0.39	0.04
<b>Financial</b>	-0.02	.62	-0.18	.00	0.21	0.16	0.15	.08	-0.14	0.48
<b>Self-declared</b>	0.48	.00	0.53	.00	-0.17	0.25	<b>0.63</b>	.00	0.55	0.00
<b>Spammer</b>	0.37	.00	0.32	.00	-0.03	0.85	0.51	.00	0.45	0.01
<b>Other</b>	0.58	.00	<b>0.66</b>	.00	-0.09	0.56	<b>0.79</b>	.00	0.44	0.02

\* Dataset 1= Myth 1 (aluminium), Dataset 2 = Myth 2 (flu vaccination), Dataset 3 = Myth 3 (vitamins), Dataset 4= Myth 4 (cannabis/marijuana)

As described in Section 4.3.3, Botometer v4 reports sub-scores for six bot types (astroturf, fake follower, self-declared, financial, spammer, and other) which come from the specialised bot classifiers and estimate how similar an account is to different types of bots. These sub-scores were used to determine the most influential bot type contributing to the overall dataset and to each myth sub-dataset using the Pearson correlation coefficient. Pearson correlation coefficient results are provided in Table 6.3. The variable P shows the statistical significance of the data. The correlation is considered as significant if  $p < .05$ . The correlation coefficient (r) is a statistical measure of the strength of the relationship. Pearson correlation coefficient (r) range between -1 and 1. It can be described as follows: .00 to .19 (00 to -.19) is very weak, .20 to .39 (-.20 to -.39) is weak, .40 to .59 (-.40 to -.59) is moderate, .60 to .79 (-.60 to -.79) is strong, and .80 to 1.0 (-.80 to 1.0) is very strong (Evans, 1996).

In the overall dataset, there is a statically significant positive moderate correlation between the overall bot score and the others and self-declared bot type ( $p < .05$ ) ( $r = .58$  and  $r = .48$ , respectively). The others type does not clarify who the responsible information spreader is, because the training dataset used for this bot type in Botometer was composed of multiple unexplained types. As described by Botometer (Section 4.3.3), the others class consists of bots from manual annotation and user feedback. Thus, others does not refer to a particular type and further exploration is recommended

to determine the exact composition of this bot type. The self-declared type dataset is based on botwiki.org, which preserves examples of interesting and creative online bots that identify themselves as bots <sup>9</sup>.

Spammers and fake follower types are also present with a statically significant positive weak correlation ( $p < .05$ ,  $r = 0.37$ ,  $r = 0.34$ ). The spammer bot type is trained on spam bots from “pron-bots” <sup>10</sup> and a subset of “cresci-17” dataset (Cresci, Di Pietro, Petrocchi, Spognardi, & Tesconi, 2017). “Pron-bots” are Twitter bots advertising scam sites and the cresci-17 dataset contains several types of spam datasets, such as spammers of scam URLs, spammers of products on sale at Amazon.com, and automated accounts spamming job offers <sup>11</sup>. The fake follower type refers to bot accounts purchased by companies to increase their followers.

A detailed analysis on the individual sub-datasets to determine the bot types responsible for spreading a particular type of myth was also performed. For Myth 2 (flu vaccination), there was no correlation between this dataset with any bot type except with astroturf type. There was statically significant positive weak correlation ( $p < .05$ ,  $r = .34$ ). Even though the correlation is weak, this could be because, as shown in Table 6.3, the data itself contained a very low percentage of bots (%17). Figure 6.2 shows tweets from a single account classified as astroturf bot type.

The overall scores of profiles in other myths (1, 3 and 4) were significantly correlated with the others bot type ( $p < .05$ ) with varying strengths of correlation. There was a positive strong correlation ( $r = .66$  and  $r = .79$ ) with Myth 1 and Myth 3 (aluminium and vitamins) and a positive moderate correlation ( $r = .44$ ) with Myth 4 (cannabis/marijuana). Similarly, the overall scores of profiles in the same data subsets were significantly correlated with the self-declared bot type ( $p < .05$ ) with varying strengths of correlation. There was a positive strong correlation with Myth 3 ( $r = .63$ ) and a positive moderate correlation ( $r = .53$ ,  $r = .55$ ) with Myth 1 and Myth 4. This suggests that the self-declared and the others type are responsible for spreading the majority of dementia related myths (Myth 1, 3, and 4).

---

<sup>9</sup><https://botwiki.org>

<sup>10</sup><https://github.com/r0zetta/pronbot2>

<sup>11</sup><http://mib.projects.iit.cnr.it/dataset.html>

## Chapter 6. Study 2: Machine Learning Experiments

All the fools rushing out to get their aluminum laden, dementia causing flu shot, enjoy Corona time as well.  
<https://t.co/SbD1SL4JjO>

It's no coincidence that dementia cases have been spiking during the same time that children and adults alike are being over-vaccinated (flu shot, anyone?) and the over-prescription of brain-altering drugs like antidepressants is prevalent.  
<https://t.co/Owl2ac20Dy>

My mother died earlier this year she didn't have COVID but died of pneumonia in April. There needs to be research into flu shots and the increase of dementia which is at an all-time high.  
<https://t.co/KyhWo264TS>

@ChuckCallesto Nope. Don't get flu shot either. My mom had dementia and many of the chemicals in last years flu shot "can contribute to dementia". I'll chance the flu and actively build up my immune system daily.

@Notyour28981739 As witnessed by the marked increase of Alzheimer's Disease and Dementia in our elderly via the aluminum, (and mercury from the 10 shot vial), in flu shots.  
@minihorses89 The flu one makes you more susceptible to the beer bug. What might the beer shot do to immunity for the next one?

Flu shot makes people ill. Modern vaccines hurt a lot of people. Ask India and Kenya. Or just cradle death. Autism, ADHD, dementia...

Figure 6.2: Examples of Myth 2 from astroturf bot.

Another prominent bot type is the spammer. There was a statistically significant positive moderate correlation between the spammer bot type and Myth 3 and 4 (vitamins and cannabis/marijuana) ( $p < .05$ ,  $r = .51$ ,  $r = .45$ , respectively). There was a statistically significant but weak correlation with Myth 1 (aluminium) ( $r = .32$ ). A manual inspection of spammer profiles in Myths 3 and 4 showed that most of the profiles were promoting products, either vitamins or cannabis. As an example, screenshots of two spammers profiles are shown in Figure 6.3. The fake follower type also showed a significant positive moderate correlation ( $p < .05$ ,  $r = .40$ ) with Myth 3 (vitamins) and a significant weak correlation with Myth 1 and 4 (aluminium and cannabis/marijuana) ( $r = .31$ ,  $r = .39$ ).

The financial and astroturf type has less presence in the dataset. The financial type is trained on bots mentioning stocks traded on the most popular US markets (Cresci, Lillo, Regoli, Tardelli, & Tesconi, 2019) and the astroturf type is trained on hyper-active political bots. Neither of these are likely to be very present in the dataset. The coefficient correlation of the financial type and the overall score is very weak in Myth 1 ( $r = -.18$ ). Likewise, the astroturf type showed a very weak negative correlation within the overall data, Myth 1, Myth 2, and Myth 3 ( $r = -.15$ ,  $-.16$ , and  $-.26$ , respectively)



Figure 6.3: Example of spammer bot profiles

and a weak positive correlation with Myth 2 ( $r = .34$ ).

In summary, a moderate and strong correlation was found between the overall bot score and some bot types in different myth subsets. This may help to identify the bot types in the diffusion of myths. A strong correlation between overall score and others was shown in Myths 1 and 3 (aluminium and vitamins) as well as between overall score and self-declared types in Myth 3 (vitamins). There was a moderate correlation with spammer in Myths 3 and 4 (vitamins and cannabis/marijuana) and fake follower types in Myth 3 (vitamins).

### 6.3.2 Feature Importance Selection

The individual feature's importance was determined using three feature selection methods (ANOVA, REF, RF) on the annotated dataset after removing the duplication. The final number of false tweets was 1,009 and the number of true tweets was 726. The different feature selection techniques and the justification for using these three methods are discussed in Section 4.2.1.2. The top 30 features of each selection method are shown in Appendix M, N, and L. A feature is considered important if it is listed in the

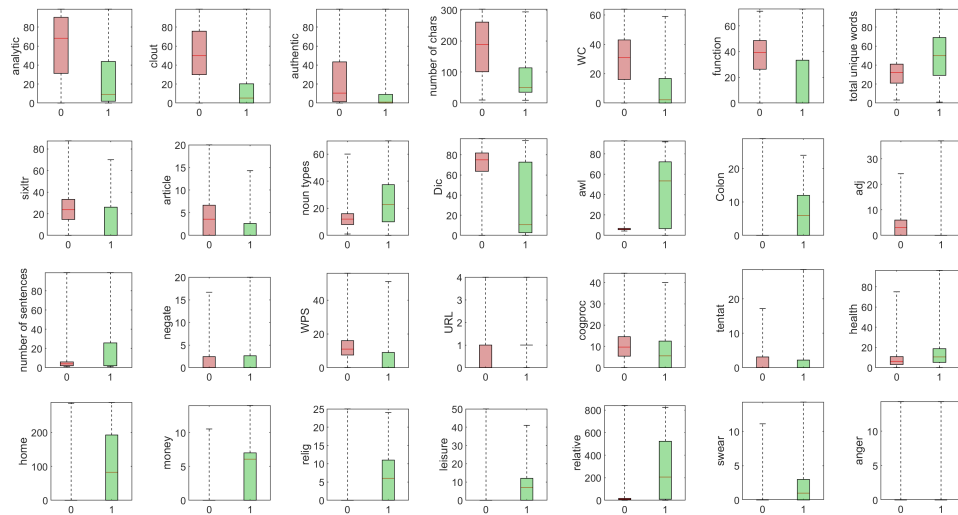


output of two or three feature selection techniques. Ultimately, 28 important features turned out to contribute most to the classification (see Table 6.4), while six of them (no of characters, no of sentences, AWL, total unique words, noun types, URLs count) were same as those resulted from the first study of content feature analysis of bot and human.

Boxplots for each feature in both classes 0 and 1 are shown in Figure 6.4. Summary statistics (count, mean - average, standard deviation, the minimum value, the maximum value 25% percentile, 50% percentile, 75% percentile) of these boxplots are provided in Tables 6.5, and 6.6. The percentile refers to how many of the values are less than the given percentile. Figure 6.4 below uses boxplots to depict the distributions of the 28 optimal features to distinguish between true information and myths. As a visualisation tool, a boxplot depicts the change in numerical data, allowing for visual comparison. Each box plot depicts the distribution of false and true instances (posts) along a feature's value scale. For instance, in the case of number of characters, the scale goes from 100 to 252. The median, or mid-point of the data, is shown by the dark line within each box. Whiskers are the lines drawn vertically from the top and bottom of a box. Outlier bounds are shown by the borders of these lines. Boxplots can show various types of statistical data, such as medians, ranges, and outliers. For the purpose of interpretation, the box lengths determine the data distribution between samples. The taller the box, the more dispersed the feature values. In contrast, a shorter or "compressed" box indicates the feature values are limited to a specific range. There is a difference between the two groups if two boxes do not overlap, such as when box A is totally above or below box B, for example with the clout feature in Figure 6.4. If boxes do overlap, the middle line is checked. If the median line of box A lies totally outside of box B, there is likely to be a difference between the two groups. This is illustrated by most features shown in the figures (e.g., number of chars, WC, Dic, function, analytic, authentic, article, and noun\_types). If boxes overlap and the median line of box A lies within box B, the length of the box is assessed. Short boxes indicate that the data lie close to the centre values, whereas taller boxes means that the data is more variable and has a wider distribution.

Table 6.4: Selected features.

Feature Category	Features	Library	Description
Domain specific	URL count	Python	Number of full web links or shortened web links
Syntactic and lexical features	Average word length	Posit	Average word length
	Number_of_chars	Posit	Number of characters
	Number_of_sentences	Posit	Number of sentences
	Total_unique_words	Posit	Number of unique words in tweets
	Noun_types	Posit	Number of noun types
	WPS	LIWC	Number of words per sentence
	WC	LIWC	Word count
	Sixltr	LIWC	Percentage of words in the text that are longer than six letters
	Adjective(adj)	LIWC	Frequency of adjectives
	Function	LIWC	Refers to words used to connect and shape other types of words in the text. Includes pronouns, articles, prepositions, auxiliary verbs, negations, conjugations, quantifiers, and common adverbs.
	Article	LIWC	Articles
	Negate	LIWC	Negations
	Colon	LIWC	Punctuation
Dic	LIWC	Percentage of target words captured by the LIWC dictionary.	
Psycholinguistic (word level)	Anger	LIWC	Words indicating the anger emotion (i.e., hate, annoyed)
	Cogproc	LIWC	Words indicating cognitive processes (i.e., cause, know, ought)
	Tentat	LIWC	Words indicating cognitive processes, specifically tentative words (i.e., maybe, perhaps)
	Health	LIWC	Words indicating biological processes (i.e., clinic, pill)
	Leisure	LIWC	Words indicating personal concerns regarding leisure (i.e. cook, chat, movie)
	Money	LIWC	Words indicating personal concerns regarding money (i.e., audit, cash, owe)
	Relig	LIWC	Words indicating personal concerns regarding religion (i.e., altar, church)
	Home	LIWC	Words indicating personal concerns regarding home (i.e., kitchen, landlord)
	Swear	LIWC	A subset of the informal category, indicating swear words (i.e., damn)
Relative	LIWC	Words that cannot be found in any other subcategories. (i.e., area, bend, exit)	
Psycholinguistic (summary)	Analytic	LIWC	A high score indicates formal, logical, and hierarchical thinking, while a low number indicates informal, personal, current/in the moment and narrative thinking.
	Authentic	LIWC	A high score indicates that the author is speaking from a position of honesty and integrity and implies open dialogue; the lower the number, the more guarded and distant the dialogue.
	Clout	LIWC	A high score indicates that the author is speaking from a position of high expertise and is confident; low numbers indicate uncertainty.



\* myth = 0, true = 1

Figure 6.4: Boxplot for selected features

There was a notable variation in the feature value distribution between the two classes, as shown in Figure 6.4. For example, boxes of all three LIWC summary variable measures (analytic, authentic, and clout) suggest that there is a difference between the two classes for each of these three features. Surprisingly, the boxes of analytic, authentic, and clout features for the myths class are taller than the boxes for the true class, suggesting that these features have the greatest data variability in the myths class. This means myth posts appear to be more analytic, authentic, and confident. The three summary features scored between 0 and 100, calculated by LIWC using a proprietary algorithm derived from past research. All are important, because they characterise user abilities and social status (Pei-Chi & Ee-Peng, 2018).

There are varying degrees of the analytic measure in the myths class compared to the true one. High values of analytic features in myths suggest formality and logicity in writing tweets, whereas lower numbers in true tweets suggest informality and narrative language in writing (Pennebaker et al., 2015). Similarly, the myths class exhibits a higher level of clout compared to the true class. This means myth posts appear more confident than the true ones. The authenticity box for the myth class suggests more

personal and honest words compared to the true class. This shows that sources who post myth tweets exhibit higher analytic thinking, authenticity and clout than true post sources.

Figure 6.2 presents 14 different linguistic features (syntactic and lexical) which were generated by both LIWC and Posit, and one domain-specific feature (URL). Features such as the number of characters, word count (WC), function, article, and dic show clear comparative box plots (not overlapped, or the middle line of one box lies out the other box) for both classes (true and myths).

The function and dic boxes for the myth post are higher, indicating more function words and dic are used in myth tweets compared to true ones. Function words are style or non-content words in a tweet, such as prepositions. Dic feature depicts the percentage of words captured by the LIWC dictionary, which indicates the technical complexity of the writing. A high level in Dic values suggests more common, less difficult and non-technical words. Figure 6.2 shows that myth posts have a larger range and less variability in Dic feature values compared to true posts, suggesting less technical writing and less complexity in myth posts. A discrepancy in the distribution of Dic values in true posts compared to myth posts, suggests that myth posts use a limited number of common words.

Importantly, the range of unique words and noun types used by the myth posts is smaller than those used in true tweets. This suggests myth posts concentrate on a specific number of words and noun types and contain less variety in word usage compared to the true ones. Patterns of linguistic features such as number of characters, the number of unique words, and number of noun types found in the myths and true sets are similar and align with linguistic features found in the bots and human sets discussed in the previous chapter. This means that some linguistic features that are successful in identifying bots can be also used in identifying low-quality posts.

The average word length, colon, and adjective features show random distribution in one class and uniform distribution in the other one. The visualisation of some features does not provide any important information. For example, the distribution of the negation feature values for both classes is similar. However, these features could be

important for the classification when they are combined with other features.

For domain-specific features (URLs), there is a uniform distribution of URLs used within the myths class, while there is a random distribution (outliers) of the total number of URLs used in the true class. This indicates clear usage of URLs in myths class.

Figure 6.2 also presents boxes of features for the psycholinguistic (word-based features) categories. These word feature scores reflect the percentage of words in the analysed text that refer to various psychological constructs, personal concern and informal language markers as described in LIWC. Cogproc and tentative as cognition process features and health as a biological process features show a similar pattern in Figure 6.2. This suggests a similar number of fixed words from these two categories are used for both classes (true and myths).

Personal concerns related features, such as home, money, religion and leisure, and informal speech features, such as swear words and relativity features, show similar patterns. The box of true posts is clearly plotted, whereas the boxes are compressed for the myth posts, due to extreme outliers being accommodated. This means the values of the psycholinguistic features related to personal concerns and informal speech are uniformly distributed in the true tweets, but randomly in the false ones. This could indicate that words related to these categories are more likely to be used by true tweets. Lastly, anger words have a completely random spread in both types of tweets. This shows both types convey negative emotion in their tweets.

The main objective of this section is to visualise the important features selected by various selection techniques. The visualisation shows clear differences between the low-quality information (myths) and high-quality information (true) posts, based on LIWC's summary of psycholinguistics features (e.g., authenticity, analytics, clout), LIWC's psycholinguistics word level (e.g., cogproc), syntactic and lexical features (e.g., number of characters), and domain-specific features (URLs).

Table 6.5: Summary statistics for class 0

	<b>analytic</b>	<b>clout</b>	<b>authentic</b>	<b>number_of _chars</b>	<b>WC</b>	<b>function</b>	<b>total_unique _words</b>	<b>Sixltr</b>
count	1,009	1,009	1,009	1,009	1,009	1,009	1,009	1,009
mean	60.03	51.69	25.63	176.57	28.89	35.82	32.68	24.40
std	32.63	29.91	30.40	87.28	16.15	17.28	16.51	14.74
min	0.00	0.00	0.00	9.76	0.00	0.00	3.22	0.00
25%	31.30	29.92	1.40	101.00	16.00	26.32	21.00	14.71
50%	68.29	50.00	10.47	189.00	31.00	39.39	32.00	23.81
75%	90.36	75.74	43.37	261.00	43.00	48.65	41.00	33.33
max	99.00	99.00	99.00	303.00	64.00	71.43	99.00	87.50
	<b>article</b>	<b>noun_ types</b>	<b>Dic</b>	<b>awl</b>	<b>Colon</b>	<b>adj</b>	<b>number_of_ sentences</b>	<b>negate</b>
count	1,009	1,009	1,009	1,009	1,009	1,009	1,009	1,009
mean	4.00	14.31	67.17	13.43	1.61	3.60	7.93	1.53
std	3.91	10.85	24.02	21.34	4.02	3.69	16.97	2.59
min	0.00	1.00	0.00	4.25	0.00	0.00	1.00	0.00
25%	0.00	8.00	63.64	5.59	0.00	0.00	2.00	0.00
50%	3.57	12.00	75.00	6.11	0.00	3.03	4.00	0.00
75%	6.67	16.00	81.82	6.81	0.00	6.00	6.00	2.44
max	20.00	60.00	96.15	92.86	29.00	24.14	99.00	16.67
	<b>WPS</b>	<b>URL</b>	<b>cogproc</b>	<b>tentat</b>	<b>health</b>	<b>home</b>	<b>money</b>	<b>relig</b>
count	1,009	1,009	1,009	1,009	1,009	1,009	1,009	1,009
mean	12.71	0.50	10.40	1.87	8.06	20.86	1.01	1.24
std	9.39	0.64	7.12	2.83	7.77	64.50	2.24	3.64
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	7.50	0.00	5.41	0.00	3.12	0.00	0.00	0.00
50%	11.00	0.00	9.68	0.00	6.25	0.00	0.00	0.00
75%	16.00	1.00	14.55	3.12	10.91	0.00	0.00	0.00
max	56.00	4.00	44.44	17.24	75.00	286.00	10.53	25.00
	<b>leisure</b>	<b>relative</b>	<b>swear</b>	<b>anger</b>				
count	1,009	1,009	1,009	1,009				
mean	1.76	44.02	0.39	0.69				
std	4.41	133.16	1.18	1.77				
min	0.00	0.00	0.00	0.00				
25%	0.00	5.56	0.00	0.00				
50%	0.00	10.53	0.00	0.00				
75%	0.00	16.67	0.00	0.00				
max	50.00	841.00	11.11	14.29				

Table 6.6: Summary statistics for class 1

	<b>analytic</b>	<b>clout</b>	<b>authentic</b>	<b>number_of _chars</b>	<b>WC</b>	<b>function</b>	<b>total_unique _words</b>	<b>Sixltr</b>
count	726	726	726	726	726	726	726	726
mean	24.58	17.22	12.37	87.57	10.60	15.21	48.98	12.65
std	31.20	25.70	23.98	78.39	15.31	21.35	25.72	18.33
min	0.00	0.00	0.00	9.09	0.00	0.00	1.00	0.00
25%	1.92	0.00	0.00	34.91	0.00	0.00	29.23	0.00
50%	9.09	5.26	1.00	50.00	2.20	0.00	50.00	0.00
75%	43.78	20.24	9.09	113.50	16.75	33.33	69.03	26.09
max	99.00	99.00	99.00	294.00	59.00	73.08	99.00	70.00
	<b>article</b>	<b>noun_ types</b>	<b>Dic</b>	<b>awl</b>	<b>Colon</b>	<b>adj</b>	<b>number_of_ sentences</b>	<b>negate</b>
count	726	726	726	726	726	726	726	726
mean	1.65	24.23	31.33	43.32	6.88	1.29	31.09	1.78
std	2.93	15.73	34.86	30.97	6.13	2.96	36.71	3.35
min	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
25%	0.00	10.00	2.63	6.61	0.00	0.00	2.00	0.00
50%	0.00	22.65	10.67	53.48	6.00	0.00	25.77	0.00
75%	2.63	37.41	72.73	72.22	12.00	0.00	25.77	2.63
max	14.29	69.77	94.12	91.89	24.00	37.04	99.00	20.00
	<b>WPS</b>	<b>URL</b>	<b>cogproc</b>	<b>tentat</b>	<b>health</b>	<b>home</b>	<b>money</b>	<b>relig</b>
count	726	726	726	726	726	726	726	726
mean	5.29	0.95	7.50	1.81	13.24	100.55	4.44	6.37
std	8.01	0.64	7.80	3.68	11.16	96.62	3.38	5.93
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	1.00	0.00	0.00	5.26	0.00	0.00	0.00
50%	0.00	1.00	5.56	0.00	10.62	82.50	6.07	6.00
75%	9.09	1.00	12.50	2.21	18.75	192.50	7.00	11.00
max	51.00	4.00	40.00	28.57	96.30	288.00	14.00	24.00
	<b>leisure</b>	<b>relative</b>	<b>swear</b>	<b>anger</b>				
count	726	726	726	726				
mean	7.82	275.35	1.61	0.38				
std	7.40	276.46	1.80	1.42				
min	0.00	0.00	0.00	0.00				
25%	0.00	9.09	0.00	0.00				
50%	7.00	206.00	1.00	0.00				
75%	12.00	523.50	3.00	0.00				
max	41.00	824.00	14.29	14.29				

### 6.3.3 Machine Learning Classification Analysis

To address RQ5, on the most effective features to improve the automated assessment of dementia information quality, five common supervised classification algorithms (RF, DT, SVM, LR, and KNN, with grid search using stratified 10-fold cross-validation, these are described in Section 4.2.1) are used to test the influence of different combinations of features on the classification task.

Initially, the impact of different sets of features on classification accuracy was evaluated in different experiments. Results of all algorithms obtained through separate features and combinations of different feature sets are given in Table 6.7. The table provides different performance metrics for all classifiers, including precision, recall, accuracy, MCC, and execution time in seconds). Moreover, the precision, recall and F1-score are also provided for each of the two classes for the given classification algorithm. All evaluation metrics are explained in Section 4.2.1.3.

Generally, accuracy and MCC of classification improved from 68% to 86%, 0.32 to 0.72, respectively, using different combinations of features. The recall value for the true class (1) increased significantly, from 38% to 77%, by using different combination of feature. All features are introduced in Section 6.2.5

In the first experiment, the classification accuracy ranged between 68% and 71% using only three domain-specific features (URL, mentions, hashtags). MCC values ranged from 0.32 to 0.40, indicating weak performance of all five classifiers. Support vector machine had the highest execution time for classification using the three features. The precision and recall values for the true class (1) were relatively low, indicating the classifier does not correctly classify the true tweets. However, on average, a better recall for the myth class (0) was found.

Similarly, using only four profile features (verified, friends\_count, favourites\_count and geo\_enabled), in experiment 2, the accuracy and MCC values are not improved. Notably, there is an increase in recall value for the true class (1) in experiment 2 by (RF, DT, and KNN) compared to experiment 1. The highest recall for class 1 in experiment 2 is 62% whereas the highest recall value for class 1 in experiment 1 is 46%. Overall, in both experiments, recall scores for the true class (1) produced by the classifiers were



low compared to those for the false class (0).

Using 27 linguistic features (syntactic and lexical) from the Posit toolset (experiment 3) with different classifiers resulted in a precision, recall and accuracy value of 79% with all classifiers except KNN, which attained 78% for recall and accuracy. The execution time of all classifiers was comparable, with the exception of SVM, which had the highest execution time of 65.89 seconds. MCC values for all classifiers were in the moderate range (0.54-0.57). The recall value for the true class(1) was between 63% to 70% with all algorithms, whereas recall values were between 63% and 70% for the false class (0). The results show a reduced difference in recall values of the two classes for different algorithms. This suggests that Posit's linguistic features are effective to classify both classes compared to the domain or profile features used in experiment 1 and experiment 2.

Using 90 LIWC features (syntactic, lexical and psycholinguistics) in experiment 4, the precision, recall and accuracy values are in the range of 79% to 82% for all classifiers. Matthews' correlation coefficient values for all classifiers are in the moderate range (0.56-0.63). The recall value for the true class (1) is greater than 70% and for the false class (0), it is greater than 80%. K-nearest neighbour achieved the highest accuracy value (82%), MCC (0.63) and recall (false class = 81%, true class = 83%). This demonstrates that LIWC's features produced the best result compared to the other features used in experiment 1, 2, and 3.

In experiment 5, a combination of profile, Posit, LIWC, and domain features, totalling 124 features, was applied to different algorithms, which produced the highest accuracy of 83%, an MCC value of (0.66), and recall values of 81% for the true class (1) and 85% for the false class (0), using the KNN classifier. This shows a slight improvement in the results compared to experiment 4 using LIWC features only.

Considering the large number of features in experiment 5, a set of features was selected from the overall features set Table 6.4 (techniques used for features selection is explained in Section 6.2.5). The 28 selected features were tested with the same ML algorithms in experiment 6. The results show the accuracy was improved for all algorithms except KNN, by 4% for SVM and LR, 2% for RF, and 1% for DT compared

with the accuracy achieved using all 128 features (experiment 5). Similarly, the MCC value of all classifiers improved (experiment 6) compared with the MCC value of all classifiers (experiment 5). The highest accuracy value achieved was 84% for RF, SVM and LR (experiment 6) and the highest MCC value was 0.68 by RF. Recall values in the true class are greater than 72 for experiment 6. Overall, the results suggest that RF achieves the best result (accuracy = 84%, MCC = 0.68) using fewer features (experiment 6), with KNN achieving a similar result (accuracy = 83%, MCC = 0.66) using a higher number of features (128) (experiment 5).

In the last experiment, the bot-flag value is added as a feature along with the 28 features identified in experiment 6. As discussed previously, false tweets generated by a bot account can spread on a wide scale. This is an example of a low-quality post from a source with high bot score (4.6): ‘There are long term ramifications to many of these vaccines. Consider the fact they contain Formaldehyde, Mercury & Aluminum: which passes through the blood brain barrier and causes neurological damage, from lower IQs to eventual serious conditions such as Alzheimer’s & dementia’. Experiment 7 aimed to investigate whether adding a bot score feature as a bot/human flag, indicating how likely the account is to be human or bot, would change the classifier’s performance in classifying a post as true or false. Bot scores were scaled to three values for the flag: 0 for scores in the range 0 - 1.9, 1 for 2 - 3.5, and 2 for  $\geq 3.5$ . Then the 29 features were tested with different algorithms. The results show an improvement in the accuracy values of RF (2%) and LR (1%). However, the accuracy dropped by 2% for DT and by 1% for SVM. The accuracy of KNN remained the same. The MCC value reached 0.72 for RF, which indicates strong performance, implying that the bot flag correlated with important features, including the linguistic features of tweets. This helped to improve the accuracy. This suggests that this single feature led to a more accurate classification and analysing dementia information quality for some algorithms such as RF and LR.

Lastly, some steps were implemented to mitigate model bias and overfitting due to the class imbalance and dataset size. These steps included feature selection (which dealt with the problem of overfitting, making the ML models more robust, (Ammu & Preeja, 2013), stratified 10-fold cross-validation (Berrar, 2019), as well as reporting on

different model metrics, including MCC. However, it would be valuable to also report the performance of the model on unseen topics. One way to achieve this is to train the model on all but one topic and then test it on the unseen topic. Therefore, different experiments were conducted to explore the variation in classification performance of the model when trained on all three dataset topics and tested on the fourth unseen topic using the RF classifier with the most important 27 features. Table 6.8 lists the classifier performance of each topic that is eliminated from training the classifier and then used for testing. In general, the model performance is lower when one topic is excluded from training. When the model was tested on the excluded sub-datasets (namely flu vaccination, cannabis/marijuana, and aluminium), the precision, recall, F1 and accuracy were higher than .50, however, the MCC values were close to 0 (0.02, 0.14, 0.27 for flu vaccination, cannabis/marijuana, and aluminium, respectively). As discussed earlier, MCC is an informative metric for measuring classification performance on unbalanced data; it takes a value in the range  $[-1, +1]$ . If it is close to  $+1$ , all four fundamental rates of the confusion matrix have consistently high values, but when it is closer to  $-1$ , it means the model prediction is random. The reason for the low model performance on these three topics could be the unbalanced data sizes of classes (0 and 1) in the individual sub-datasets, so the model was unable to learn enough from the data in the training set when a sub-dataset was eliminated from the training dataset. However, the model tested on the 'Vitamin' sub-dataset showed moderate performance (MCC = 0.51, accuracy = 0.80), which is comparable to the performance of the model trained on all datasets (MCC = 0.64, accuracy = 0.84). This could suggest that the model is quite topic independent. Yet, a more detailed study with a larger dataset and balanced classes is required to verify these findings.

Chapter 6. Study 2: Machine Learning Experiments

Table 6.7: Performance of the ML algorithms using different features sets

Classifier	Precision	Recall	Accuracy	MCC	Time	Class	Precision	Recall	F1-Score	Support
Experiment 1: Domain features (3)										
RF	0.71	0.70	0.70	0.37	9.57	0	0.69	0.88	0.77	252
						1	0.73	0.45	0.56	182
DT	0.69	0.68	0.68	0.32	.55	0	0.67	0.89	0.76	252
						1	0.71	0.38	0.50	182
SVM	0.72	0.71	0.71	0.40	24.7	0	0.70	0.89	0.78	252
						1	0.75	0.46	0.57	182
LR	0.71	0.71	0.71	0.38	0.05	0	0.69	0.88	0.78	252
						1	0.74	0.46	0.56	182
KNN	0.69	0.68	0.68	0.33	0.32	0	0.67	0.90	0.77	252
						1	0.73	0.38	0.50	182
Experiment 2: Profile features (4)										
RF	0.69	0.69	0.69	0.37	11.08	0	0.74	0.73	0.73	252
						1	0.63	0.64	0.64	182
DT	0.69	0.69	0.69	0.36	1.02	0	0.73	0.73	0.73	252
						1	0.63	0.62	0.62	182
SVM	0.66	0.59	0.59	0.11	7.54	0	0.59	0.99	0.74	252
						1	0.75	0.05	0.09	182
LR	0.62	0.59	0.59	0.08	0.6	0	0.59	0.98	0.74	252
						1	0.67	0.04	0.08	182
KNN	0.67	0.66	0.68	0.33	0.05	0	0.71	0.76	0.73	252
						1	0.63	0.56	0.59	182
Experiment 3: Posit features (27)										
RF	0.79	0.79	0.79	0.57	14.1	0	0.80	0.86	0.83	252
						1	0.78	0.70	0.74	182
DT	0.79	0.79	0.79	0.56	11.1	0	0.78	0.88	0.83	252
						1	0.80	0.66	0.72	182
SVM	0.79	0.79	0.79	0.57	65.89	0	0.78	0.88	0.83	252
						1	0.81	0.66	0.73	182
LR	0.79	0.79	0.79	0.56	0.37	0	0.77	0.90	0.83	252
						1	0.82	0.63	0.71	182
KNN	0.79	0.78	0.78	0.54	0.22	0	0.78	0.87	0.82	252
						1	0.79	0.66	0.72	182
Experiment 4: LIWC features (90)										
RF	0.80	0.80	0.80	0.60	16.81	0	0.82	0.85	0.83	252
						1	0.78	0.74	0.76	182
DT	0.81	0.81	0.81	0.61	8.42	0	0.83	0.85	0.84	252
						1	0.79	0.75	0.77	182
SVM	0.81	0.81	0.81	0.61	173.80	0	0.81	0.88	0.85	252
						1	0.82	0.71	0.76	182
LR	0.79	0.79	0.79	0.56	0.24	0	0.81	0.83	0.82	252
						1	0.76	0.73	0.75	182
KNN	0.82	0.82	0.82	0.63	0.22	0	0.86	0.83	0.84	252
						1	0.77	0.81	0.79	182
Experiment 5: All features (124)										
RF	0.82	0.82	0.82	0.62	16.7	0	0.82	0.88	0.85	252
						1	0.81	0.74	0.77	182
DT	0.80	0.80	0.80	0.59	41.44	0	0.79	0.90	0.84	252
						1	0.82	0.67	0.74	182
SVM	0.80	0.80	0.80	0.59	4.85	0	0.82	0.88	0.85	252
						1	0.82	0.74	0.77	182
LR	0.80	0.80	0.80	0.60	0.6	0	0.82	0.84	0.83	252
						1	0.77	0.75	0.76	182
KNN	0.83	0.83	0.83	0.66	0.23	0	0.86	0.85	0.86	252
						1	0.80	0.81	0.80	182
Experiment 6: Selected features (28)										
RF	0.85	0.84	0.84	0.68	5.48	0	0.83	0.92	0.87	252
						1	0.87	0.74	0.80	182
DT	0.81	0.81	0.81	0.61	14.12	0	0.82	0.87	0.84	252
						1	0.81	0.73	0.77	182
SVM	0.84	0.84	0.84	0.66	24.18	0	0.84	0.89	0.86	252
						1	0.84	0.76	0.80	182
LR	0.84	0.84	0.84	0.67	0.49	0	0.84	0.89	0.87	252
						1	0.84	0.76	0.80	182
KNN	0.83	0.83	0.83	0.65	0.22	0	0.84	0.88	0.86	252
						1	0.82	0.76	0.79	182
Experiment 7: Selected features with bot-flag feature (29)										
RF	0.87	0.86	<b>0.86</b>	<b>0.72</b>	5.54	0	0.85	0.93	0.89	252
			1	0.89		0.77	0.83	182		
DT	0.79	0.79	0.79	0.56	13.93	0	0.82	0.84	0.83	252
						1	0.77	0.74	0.75	182
SVM	0.83	0.83	0.83	0.65	23.76	0	0.83	0.89	0.86	252
						1	0.83	0.75	0.79	182
LR	0.85	0.85	0.85	0.69	0.2	0	0.84	0.90	0.87	252
						1	0.85	0.77	0.81	182
KNN	0.83	0.83	0.83	0.65	0.22	0	0.84	0.87	0.86	252
						1	0.81	0.77	0.79	182

\*Highest accuracy score is in bold

Table 6.8: RF classification performance on each topic eliminated from training the classifier and then used for testing using 27 selected features.

Unseen Topic	Precision	Recall	F1	Accuracy	MCC
Topic 1	0.73	0.58	0.59	0.83	0.27
Topic 2	0.57	0.52	0.51	0.79	0.07
Topic 3	0.81	0.72	0.74	0.80	0.51
Topic 4	0.58	0.56	0.57	0.76	0.14

\* Topic 1= Myth 1 (aluminium), Topic 2 = Myth 2 (flu vaccination), Topic 3 = Myth 3 (vitamins), Topic 4= Myth 4 (cannabis/marijuana)

## 6.4 Discussion

The research questions for this chapter were: RQ4: To what extent, if at all, do the most active bots contribute to spreading low-quality dementia-related information on Twitter? Which bot types have the greatest involvement in the spread of low-quality dementia-related information on Twitter? RQ5: What are the most effective features to improve the automated assessment of dementia information quality?

The results show that bots play a significant role in spreading low-quality dementia information (myths) on Twitter. The findings show that 59% of myth authors are bots and 41% comes from human-likely profiles. The negative impact of bots is higher due to their ability to amplify the information faster using automated methods (e.g., replies, retweets). The bot types others, self-declared, spammer, and fake followers are dominant in spreading dementia myths, while the financial and astroturf bots contributed only minimally. The fake followers type is an amplification bot that can be used to increase the reach of a particular account (Jamison, Broniatowski, & Quinn, 2019). It is not clear what the exact purpose is of the first two bot types, other and self-declared, however, spammers and fake followers have a presence in most of the myths datasets as well and their purpose is easier to ascertain. Such bot types may be especially effective in promoting low-quality information or products targeted at vulnerable populations (such as PWD, their families, or caregivers) who might have low medical literacy.

This study expands on existing research on bot evaluation, which has been largely focused on the political domain (Bessi & Ferrara, 2016; Ferrara, 2017; Shao et al., 2018),

to include the health domain, particularly that of dementia, which is a major public health problem (Selbæk, 2021). Low quality information may lead to serious health damage and economic burdens on dementia patients, their families and governments.

The findings of the current study also align with the call made by (Jamison et al., 2019) for studies to analyse the involvement of bots in the propagation of health misinformation across numerous platforms. Researchers, practitioners and policymakers need to be enabled to mitigate the negative effects of bad actors in the public health domain (Jamison et al., 2019). Understanding the diversity of malicious actors is necessary to properly comprehend how public opinions are manipulated on Twitter (Jamison et al., 2019). The present study investigated the use of bots in dementia contexts. The results indicate more than a third of low-quality information authors are bots. Finally, these findings could also encourage researchers to improve bot detection, particularly those researchers who recently started specialising in the health domain (Davoudi, Klein, Sarker, & Gonzalez-Hernandez, 2020), to detect malicious health bots and to prevent their negative impact.

To determine the most effective features to improve the automated assessment of dementia information quality, the principal bot features found in the first study were expanded to include a variety of linguistic features. The current study introduced a new dementia related dataset, which was retrieved and labelled manually. Experiments were then conducted on this dataset. Different ML classifiers using GridSearchCV with stratified 10-fold cross-validation techniques were applied to different combinations of features from profile, domain, and linguistics levels. Initially, the bot and human features identified in the first study were used as candidates for the analysis, however, the linguistic and domain features were enhanced to include psycholinguistic features, hashtags and mentions. The impact of different sets of features on the accuracy of classification was tested separately. While much of the existing research has compared the outcomes of different types of classification algorithms using one feature selection method, as in (Sicilia, Giudice, et al., 2018) and (Sicilia, Merone, et al., 2018), the current study explored the selection of different features by combining feature selection methods, whereafter the top features were selected. These were then tested on

different classification algorithms in order to achieve a high accuracy performance in the classification as suggested by (Salih & Abdulrazaq, 2019).

The results show that the RF model applied with 28 selected features had the highest accuracy (84%) and a moderate MCC value (0.68) compared to the other models. The results demonstrate the effectiveness of the selected features, which can correctly classify about 83% of the myths with 92% precision. The selected features did not include any profile features. Twenty-seven features are linguistic features (lexical, syntactic, and psycholinguistic) and one is a domain feature, URL. The findings suggest that the automated assessment of the text quality relied mainly on linguistic features. Although features reported in other studies (see Section 3.3) are effective, they are more likely to be dynamic and subject to bot manipulation. For example, profile features (e.g., favourites count) and/or post popularity features (e.g., retweets) are frequently manipulated by bad actors such as bots, who can thus create an entirely false impression of credibility (Qureshi et al., 2021).

The current study also explored the effectiveness of the bot score feature with other features when automatically analysing the tweet quality. No research has yet attempted to use this feature for quality assessment of health information on the single tweet level. The results reveal that the bot score feature is useful in quality assessment, as it has good discriminative power that enhances myth identification. The accuracy and MCC of the best model (RF) using different combination of features improved by 2% to 86% and by 4% to .72, respectively, reflecting very good performance. Even though bot profile features revealed in the first study do not contribute to classifying low-quality features, other content features such as linguistics and domain-specific features identified in the first study have proven to be effective in identifying low-quality information. This indicates that malicious actors or bots must be identified and addressed for quality assessment.

## Chapter Summary

Previous studies have documented the effectiveness of ML models when automatically analysing or predicting information quality on social platforms in different contexts, yet the health context has been less investigated. To build ML models, a variety of hand-crafted features or features reported in literature (e.g., temporal, popularity, propagation, as well as text-based features) are used. Despite the similarity of features utilised in this and existing studies for automating the assessment of dementia information quality and credibility on SM, where the use of these features achieved good results, more explanation of factors contributing to the automation assessment is needed. Therefore, the primary goal of this study is to analyse bot features in the specific context (dementia) and how they can inform the text quality classification. To achieve this, a systematic investigation of different bot features (including profile and content features) was conducted in Study 1. The most important features in evaluating bots were revealed to be: four features of user profiles (verified, friends count, geo data and favourites count), 27 linguistic features (syntactic and lexical), and domain features (URL). These principal bot features found in the first study were expanded to include diverse linguistic features, in order to improve the automated classification. The RF algorithm is able to correctly predict whether the information provided in a tweet is true or false with an accuracy level of 85% and an MCC value of 0.68. It is also evident from the classification results that profile features have little effect, whereas linguistic features significantly improve the performance of the classification. Importantly, six of the selected features, such as number of characters, number of sentences, AWL, total unique words, noun types, and URL count) in the binary classification of dementia information quality were the same as those that resulted from the first study of feature analysis of bot and human profiles. These results suggest that employing bot features can be an effective strategy for training a ML classifier for information quality classification.



# Chapter 7

## Study 3 : User Study

### 7.1 Introduction

This chapter addresses the following research question:

- **RQ6:** What are the factors used by information consumers to assess the credibility of dementia information on Twitter?

As discussed in Chapter 3 Section 3.1.1, few scholars have studied the credibility of health information on Twitter, particularly taking into consideration the prevalence of bots. Various quantitative methods (e.g., surveys and questionnaires) have been applied to determine the factors influencing people's assessment of the credibility of information on SM. These approaches create or test theories that include one or two credibility components as a dependent variable. However, because the constructs used to establish credibility do not have a standard definition, and the connection between variables is ambiguous, it is difficult to get consistent results (Sbaffi et al., 2017). Thus, more qualitative studies are required (Sbaffi et al., 2017). Additionally, most previous studies have used screenshots of manipulated Twitter feeds rather than live views (Spence et al., 2019; Edwards et al., 2014) and it is likely that there would be a difference in perception depending on whether the user views a static view or a live view of the feed (Edwards et al., 2014). It is therefore essential to examine participants' perceptions using live feeds. This research fills this important gap by developing qualitative studies

to gain a deeper understanding of health information judgements on SM platforms, while also considering the role of bots. Understanding the factors that impact the quality assessment of online health information is important to assist with creating health education programmes, information content, and systems, and plotting patient-supplier communications (Sbaffi et al., 2017), as well as to automatically evaluate the information quality.

This chapter is a revised version of (Alhayan, Pennington, & Ruthven, 2022).

## 7.2 Methods

In the current study, data was collected through a concurrent think-aloud session followed by a retrospective interviewing session (post-task interview) and was used in the form of semi-structured interviews.

As described in Section 4.2.2, the concurrent session involves asking participants to verbalise their thoughts while the task is being conducted and the retrospective session involves asking the participants about their thoughts after conducting the task. The reason for employing concurrent think-aloud is that it aids in obtaining immediate thoughts while carrying out the task, whereas retrospective interview is useful when participants do not articulate the ideas sufficiently.

Think-aloud interviews is commonly used to reveal possible factors influencing the credibility judgement of health-related information on websites and during online search (Ghenai et al., 2020; Kattenbeck & Elswailer, 2019; Klawitter & Hargittai, 2018; Muntinga & Taylor, 2018).

### 7.2.1 Study Design

The researcher selected six Twitter profiles for participants to consider. These profiles were sampled from the profiles collected through the second study. Social media users typically look at the whole profile rather than relying on individual tweets only. Therefore, the assumption is that a complete examination of the whole profile with all its features would enable participants to reach a decision on credibility in a more realistic

way. Thus, total profiles were provided rather than showing individual tweets only. The profiles were chosen to ensure that 1) most tweets on the profiles were dementia centric, as determined by two independent assessors, and 2) there were two publicly available profiles from each of the following categories: *organisations*, *professionals*, and *individuals*. The researcher chose these because they embody the primary profile types that tweet about dementia (Alhayan & Pennington, 2020). *Organisations* contained two dementia-related organisations, *professionals* featured two dementia researchers, and *individuals* comprised two partners/caregivers of PWD. Categorisation was based on the Twitter bio. In each category, one profile had an extremely high bot score and the other had an extremely low bot score, as calculated by the Botometer API (K.-C. Yang et al., 2019) (for more details about Botometer see Chapter 4, Section 4.3.3). Accounts with high bot scores mostly included automatically generated tweets or retweets. While some sources were reputable, other sources' posts had misleading information; for instance, information on how certain fruits or a particular exercise can prevent the risk of dementia or memory loss. A sample screenshot of tweets from an individual profile with a high bot score appears in Appendix O. Providing a chosen list of webpages is common in think-aloud web studies; for example, Kattenbeck and Elswailer (2019) selected eight search engine results page (SERP) listings (four credible, four non-credible) to study credibility judgements on three controversial topics: topics: the safety of autonomous vehicles, the legalisation of cannabis and the healthiness of a vegan diet. The research presented in Ghenai et al. (2020) also used SERP listings, representing either correct or incorrect information to understand factors affecting online health search.

All six profiles contained a biography, profile photos, location, the year the user joined Twitter, and dementia-related tweets. Participants were asked to think aloud while assessing the pair of profiles in each category, potentially choosing one of the accounts or neither as credible, and providing reasons for their choice. Participants were free to navigate the profile content without time constraints, reading as many tweets as they preferred, and going back to profiles whenever they wanted. The longest think-aloud session was about 15 minutes. Participants were also asked to rate each

account's credibility on a Likert scale, with 1 being the least credible and 7 the most credible, to determine their confidence in choosing a profile.

### 7.2.2 Participants and Recruitment

The researcher recruited a purposive sample of formal and informal caregivers of PWD who live in the UK and use Twitter. Criterion-based participant sampling and the decision to sample caregivers was explained in Section 4.1. The researcher posted a flyer on Twitter with a link to the study registration form. UK dementia organisations were invited to share the flyer by retweeting or sharing by email. Participants read the information sheet, signed the consent form, and provided their preferred interview time and contact details. They received a confirmation email within 24 hours, containing the interview date, time, and a secure Zoom link.

A day before the interview, participants received a link to an online questionnaire (see Appendix P) as well as Zoom interview instructions by email. The questionnaire gathered basic demographics, frequency of Twitter and other SM usage, and general questions regarding the profiles and information types usually read on Twitter. The Zoom interview instructions came in the form of a five-minute video on computer requirements and Zoom screen sharing.

Participants were recruited until data saturation was reached; the last two participants did not reveal new insights (O'reilly & Parker, 2013). Rich and in-depth data are the focus of think-aloud studies, and sample sizes are fairly small (Van Someren et al., 1994). The final sample included six formal and seven informal caregivers for PWD at different stages. Twelve caregivers were female, and one was male, with ages ranging between 21-35 (3), 36-50 (2), and 51+ (7). Education levels spanned undergraduate (5), some college (6), and postgraduate (2). Twelve participants had used Twitter for over a year. Eight participants used Twitter daily, one weekly, two monthly, one occasionally, and one was not sure. Participants received a £20 e-gift voucher after completing the interview; this was later increased to £40 to encourage participation.

### 7.2.3 Study Procedure

First, the interviewer-initiated discussion about the participant's questionnaire responses about their preferred Twitter sources, categories (e.g., organisations, professionals), and information types. To start the CTA part, the interviewer sent a link via Zoom's chat box with the task scenario and task instructions (see Appendix Q). The participant was asked to read the task scenario and instructions and then shared their screen for the researcher to observe their interactions with the profiles. Participants engaged in an approximately 15-minute CTA that entailed assessing live views of Twitter profiles and they had a chance to ask questions.

As soon as the participants had completed the task, they took part in retrospective session (a post-task semi-structured interview) to elaborate on their statements during the think-aloud session. In the RTA part, the participants had access to the same profiles they browsed. The participants were asked for their assessments of the profiles they did not select and what they felt constituted a credible source generally (e.g., in your opinion, what do you think about other accounts that you did not select? Can you explain in your own words what a credible source is on Twitter?). Interviews lasted 15-20 minutes. RTA and CTA data was captured through audio recording of the participants using a Zoom recording.

## 7.3 Data Coding and Analysis

Data collected through the concurrent session and retrospective session (semi-structured interviews) was transcribed and analysed using conventional qualitative content analysis, in which coding categories are gained directly from the text (Hsieh & Shannon, 2005) (see Chapter 4 Section 4.2.1 for more details).

First, the names of the participants were masked to hide all personal identifiable information from text files (Gibbs, 2018). The Zoom auto transcription was checked to ensure transcription accuracy. Then, the data was organised, which included open coding and creating categories. The open coding process started with a small number of transcriptions of interviews, four only. Four transcripts were given to a coder (another

PhD colleague) for independent coding. After the first initial coding by both the researcher and the coder, an in-person meeting was held to discuss the similarities and differences in the codes and to resolve issues as they occurred. A consensus on the main categories was established. Both agreed that themes can be grouped to three main categories: source, content, and user. However, a slight modification was made to some codes' definitions and in categorising the sub-categories. Specifically, the researcher and the coder disagreed in two places related to the sub-category results. "Interactivity" related quotes were defined as a separate sub-category for the source by the coder. After the discussion, the coder and researcher agreed that quotes under "interactivity" indicate the source presence and can be a subgroup of "social presence". The second difference of opinion was about the reader's "relevance" and "knowledge" which was coded as one category by the researcher. However, the coder suggested that there should be two different themes based on different participants' quotes. Then, based on the discussion, the refined coding scheme was used by both the researcher and coder to code the same transcripts, reading each transcript line by line, to refine the coding scheme by exchanging opinions immediately when any disagreements or discussion points emerged. Following Gibbs (2018), the process was iterated until both researcher and coder agreed on the given themes and codes. The researcher used the final coding scheme as refined in the third phase to code the rest of the interviews. In the final stage, the researcher reported the results with the main themes and their sub-themes, as shown in the following section.

In order to enhance the overall validity of the qualitative study, every interview transcript text was read multiple times. A senior researcher was asked to provide an objective assessment of coding results and the interpretation of the interview data.

## 7.4 Findings

The analysis resulted in three main categories of credibility assessment dimensions: source, content, and user, as well as 13 subcategories that support different main categories, as illustrated in Figure 7.1. Each of the categories is described in more detail

Table 7.1: Participants' credibility ratings.

Profile type	Ratings		
	Unsure (1-2)	Indecisive (3-5)	Sure (6-7)
Bot-likely	8	24	7
Human-likely	7	18	14

below in Sections 7.4.1 to Section 7.4.3.

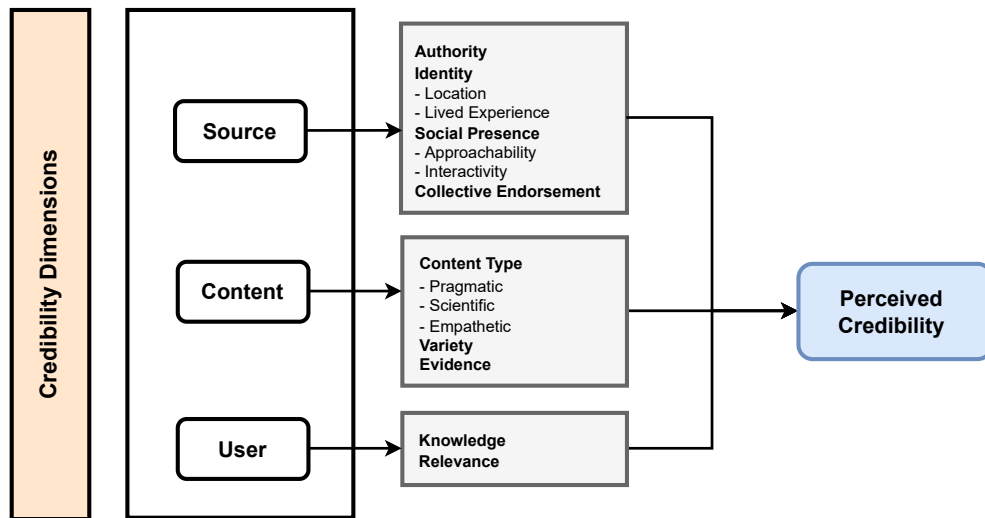


Figure 7.1: Qualitative results.

Table 7.1 reports the participants' credibility ratings collected from participants during concurrent session. Significantly, a comparatively high number of profiles were rated between 3-5, indicating indecision. Fourteen human-likely profiles were rated as credible, and only seven bot-likely profiles were rated credible. This pattern suggests that the participants are frequently perplexed as to the credibility of dementia-related material found on SM.

### 7.4.1 Source

This category captures participants' assessment criteria for determining the credibility of the source. Source refers to the Twitter profile and is based on features such as profile description, location and picture, which the profile owner chooses to share publicly.

Source refers to Twitter profile or Twitter user (see Chapter 1.3 for more details about Twitter terms). Subcategories for source are: authority, identity, social presence, and collective endorsement.

#### 7.4.1.1 Authority

Ten participants considered the source as credible if they recognised the author of the tweets as an authoritative source. Participants saw authority as subject-related experts with a scientific background and/or qualification, an official entity, or a well-recognised public account. They also noted the blue badge (verified feature) assigned by Twitter, which validates an account of public interest as authentic. Some points of view on authority from participants are provided below. The + or – after each quote indicates whether the related statement positively or negatively affected the participants' perception. In general, authority had a positive effect on source credibility assessment.

**P6:** 'I would find that [profile] less credible than someone with a qualification or a recognised organisation' (-)

**P7:** 'If it has a blue tick, then it's a credible account' (+)

**P10:** 'Just to make sure that it was all credible. What their qualifications were. What they'd studied' (+)

#### 7.4.1.2 Identity

Participants discussed identity when the source declared their geographic location and lived personal experience as patient or caregiver. Location is declared either in the source's profile or in their posts. The participants identified the account holder's location as an identity cue when they browsed a profile. Nine participants believed that profiles of individuals and organisations from the same country were more credible.

**P8:** 'It seems a credible source, but because it's not in the UK, I wouldn't probably follow it' (-)

**P6:** 'I wouldn't be interested in that because that was in the USA' (-)

**P10:** 'I would be more inclined to go for a UK based one' (+)



A second type of identity cue is lived personal experience, defined as being a caregiver or a PWD as stated either in profile descriptions or posts. Six participants showed an interest in finding people with the same situation, either living with dementia or caring for a person with dementia.

**P3:** ‘Just because she seems to be dealing more with what I deal with on a daily basis’ (+)

**P5:** ‘I think perhaps in terms of experiences, personal experiences, I might go with it, who either has lived experience and has done something with that?’ (+)

### 7.4.1.3 Social Presence

Participants expressed the social presence of the source in two different ways: approachability and interactivity.

Approachability is the feeling that the source is human and there is a possibility of direct interaction with the source. During the think-aloud session, participants experienced greater feelings of social presence while evaluating a source in the professional category; this source was more likely to be perceived as a social actor. Some of the profile posts included their contact details (e.g., telephone number). The following comments were made by participants while assessing that source:

**P5:** ‘Okay, so she’s posting things that make her more human’ (+)

**P9:** ‘It sounds more friendly . . . she said, you know, you can call, you can chat. Which I think is good, whereas the others no. I didn’t get that impression of the others.’ (+)

**P4:** ‘I like her approach. I like the fact that she was actually asking for experiences. It was much more tailored to the needs of someone caring for someone with dementia’ (+)

Interactivity is expressed in terms of ongoing tweeting activity, either writing one’s own tweets or retweeting. One participant, for example, felt a stronger social presence when evaluating a user who tweets more than retweets:

**P11:** ‘She seems more human. The other one . . . . a lot of retweets and not really about him, as such . . . . Can see a human behind the feed whereas the other link has no personal feel . . . whereas there . . . . further wasn’t enough evidence of the human

behind any of those accounts and that might suggest that it's a bot or not an actual human being and have got thing about people that only retweet. I like people that give their own opinion.' (-)

In contrast, another participant did not sense a strong social presence from someone who tweets more than retweets:

**P2:** 'Wouldn't say is particularly good because it's not taking into account like of a wide range of people's opinions.' (-) **P2:** 'Uses fellow colleges and alike as well as himself, not self-involved and purely wants good information out there ... It's not self-involved. So, it's, it seems more like actual information output rather than a personal account that's kind of trying to like glorify themselves.' (+)

Two other participants experienced greater feelings of user presence from users who tweet regularly:

**P7:** 'Yet they haven't posted since June and July. So, I want to be kept up to date. They're not active, they're not recent.' (-)

**P8:** 'A more credible account. Even though it was in America. They are actively sharing information daily.' (+)

#### 7.4.1.4 Collective Endorsement

Participants considered how other people view the source. This was measured by the number of followers the source has, who the followers are, the number of likes on the source's posts, whether they and the participants have mutual friends, or even through a recommendation from a trusted person offline. The endorsement could be from one user or a group of people:

**P1:** 'I think the more followers you have probably the more credible, the more reliable source.' (+)

**P5:** 'Number of followers is a factor. It's a conscious factor I should say that.' (+)

**P5:** 'If I go back a step, this profile is followed by [redacted]. [Redacted] is a dementia expert by experience. He has dementia and I've followed him and attended some conferences he's spoken at prior to COVID so that immediately I like in terms of reliability.' (+)

## 7.4.2 Content

The posts on profiles also provided meaningful insight into the participants' credibility perspective. Content is the source timeline, or aggregated stream of tweets, retweets, replies, and quote tweets (see Section 1.3 for more Twitter terms and definitions). Content category is subdivided into content type, variety, and evidence.

### 7.4.2.1 Content Type

Content type refers to what main aspect the message discusses. Three different content types, scientific, pragmatic, and empathetic, were identified while assessing the information credibility. Six participants showed interest in sources providing scientific information, in other words, sources where most of the tweets included medical or scientific research findings. Generally, these tweets include links to external websites or cite information on external platforms.

**P12:** 'He would be the more credible source. He seems to be sharing, you know, this is sharing more of studies . . . The other one just raising awareness.' (+)

In contrast to the purely scientific information in some sources, seven participants showed interest in sources' pragmatic information. This means practical ideas or tips that can be usefully applied in caregivers' everyday lives.

**P7:** 'Account offers more practical ways of making life more bearable for people living with dementia.' (+)

**P11:** 'Appears to share more practical information and tweets about own experience which may be more helpful than seeing scientific papers in second link.' (+)

Participants also viewed sources providing empathetic content as more credible. Empathetic content conveys expressions of emotion or feeling, including caring, helping, appreciating, and supporting, as well as faith-based support.

**P7:** 'Going to assume an accredited account of course everything they have to say matters to me as a caregiver and especially because they are offering support for their carers.' (+)

**P10:** 'We're here to help carry the burdens that would be brilliant.' (+)

**P4:** 'The faith based one I liked, but I wasn't sure how applicable it would be to

me because we have Christians and think that I would like a faith-based support, but what. . . there isn't a Christian tradition.' (+)

#### 7.4.2.2 Variety

Variety in the content refers to topic diversity, whether focused on one or several aspects. Two different points of view emerged about the content variety. One participant, for example, showed an interest in accepting an information source as credible if it is focused only on one topic rather than discussing a range of topics.

**P12:** 'I would pick him because it tends to be more of a focus on actual dementia.' (+)

However, other participants expected a wide range of information from a source and rated the source as credible if the content covered different topics such as drugs and medicines or research articles.

**P10:** 'It just seems to focus on the one drug, and it doesn't seem to sort of focus on too many you know, other things.' (-)

#### 7.4.2.3 Evidence

Another criterion referred to evidence for scientific claims: if evidence was available within the content, the information source was deemed credible. Evidence refers to links (URLs) to other sources supporting the information.

**P13:** 'Credible source showing links to other sources.' (+)

**P12:** 'If someone's just saying, like, fruit can help prevent dementia. Well, I need to know why that you're saying that what's your rationale behind its kind of thing, where have they got this information from because the internet is not reliable at all.' (+)

Another group of participants questioned the quality and reliability of the provided references and links, believing that information based on authentic, reliable, high-quality sources would be credible.

**P5:** 'She's posting information or links to information that looked like it would be very useful and that from reliable sources as well. So, I see University of [redacted], I recognise that name Centre for Dementia Studies, Alzheimer's Society. I know. So,

I'm drawn to this.' (+)

**P10:** 'Links they are sharing and go to the links and see what the quality information that they were receiving; were the links they're sharing from credible sources?' (+)

Participants also mentioned that if the majority of users' tweets includes links to the same websites, these could not be trusted.

**P2:** 'Kind of links they have are from the same website...because if it's from the same websites, it's most likely the same people who are writing articles... It doesn't seem like a bad organisation, in any way; it just seems as not as reliable as something else would be. I probably would not use it.' (-)

### 7.4.3 User

This refers to the participant who has been asked to assess the profile rather than the profile itself. In this study, users were formal and informal caregivers. Some characteristics related to the users influence their assessment of credibility, including relevance to the user personally and the user's prior knowledge. Relevance refers to the participant's interest in the content provided by the source. Relevance was frequently mentioned during profile assessments.

#### 7.4.3.1 Relevance

Relevance refers to the participant's interest in the content provided by the source.

**P4:** 'They're just too medical, they're not they're not something that I'm that interested in to be honest. As I said, I am more interested about the care and support that one might need after diagnosis.' (-)

#### 7.4.3.2 Prior Knowledge

Prior knowledge refers to the participant's ability to understand and interpret the content provided by the source.

**P5:** 'Interesting material that I have some knowledge of and can understand.' (+)

Lack of knowledge was mentioned as a reason for not being able to understand the information; for example, scientific terms used in most of a profile's tweets:

**P5:** ‘I wouldn’t follow [user1] purely and simply because I wouldn’t have a clue what he’s talking about and to be honest, I don’t have the time to go in and look it up.’ (-)

**P13:** ‘I would have difficulty making a decision whether that’s credible ... it looks way above anything that I intellectually could understand.’ (-)

## 7.5 Discussion

Most source-related heuristics revealed by the participants in this study – authority, identity, collective endorsement, and social presence – are in line with the agency affordance heuristics provided in the MAIN model (Sundar, 2008), as discussed in Section 2.3. In addition, the role of content in credibility assessments and user characteristics as uncovered in this study is found to be complementary to technological agency affordances in the MAIN model (Sundar, 2008). As discussed previously, the MAIN model proposes four classes of technological affordances that can trigger cognitive heuristics which affect credibility judgements: modality (M), agency (A), interactivity (I), and navigability (N). Modality deals with the medium through which data is presented, interactivity implies both interaction and activity with devices, and navigability focuses on interface cues helping with navigation in cyberspace. The agency affordance deals with the source of information on digital media such as websites, a poll of friends on SM, or a person having a profile on an online platform.

In this study, participants identified 13 different factors supporting three main credibility dimensions, namely source, content, and user characteristics. Four of the 13 factors are source related heuristics and they were explained with respect to the agency affordance in the MAIN model (Sundar, 2008). First, participants in this study identified a source as an authority when the source was a domain expert or an official entity (Sundar, 2008). This aligns with the general findings in the credibility literature showing that authority impacts credibility evaluation (Rieh, 2010; Lin et al., 2016). Second, the identity heuristic is likely to be triggered whenever the affordance enables users to express themselves through manipulating content and asserting their identity. The user interface of SM platforms can be designed to generate different verifications of

identity, and potential followers may use these for their own evaluation of a profile. People can evaluate a source's name, location, profile photo, or other identifiers. Various identity-related parameters for credibility assessment have been used in prior research for different contexts, such as profile pictures for evaluating online news comments (Lin, Kaufmann, Spence, & Lachlan, 2019) and LinkedIn profiles (Edwards, Stoll, Faculak, & Karman, 2015). Another identity cue, nationality, is used during online shopping; if someone from the consumer's own country has provided a review, this is trusted more than a review from someone in a different country (Bracamonte & Okada, 2015). The current study shows that identity was perceived in two forms: source location and personal experience as a caregiver or PWD. All participants were UK residents, and they evaluated sources located outside of the UK as less credible. Third, social presence as a concept means feeling the presence of other people irrespective of technology use (K. M. Lee, 2004). Social presence heuristics, triggered by agency cues, may provoke feelings of the presence of another entity (Sundar, 2008). Feeling the social presence of other entities (human or machine) develops trust in the system (K. J. Kim, Park, & Sundar, 2013; Lu, Fan, & Zhou, 2016). Prior studies have demonstrated different cues of social presence on Twitter influencing users' credibility perceptions. Son, Lee, Oh, Lee, and Woo (2020) identified Twitter account age as a cue for social presence in disaster situations. Other research investigated dynamic features, such as the relationship between levels of the source's timeline interactivity, as social presence cues: if a source's timeline is highly interactive (e.g., a political figure), expressed by the number of replies provided to followers, this results in greater social presence (E.-J. Lee & Shin, 2012). Dialogic retweets, or retweets of users who mentioned the organisation, produced a higher level of social presence compared to monologic or "one-way" tweets from the organisation (Lim & Lee-Won, 2017). A key finding of the current study is that participants use social presence to detect human characteristics. It has thus extended past research by incorporating different perspectives on a source's profile interactivity and added approachability as another lens for social presence. Interactivity involved frequent tweeting. Contradictory perceptions of sources' tweeting interactivity were also shown. For some participants, the source who tweeted more than retweeted generated

a greater sense of social presence, while this was not the case for other participants. These varied perceptions of social presence should be further explored, because bots can be set to retweet as well as automatically reply to tweets, pretending to be real people. Fourth, bandwagon heuristics in the agency affordance reflects a group endorsement of the source's reputation, which impacts on its credibility. Collective endorsement has been observed as a factor affecting credibility perception in news (Q. Xu, 2013) and on online health forums (Jucks & Thon, 2017). Similarly, endorsements disclosed in this study included the number of followers the source has, who the followers are, the number of likes on the source's posts, and if they have mutual friends. However, one participant included a recommendation by a trusted person offline.

The other dimension for credibility evaluation relates to content features. Participants also evaluated the profile based on content type, variety of content, and evidence. Referring to type of content, empathetic content is an important cue for the evaluation of credibility. Interestingly, faith-based content also contributed to enhanced credibility perceptions. Although empathetic content has been identified as an important aspect in evaluating general web health information sources by studies such as (Neal & McKenzie, 2011), no work has directly examined the relationship between individual perception of health information and faith-based content or religiosity on SM.

Mixed perceptions were observed regarding the variety of content. Although a few participants were interested only in topic-focused information, some showed interest in multiple types of information. Participants also assessed content in light of evidence or references provided with the tweet. Some participants did not believe in the contents if they were posted by the same web source that also ran the Twitter account.

Users' knowledge and the relevance of the contents also play a vital role in the evaluation. If the contents were related to the participant's needs or experiential background, the source was identified as more credible. This is unlike the results in (Unkel & Haas, 2017), where knowledge did not impact the participants' credibility perceptions of search engine results, yet it agrees with topic knowledge of many credibility assessment models on the general web (Rieh & Belkin, 2000; Lucassen & Schraagen, 2011).



## Chapter Summary

Given the growing popularity of Twitter bots and users' ability to share content without gatekeepers' filters, it is imperative to understand what factors affect credibility assessments of information on social platforms. The study provided a step towards a qualitative assessment of user perceptions of SM health information and suggests a direction toward generalising for other domains.

Although some of these findings are common in the previously discussed models, this study cannot be directly connected to any one credibility assessment model. Allowing participants to evaluate the full "live" profile using the think-aloud method provided a means to observe that users do not only rely on source heuristics. The findings demonstrate the importance of qualitative studies that help establish the role of users' prior knowledge and relevance in information processing. It has also shown the essential requirement of adopting systematic processing metrics for credibility evaluation in health information on SM along with source heuristics. Most existing models and frameworks for credibility assessment have been developed with the perspective of information available on static web resources. In conclusion, the findings indicate the necessity to incorporate all three credibility facets in order to understand participants' perceived credibility.

## Chapter 8

# Conclusions

‘Computers may be clever, but  
human beings are much smarter.  
We invented the computer’

---

Jack Ma

This chapter concludes the thesis and summarises the research results of the previous chapters. Based upon the answers to the RQs, a comprehensive framework for assessing dementia information credibility on Twitter is proposed. Additionally, this chapter describes the key contributions and practical implications of the study, and highlights limitations and possible directions for future work.

### 8.1 Key Research Findings

The research investigates the credibility of health information available on social platforms. The scope of this research was limited to information related to dementia on Twitter. The research began by reviewing literature, which revealed that most of the earlier studies on information credibility assessments on social platforms focused on measuring one component of credibility such as source expertise, source trustworthiness, or information quality (Chapter 2) provides a detailed discussion of these components) with very little focus on information about health conditions. Initial studies regarding the credibility of health information on social platforms focused on one as-

pect; namely investigating information quality based on the ML approach. Different automatic models to evaluate health information quality on SM using various types of features were proposed. In these studies, however, malicious bots were disregarded as an indicator of quality. In general, bot identification was found to be the least investigated aspect of credibility assessment methods (Qureshi et al., 2021). Therefore, this study aimed to analyse bot features and investigated how these features would inform a better information quality automated classification on SM. Furthermore, reflecting on what users consider as most significant is important to design automatic models to assess online health information (Al-Jefri, 2019). Yet, very few studies have investigated the perceived credibility aspect of health information on SM from user perspectives. Most existing work has examined the quantitative relationship between specific credibility cues related to specific credibility components and users' perceived credibility assessments, but the health context is the least investigated context.

Thus, the research aims to develop a framework to automatically assess dementia information credibility. This research focuses on two important aspects of information credibility: quality and perceived credibility by information consumers that can complement each other. The quality component deals with two facets: the text quality (e.g., quantifying the syntactic and lexical features) and spam as an indicator of bad quality (e.g., malicious bots) (Ginsca et al., 2015). Perceived credibility refers to the consumer's belief in the credibility of the information based on how they assess various credibility aspects. Therefore, motivated by the existing knowledge gap, this thesis answers the following research questions, using evidence collected from three empirical studies.

- **RQ1:** What profile types participate in dementia-related discussions on Twitter? (Study 1)
- **RQ2:** Are there bot activities in the context of dementia information dissemination on Twitter? If so, what is the relationship between bot patterns and different profile types? (Study 1)
- **RQ3:** What profile features and content features contribute most to demonstrat-

ing bot-like behaviour? (Study 1)

- **RQ4:** To what extent, if at all, do the most active bots contribute to spreading low-quality dementia-related information on Twitter? Which bot types have the greatest involvement in the spread of low-quality dementia-related information on Twitter? (Study 2)
- **RQ5:** What are the most effective features to improve the automated assessment of dementia information quality? (Study 2)
- **RQ6:** What are the factors used by information consumers to assess the credibility of dementia information on Twitter? (Study 3)

This research was conducted using an explanatory sequential mixed methods approach in three phases. It began by exploring the profile types and the role of bots sharing dementia information by quantifying them to better understand the research problem (Study 1). The association between groups of profile types, which were defined based on their profile descriptions and bot likelihood, was investigated. Further investigations involved examining features that characterise humans and bots, focusing on linguistic and profile features. To address RQ1, inductive coding was applied to profile descriptions of the collected sample, which resulted in eight distinct profile types, with the *individual* category being the largest, and the *apps/books* category the smallest. Descriptive analysis was conducted to answer RQ2, which revealed the presence of bots in dementia discussions on Twitter, although there was a general tendency towards human profiles. The *care providers* and *apps/books* categories had a relatively high average bot score compared to other categories. To address RQ3, various statistical tests were applied to two groups of profiles (human-likely and bot-likely profiles) to examine the principal features characterising each. Differences in both profile and content features of bot-like and human-like profiles participating in dementia information were found. Key distinctive features between both profile types are user profile metadata features and content features, including linguistic and URL frequency features. Four profile features, *verified*, *friends\_count*, *geo\_data* and *favorites\_count*, are contributing features

when evaluating bots. The content of bot profiles was also very different in linguistic features and URL counts compared to the content in human profiles. Linguistic features include word-level and sentence-level features (number of words, characters, sentences, unique words, average word length and average sentence length), POS (nouns, verbs and personal pronouns), types of POS (nouns types).

The second empirical study examined the extent to which bots contribute to disseminating myths related to dementia, to answer RQ4. The descriptive analysis showed that both human and bot profiles contribute to false information dissemination; 59% of dementia myth authors are bots whereas 41% are human authors. Most bots belong to the other and self-declared bot types ( $P < .01$ ,  $r = .58$  and  $r = .48$ , respectively). To answer RQ5, the study used ML algorithms to measure the degree to which the features defined in the first study, along with other linguistic and domain features, could aid in classifying tweets based on their quality. To select relevant features contributing to the distinction of myth tweets, feature selection methods were applied, resulting in 28 features. Out of the 28 features, six features (no of characters, no of sentences, AWL, total unique words, noun types, URLs count) were similar to those displayed in the first study's content feature analysis of bot and human. Text quality analysis using different combinations of features employing ML algorithms showed that when relying solely on linguistic features (static features) and domain features (URLs), RF, SVM, and LR, the best performance is attained in terms of accuracy (84%), whereas RF achieved the best result in terms of MCC across all models. The bot-flag feature increased the prediction accuracy of RF and LR, but did not increase the prediction accuracy of DT, SVM, and KNN. However, RF performed better than LR in terms of both accuracy (86% vs. 85%) and MCC measure (0.73 vs. 0.62), respectively.

The third empirical study further elaborates on the findings of the first and second studies for additional explanation of credibility assessments from user perspectives. This was motivated mainly by two facts: First, little research has to date concentrated on people's reliability judgements of health-related information on SM (Keshavarz, 2020), in contrast to the numerous research projects on user assessments of databases and static websites. Second, this study aimed to bridge the gap between the quality

and perceived credibility aspect. To address RQ6, in Chapter 7, factors affecting user perceptions of the credibility of dementia information were investigated by showing “live” feeds of different profile types, including likely-bot and likely-human profiles, to participants. This study used the think-aloud interviews. The profiles used in the study were sampled from the profiles collected in the second study. Automatically generated tweets or retweets were mostly found in accounts with high bot scores. Low-quality information, such as, for example, information on how certain fruits or a particular exercise can prevent dementia or memory loss, was found in some sources, yet some sources were reputable. Social media users typically look at the whole profile rather than relying on individual tweets only. Therefore, the assumption is that a complete examination of the whole profile with all its features would enable participants to reach a decision on credibility in a more realistic way. Thus, complete profiles were provided rather than showing individual tweets only. The analysis of collected data revealed participants relied on 13 credibility criteria, which were grouped into three categories: source, content, and user characteristics (see Figure 7.1). Source related heuristics are found to be in line with the agency (source) technological affordances explained the MAIN model (Sundar, 2008), which suggests that different technological affordances contribute to increasing or decreasing the credibility assessment (Chapter 2 Section 2.3). The study also showed the necessity of adopting systematic processing metrics for content credibility evaluation along with source related heuristics. Importantly, it was shown that users’ prior knowledge and the relevance of the information had a significant impact on the evaluation process.

Even though the ML approach used in Study 2 focuses on the quality of individual tweets, while human perception in Study 3 applies to the whole profile, some disparities and similarities between algorithms and humanly perceived credibility came to light. The presence of evidence in the form of URLs was relevant for both quality and perceived credibility assessment by both ML and humans. URL count was also one of the features used by ML to assess the information quality. Participants’ credibility criteria also included the presence of URLs, especially for scientific claims: if evidence (in the form of a URL) was available within the post, the information was deemed credible

by the participants. In addition, linguistic and lexical features were pertinent in forming decisions by ML, while this was not the case with humans. Humans rather apply systematic processing to thoroughly understand the content or information type (scientific, pragmatic, and empathetic) as a factor in their credibility assessment. Although perceived credibility by human participants could be influenced by the relevance of the information based on information needs and interest in the profile content, this did not apply to ML, where the results are based on the training dataset and features that were utilised. Lastly, when users had to rate their certainty about their assessment (Table 7.1), more than half of profile ratings (42 out of 78) indicated indecision, which means participants could not process the information and determine the credibility in most cases. On the other hand, ML can assess low-quality information in binary classification with reasonable accuracy (84%).

### 8.1.1 The Proposed Framework

This research has connected computational and human-centred approaches to develop a comprehensive framework to automatically assess health information credibility. The proposed framework (Figure 8.1) is based on the criteria deduced from evaluated features by ML algorithms, as presented in Chapter 6, along with features from consumer perspectives as presented in Chapter 7. These features are called predictor features and are structured into two levels: post level and profile level, with various dimensions each. It is notable that all features are independently integrated in the framework and change in value if one attribute of one component does not impact the values of any of the other features. The aggregated predictor features can be supplied as input for supervised ML classifiers to determine the degree of the given information credibility. The framework will be initially automated by generating 28 post-level features, including linguistic and domain features related to a single post. Linguistic features are categorised into syntactic, lexical, and psychological categories, in addition to a domain feature including URLs (see Table 6.4). Automatic scores for these post-level features can be generated by various linguistic analysis tools (e.g., LIWC, Posit). Following that, profile-level features are weighted features whose values can be obtained from human evaluators

in order to assess the perceived credibility of a profile's information. The weighting procedure is guided by 11 different qualitative features generated from the user study, categorised into two dimensions: profile features and profile content features (Figure 7.1 illustrates profile-level features). Two characteristics of a profile evaluator are taken into account: the evaluator's prior knowledge of the profile content and to what extent the information of the profile is relevant. Section 8.1.2 proposes a conversion model to calculate the weighted credibility score of these individual profile-level features, and explains how the weighting procedure is arrived at. The next step is to feed a set of the calculated post and profile features into ML classification models in order to perform automatic predictions of information credibility and then evaluate them to identify the best performance.

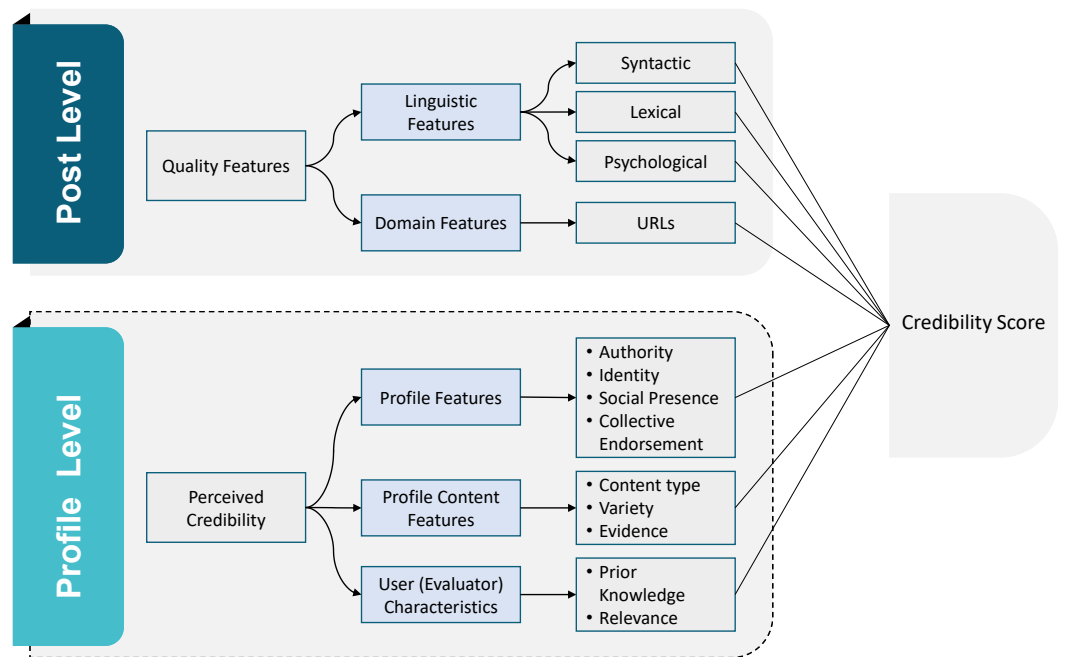


Figure 8.1: Proposed framework for automatically assessing the credibility of dementia information.



### 8.1.2 Conversion Model

As mentioned earlier, it has been established in prior research on Twitter credibility that understanding the features that end users use to determine credibility enhances automatic credibility classification by, for example, advising adjustments to feature weightings in ML approaches by deciding which feature to emphasise or minimise in crowdsourcing tasks (Morris et al., 2012). In ML, crowdsourcing entails the extensive recruitment of data annotators to annotate training data sets for supervised learning. Between two to five evaluators are usually assigned to annotate specific units of data (e.g., profile, tweet). Then, the final label for each annotation unit is estimated by combining annotations of all evaluators through simple majority voting. The final aggregated label for the annotation unit is determined by the credibility class with the most votes. Majority voting is a common approach applied in Twitter credibility studies (C. Castillo et al., 2011; Gupta & Kumaraguru, 2012) which aim to develop automatic credibility classification models. However, although simple majority voting is very effective for tasks where there is a high degree of evaluator agreement on the responses, the quality of the results cannot be ensured if there are clear differences in opinion among the evaluators (Yue, Yu, Shen, & Yu, 2014). Due to varying levels of bias, expertise and differences among individuals, disagreements are bound to arise between evaluators (Al Mansour, 2016). Therefore, a better understanding of factors that affect the crowd is important (P. Thomas, Kazai, White, & Craswell, 2022). The proposed conversion model, as explained in this section, introduces a weighting approach to quantify the qualitative features of consumers' perceived credibility that emerged from Study 3. In other words, the proposed model calculates the weight of a feature that influences evaluator assessment when labelling the same profile, while considering evaluator characteristics like prior knowledge and information relevance. The overall rating of a profile assessed by different evaluators is also proposed. Terms used in the model are defined as follows:

1. **Evaluators:** The human evaluators who evaluate the profiles are represented by  $E$  in the model. Evaluators can be lay users, volunteers, or health professionals.

Two characteristics of the evaluator, namely prior knowledge and relevance, are represented by K and R, respectively. Prior knowledge refers to the user's ability to understand and interpret the content provided by the source. Relevance refers to the interest of the user in the content provided by the source. The values for K and R range from 1 to 3, The lowest value of the two characteristics is 1 and the highest is 3. The assigned value indicates to what extent the evaluator has prior knowledge about the profile's content and how likely it is that the profile content is relevant to the evaluator. The qualitative study revealed the importance of these two characteristics as factors enabling users to assess a profile. Accordingly, higher attainment of these two attributes will amplify the overall assessment.

2. **Features:** These are the 11 different attributes related to the profile or content (as defined in Table 8.2 below). A feature is represented by A where it represents the attribute value of ith. The value of each attribute is [0,1], 0 meaning the feature has no impact on the evaluation, and 1 meaning it has an impact.
3. **Weight:** The weight given to each A with value 1 is represented by W. Different evaluators will assign a weight to each feature. Weights can be collected via a questionnaire, which can be developed using the proposed criteria in Table 8.2. A seven-point Likert scale is utilised in various credibility studies (Johnson, 2011; Gupta et al., 2014), hence, it can be applied to assign the weight value, where 1 represents the least important feature and 7 the most important. The feature weight will be amplified if higher attainment of the evaluator's prior knowledge or information relevance is shown. The feature value can then be determined by calculating the average of all evaluators' ratings (assigned weights).

The evaluation of a profile containing x features by y number of evaluators is as follows:

$$\left( (K_y + R_y) \begin{bmatrix} A_{i1} & A_{i2} & \dots & A_{ix} \\ \dots & \dots & \dots & \dots \\ A_{y1} & A_{y2} & A_{yx} & \dots \end{bmatrix} \begin{bmatrix} W_{i1} \\ W_{i2} \\ \dots \\ W_{ix} \end{bmatrix} \right) = \begin{bmatrix} E_1 \\ E_2 \\ \dots \\ E_y \end{bmatrix} \quad (8.1)$$

Consider the case of a profile rating by an evaluator having values K = 1 and R

= 1. If the values of 11 attributes assigned by the evaluator are 1,0,1,1,0,0,0,0,0,0,0 and weights of these attributes are 2,0,4,5,00,0,0,0,0,0, respectively, the profile rating by evaluator E can be computed as follows:

$$E = (1 + 1)(1 \times 2 + 0 \times 0 + 1 \times 4 + 1 \times 5 + 0 \times 0) = 22$$

The average score  $S_z$  of a feature x by all evaluators y for a given profile can be calculated as follows:

$$S_z = \frac{\sum_{i=1}^y K_i + R_i(A_{ix} \cdot W_{ix})}{y} \tag{8.2}$$

For the average score of an attribute A, (authority, for example), given by five different evaluators  $E_y$ , consider the following table:

Table 8.1: Proposed score (weight) calculation for a feature.

$K_y$	$R_y$	A	W	$E_y$
1	2	1	4	12
3	1	0	0	0
3	2	1	2	10
1	1	1	5	10
2	1	1	2	6
<b>Sum</b>				38
<b>(<math>S_z</math>)</b>				38/5 = 7.6

The credibility score assigned to profile I by evaluator  $E_1$  can be also calculated as follows:

$$E_1 = (K_1 + R_1)(A_{i1} \cdot W_{i1} + A_{i2} \cdot W_{i2} + A_{i3} \cdot W_{i3} + \dots + A_{ix} \cdot W_{ix})$$

The profile rating of a profile having x attributes by an evaluator  $E_y$  can be represented as follows:

$$E_y = (K_y + R_y) \sum_{j=1}^x A_{yj} \cdot W_{yj} \tag{8.3}$$

Table 8.2: Proposed criteria for profile credibility evaluation established in Phase 2

Features	Definitions	Criteria to evaluate profile credibility*
Authority	Subject-related experts with a scientific background and/or qualification, an official entity, a well-recognised public account, the blue badge or verified feature assigned by Twitter.	<ul style="list-style-type: none"> <li>• Does the source have authority?</li> </ul>
Identity	Location of the source is declared	<ul style="list-style-type: none"> <li>• If the profile reveals their location, is this similar to that of the evaluator?</li> </ul>
	The source declares their lived personal experience as patient or caregiver.	<ul style="list-style-type: none"> <li>• Does the profile indicate lived personal experience as patient or caregiver?</li> </ul>
Approachability	The feeling that the source is human and there is a possibility of direct interaction with the source.	<ul style="list-style-type: none"> <li>• Does the profile show the possibility of direct interaction such as contact details?</li> </ul>
Interactivity	Subject-related experts with a scientific background and/or qualification, an official entity, a well-recognised public account, the blue badge or verified feature assigned by Twitter.	<ul style="list-style-type: none"> <li>• Does the profile tweet regularly?</li> <li>• Does the profile tweet more than retweet?</li> <li>• Does the profile retweet more than tweet?</li> </ul>
Collective endorsement	Collective endorsement is measured by the number of followers of the source, who the followers are, the number of likes on the source's posts, whether the source and user have mutual friends, or even through a recommendation from a trusted person offline. The endorsement could be from one user or a group of people.	<ul style="list-style-type: none"> <li>• Does the profile have a high number of followers?</li> <li>• Does the profile have a high number of likes on their tweets?</li> <li>• Does the user have mutual friends with this profile?</li> <li>• Has anyone ever recommended this profile to the evaluator before?</li> </ul>
Content type	Pragmatic: Pragmatic information means practical ideas or tips that can be usefully applied in caregivers' everyday lives.	<ul style="list-style-type: none"> <li>• Do most source tweets provide pragmatic content?</li> </ul>
	Scientific: Medical or scientific background of presented research. Generally, such information on tweets is provided by including links to external websites or citing information on external platforms.	<ul style="list-style-type: none"> <li>• Do most source tweets provide the medical or scientific background of presented research?</li> </ul>
	Empathetic: Empathetic content contains emotional or feeling expressions, such as caring, helping, appreciation, supporting, as well as faith-based support.	<ul style="list-style-type: none"> <li>• Does the profile provide empathetic content?</li> </ul>
Variety	Variety in the content refers to topic diversity, whether focused on a single or several aspects.	<ul style="list-style-type: none"> <li>• Does the profile content focus on a single topic?</li> <li>• Does the profile content discuss several topics?</li> </ul>
Evidence (URLs)	Evidence refers to links (URLs) to other references supporting the information. Another criterion is that the information source, if available in the content, is deemed credible. Another group of participants questioned the quality and reliability of the provided references and links, believing that information based on authenticated, reliable and quality sources could be credible. Participants also mentioned that if the contents include links from the same source or website, these could not be trusted.	<ul style="list-style-type: none"> <li>• Is evidence provided for most content?</li> <li>• Do the provided references have high quality and reliability?</li> <li>• Are the references from the same domain or website?</li> </ul>

\*To be answered yes/no (if yes, a rating of 1-7 should be assigned based on the importance for assessing the profile credibility)

## 8.2 Contributions

The findings from this research make the following contributions to the SIR credibility research area.

- **Proposing a framework for evaluating dementia information credibility:** In the absence of available instruments for determining health information quality on social platforms, a crucial goal of this thesis was to develop a framework which connects two credibility aspects: information quality features evaluated by automated methods and features deemed important by consumers, through structuring them into different dimensions for automatic processing. As a result, a comprehensive framework is proposed for automatically assessing dementia information credibility on social platforms.
- **Proposing conversion model:** The research offers a thorough analysis of the qualitative features of perceived credibility of consumers and proposes a conversion model to quantify the qualitative features. The conversion model can serve to maintain the quality of the credibility labels gathered for the ground truth of information credibility on SM platforms. Moreover, it can be utilised as features for ML systems to detect health related bots. Despite recent attention to the design of customised computational systems to detect health-related bots, such as in (Davoudi et al., 2020). Research is still in its initial stages and more focus is needed to derive features for modelling the nuances that characterise health related bots only (Davoudi et al., 2020). Thus, prior understanding of factors affecting the health context will facilitate health bot detection with better accuracy.
- **Methodological contribution:** The thesis provides complementary methods to examine two different credibility aspects in the health context. The three studies that form the core of the thesis followed two different approaches: the quantitative approach was used in the first and second studies by applying different statistical and ML tests to examine the quality aspect (machine-based). The qualitative

approach was used in the third study by conducting think-aloud sessions followed by semi-structured interviews to examine the perceived credibility aspect (user-based). These mixed (complementary) methods contributed to consistency through combining statistical analysis with qualitative analysis to gain a broader perspective. Few studies in the information credibility literature have combined qualitative and quantitative methods, thus, this research makes a methodological contribution by showing the effectiveness of complementary methods.

- **Theoretical contribution:** Most existing theoretical models and frameworks for credibility assessments have been developed from the human cognitive perspective of information available through static web resources. The findings of the third study have theoretical implications for the role of technological affordances proposed in the MAIN model (Sundar, 2008) in the credibility assessment. The MAIN model proposes different agency technological affordances (e.g., authority) which can trigger cognitive heuristics that affect users' assessment of content quality (H.-S. Kim, Brubaker, & Seo, 2015). Although previous research has found support for the MAIN model in different contexts such as news (Lin et al., 2019) and e-commerce sites (H.-S. Kim et al., 2015), the current study shows how various agency affordances are identified and perceived differently in the context of health-related information on social platforms. Furthermore, the study found the role of content and user characteristics in credibility assessment to be complementary to the agency technological affordances in the MAIN model (Sundar, 2008).

### 8.3 Practical implications

Understanding the underlying technical and human perceptions of credibility has practical implications for several practitioners.

- Even though not all bots are malicious, social bots have previously been found to spread unproven health claims on Twitter. This research also provides evidence for the existence of malicious bots that contribute to spreading misleading

dementia information. Therefore, it is clear that their existence can considerably undermine health efforts and media literacy that allows information consumers to decide whom to trust. Because previous research has provided clear evidence of Twitter being used by PWD and their caregivers, information on this platform may influence the health decisions of PWD and their caregivers. Therefore, the role of policy makers in dementia and other health organisations is to design better media literacy programmes and health information content to educate users to effectively differentiate between low-quality information and high-quality information, and to recognise automated accounts (like malicious bots; those acting as people, often spreading misleading information).

- The research results will motivate SM developers to design better functionalities and technological affordances on user interfaces that influence user cognition. This will help users to improve the evaluation and consumption of health information on SM. For example, one of the findings in the user study in Chapter7 revealed instances leading to a more positive evaluation of the information provided. This happened when the possibility of source approachability and direct interaction, based on contact details found on the user profile, led to individuals experiencing a higher sense of source presence. In turn, better design of technology affordance can provide social presence and guide users in their approach to either obtain or avoid presence.
- The research findings should serve as a warning for specialists in internet policy and governance to develop ethical guidelines. Misleading information spread by bots poses a public health concern. Therefore, since bots eventually become a part of the social media sphere, there is a need to make sure that their effect is visible and that consumers are aware of who is behind them and who controls them (Ross et al., 2019).

## 8.4 Limitations and Possible Future Work

This research calls attention to several future work directions that can be conducted to overcome of the research limitations in this study:

- **Context:** As far as the study context is concerned, the focus of this research was on dementia information only. The purpose of the study was to completely comprehend and appreciate this specific phenomenon in this specific context. However, the findings of this research might not be generalisable to other health domains; yet further investigation of the proposed framework might apply to other health domains. This will help to improve the generalisability of the proposed features and determine the extent to which they can influence bot-human characterisation in a health dataset.
- **Further exploration of additional features:** Another issue is that a limited number of feature types related to the profile and content that characterise bot and humanlike behaviour were evaluated. This was mostly a limitation of Study 1 (Chapter 5), which in turn affected the design and results of Study 2 (Chapter 6). Researchers can address this limitation by further exploration of additional features to characterise bot and human-like behaviour in a health context and determine to what extent these features affect the text quality analysis.
- **Improved classification accuracy:** An important limitation is that this study used a limited set of myths related to dementia, namely only four types of false dementia information, for text quality analysis. Creating larger labelled datasets with different types of false dementia information could enhance the results and their generalisability. Future research could even extend the dataset to include other data types (e.g., images and video) to enhance the quality analysis. Also, the ML experiments conducted in Chapter 5 only tested applications of common algorithms (RF, DT, SVM, LR, and KNN) for dementia text quality analysis. Future research should apply other advanced algorithms (e.g., deep learning) that could improve ML performance.



- **Population diversity:** Participants recruited for the user study in Chapter 7 were formal and informal caregivers from the UK. Consequently, the findings are not representative of all Twitter users seeking dementia information. Also, the potential influence of the participants' location and culture may limit the generalisability of the findings, as these factors could influence people's perceptions. Therefore, future studies should replicate the user study to examine a greater variety of participants, even involving the general public whose interest in health and wellbeing, including dementia, should also be considered. This way, differences in credibility evaluations between different user groups can be compared. A broader understanding of how culture might factor into people's perceptions would be valuable as well. Moreover, the user study can be replicated with the use of different study procedures.
- **Measuring the actual impact and influence:** The research provides evidence for the existence of low-quality dementia information and proposes a framework to cope with it automatically. However, the actual impact of this low-quality information remains unknown. Interesting results can be revealed, especially for health policy makers, to understand if there is a clear implication associated with offline behaviours or attitude caused by exposure to such information.

### Closing Remarks

Given the fact that the study of SIR credibility is evolving, this research contributes to the area of health information credibility. It aims to create a better understanding of the information quality features and credibility judgement criteria users apply when viewing health information on SM, in the context of a little discussed health issue, namely dementia. As a result, a framework is proposed to automatically assess the credibility of dementia information, by suggesting appropriate features to be used with ML techniques.

Ultimately, the researcher's endeavours will lead to not only the rapid mitigation of low-quality information on SM, but also to the improvement of quality of life for PWD and

their carers. The outcomes of the research could also allow for collaboration between health professionals, dementia organisations and data researchers to, for example, identify misleading information as well as creating awareness, especially among the younger generation that has dismissive attitudes and misconceptions regarding dementia (Farina, Hughes, Griffiths, & Parveen, 2020). Furthermore, one of the research outcomes (study 3) revealed alignment with health information technologies regarding dementia and brain health research themes targeted by the Scottish Dementia Research Consortium (SDRC, 2022).

## References

- Abbasi, R. A., Maqbool, O., Mushtaq, M., Aljohani, N. R., Daud, A., Alowibdi, J. S., & Shahzad, B. (2018). Saving lives using social media: analysis of the role of twitter for personal blood donation requests and dissemination. *Telematics and Informatics*, *35*(4), 892–912.
- Abdelminaam, D. S., Ismail, F. H., Taha, M., Taha, A., Houssein, E. H., & Nabil, A. (2021). Coaid-deep: An optimized intelligent framework for automated detecting covid-19 misleading information on twitter. *IEEE Access*, *9*, 27840–27867.
- Abuqaddom, I., Alazzam, H., Hudaib, A., & Al-zaghoul, F. (2019). A measurable website usability model: Case study university of jordan. In *2019 10th international conference on information and communication systems (icics)* (pp. 83–87).
- Addawood, A., Balakumar, P., & Diesner, J. (2019). Categorization and comparison of influential twitter users and sources referenced in tweets for two health-related topics. In *International conference on information* (pp. 639–646).
- Afsana, F., Kabir, M. A., Hassan, N., & Paul, M. (2020). Towards domain-specific characterization of misinformation. *arXiv preprint arXiv:2007.14806*.
- Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 183–194).
- Albadi, N., Kurdi, M., & Mishra, S. (2019). Hateful people or hateful bots? detection and characterization of bots spreading religious hatred in arabic social media. *Proceedings of the ACM on Human-computer Interaction*, *3*(CSCW), 1–25.
- Al-bahrani, R., Danilovich, M. K., Liao, W.-K., Choudhary, A., & Agrawal, A. (2017). Analyzing informal caregiving expression in social media. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 342–349).
- Aleroud, A., Abu-Alsheeh, N., & Al-shawakfa, E. (2020). A graph proximity feature augmentation approach for identifying accounts of terrorists on twitter. *Computers & Security*, *99*, 102056.
- Alhadreti, O., & Mayhew, P. (2018). Rethinking thinking aloud: A comparison of

## References

- three think-aloud protocols. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–12).
- Alhayan, F., & Pennington, D. (2020). Twitter as health information source: Exploring the parameters affecting dementia-related tweets. In *International conference on social media and society* (pp. 277–290).
- Alhayan, F., Pennington, D. R., & Ruthven, I. (2022). “she seems more human”: Understanding twitter users’ credibility assessments of dementia-related information. In *International conference on information* (pp. 292–313).
- Al-Jefri, M. (2019). *Automatic evaluation of online health information quality* (doctoral dissertation). University of Brighton.
- Al-jefri, M. M., Evans, R., Ghezzi, P., & Uchyigit, G. (2017). Using machine learning for automatic identification of evidence-based health information on the web. In *Proceedings of the 2017 international conference on digital health* (pp. 167–174).
- Allem, J.-P., Ferrara, E., Uppu, S. P., Cruz, T. B., Unger, J. B., et al. (2017). E-cigarette surveillance with social media data: Social bots, emerging topics, and trends. *JMIR public health and surveillance*, 3(4), e8641.
- Al Mansour, A. A. (2016). Labeling agreement level and classification accuracy. In *2016 12th international conference on signal-image technology & internet-based systems (sitis)* (pp. 271–274).
- Alqurashi, S., Hamoui, B., Alashaikh, A., Alhindi, A., & Alanazi, E. (2021). Eating garlic prevents covid-19 infection: Detecting misinformation on the arabic content of twitter. *arXiv preprint arXiv:2101.05626*.
- Al-rakhami, M. S., & Al-amri, A. M. (2020). Lies kill, facts save: detecting covid-19 misinformation in twitter. *Ieee Access*, 8, 155961–155970.
- Al-rawi, A. (2019). Twitter influentials and the networked publics’ engagement with the rohingya crisis in arabic and english. *The SAGE handbook of media and migration*, 192–204.
- Alrubaian, M., Al-qurishi, M., Alamri, A., Al-rakhami, M., Hassan, M. M., & Fortino, G. (2018). Credibility in online social networks: A survey. *IEEE Access*, 7, 2828–2855.

## References

- Alrubaian, M., Al-qurishi, M., Al-rakhami, M., Rahman, S. M. M., & Alamri, A. (2015). A multistage credibility analysis model for microblogs. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 1434–1440).
- Alsmadi, I., & O'Brien, M. J. (2020). How many bots in Russian troll tweets? *Information Processing & Management*, *57*(6), 102303.
- Alsudias, L., & Rayson, P. (2020). Covid-19 and Arabic Twitter: How can Arab world governments and public health organizations learn from social media? In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.
- Ammu, P., & Preeja, V. (2013). Review on feature selection techniques of DNA microarray data. *International Journal of Computer Applications*, *61*(12).
- Anderson, J. G., Hundt, E., Dean, M., Keim-Malpass, J., & Lopez, R. P. (2017). “the church of online support” examining the use of blogs among family caregivers of persons with dementia. *Journal of Family Nursing*, *23*(1), 34–54.
- Astell, A., Dove, E., & Hernandez, A. (2019). An introduction to technology for dementia. *Using technology in dementia care: A guide to technology solutions for everyday living*, 11.
- Attai, D. J., Anderson, P. F., Fisch, M. J., Graham, D. L., Katz, M. S., Kesselheim, J., ... Dizon, D. S. (2017). Risks and benefits of Twitter use by hematologist-oncologists in the era of digital medicine. In *Seminars in Hematology* (Vol. 54, pp. 198–204).
- Ayalon, L. (2013). Re-examining ethnic differences in concerns, knowledge, and beliefs about Alzheimer's disease: Results from a national sample. *International Journal of Geriatric Psychiatry*, *28*(12), 1288–1295.
- Bachmann, P. (2020). Caregivers' experience of caring for a family member with Alzheimer's disease: A content analysis of longitudinal social media communication. *International Journal of Environmental Research and Public Health*, *17*(12), 4412.
- Bastos, M., Walker, S., & Simeone, M. (2021). The Imped Model: Detecting low-quality information in social media. *American Behavioral Scientist*, *65*(6), 863–883.

## References

- Basu, M., Ghosh, S., & Ghosh, K. (2018). Overview of the fire 2018 track: Information retrieval from microblogs during disasters (irmidis). In *Proceedings of the 10th annual meeting of the forum for information retrieval evaluation* (pp. 1–5).
- Baxter, G., Marcella, R., & Walicka, A. (2019). Scottish citizens’ perceptions of the credibility of online political “facts” in the “fake news” era: An exploratory study. *Journal of documentation*.
- Berrar, D. (2019). *Cross-validation*. (Vol. 1). Elsevier.
- Beskow, D. M., & Carley, K. M. (2018). Bot-hunter: a tiered approach to detecting & characterizing automated activity on twitter. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation* (Vol. 3, p. 3).
- Beskow, D. M., & Carley, K. M. (2020). You are known by your friends: Leveraging network metrics for bot detection in twitter. In *Open source intelligence and cyber crime* (pp. 53–88). Springer.
- Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion. *First monday, 21*(11-7).
- Bhattacharjee, U., Srijith, P., & Desarkar, M. S. (2019). Leveraging social media towards understanding anti-vaccination campaigns. In *2019 11th international conference on communication systems & networks (comsnets)* (pp. 886–890).
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993–1022.
- Block, B. R., Albanese, S. G., & Hume, A. L. (2021). Online promotion of “brain health” supplements. *The Senior Care Pharmacist, 36*(10), 489–492.
- Bouadjenek, M. R., Hacid, H., & Bouzeghoub, M. (2016). Social networks and information retrieval, how are they converging? a survey, a taxonomy and an analysis of social information retrieval approaches and platforms. *Information Systems, 56*, 1–18.
- Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one, 12*(6), e0177678.
- Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and

## References

- scholarship. *Journal of computer-mediated Communication*, 13(1), 210–230.
- Boyer, C., Selby, M., Scherrer, J.-R., & Appel, R. (1998). The health on the net code of conduct for medical and health websites. *Computers in biology and medicine*, 28(5), 603–610.
- Bracamonte, V., & Okada, H. (2015). Impact of nationality information in feedback on trust in a foreign online store. *Journal of Socio-Informatics*, 8(1), 1–12.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Brennen, J. S., Simon, F. M., Howard, P. N., & Nielsen, R. K. (2020). *Types, sources, and claims of covid-19 misinformation*. Reuters Institute for the Study of Journalism. Retrieved from <https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation>.
- Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., ... Dredze, M. (2018). Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American journal of public health*, 108(10), 1378–1384.
- Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative research*, 6(1), 97–113.
- Buhmann, M. D. (2000). Radial basis functions. *Acta numerica*, 9, 1–38.
- Burfoot, C., & Baldwin, T. (2009). Automatic satire detection: Are you having a laugh? In *Proceedings of the acl-ijcnlp 2009 conference short papers* (pp. 161–164).
- Byrd, K., Mansurov, A., & Baysal, O. (2016). Mining twitter data for influenza detection and surveillance. In *Proceedings of the international workshop on software engineering in healthcare systems* (pp. 43–49).
- Caracelli, V. J., & Greene, J. C. (1993). Data analysis strategies for mixed-method evaluation designs. *Educational evaluation and policy analysis*, 15(2), 195–207.
- Cartwright, B., Weir, G. R., & Frank, R. (2019). Fighting disinformation warfare with artificial intelligence. *CLOUD COMPUTING 2019*, 83.
- Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 675–684).
- Castillo, L. I., Hadjistavropoulos, T., & Brachaniec, M. (2021). The effectiveness of

## References

- social media in the dissemination of knowledge about pain in dementia. *Pain Medicine*, *22*(11), 2584–2596.
- Cations, M., Radisic, G., Crotty, M., & Laver, K. E. (2018). What does the general public understand about prevention and treatment of dementia? a systematic review of population-based surveys. *PloS one*, *13*(4), e0196085.
- Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology*, *39*(5), 752.
- Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.
- Charernboon, T., Lerthattasilp, T., & Supasitthumrong, T. (2021). Effectiveness of cannabinoids for treatment of dementia: A systematic review of randomized controlled trials. *Clinical Gerontologist*, *44*(1), 16–24.
- Chavoshi, N., Hamooni, H., & Mueen, A. (2016). Debot: Twitter bot detection via warped correlation. In *Icdm* (pp. 817–822).
- Chen, L., Wang, X., Peng, T.-Q., et al. (2018). Nature and diffusion of gynecologic cancer-related misinformation on social media: Analysis of tweets. *Journal of Medical Internet Research*, *20*(10), e11515.
- Cherak, S. J., Rosgen, B. K., Amarbayan, M., Plotnikoff, K., Wollny, K., Stelfox, H. T., & Fiest, K. M. (2020). Impact of social media interventions and tools among informal caregivers of critically ill patients after patient admission to the intensive care unit: A scoping review. *PloS one*, *15*(9), e0238803.
- Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 1–13.
- Choi, W. (2015). New framework of web credibility assessment and an exploratory study of older adults' information behavior on the web.
- Choi, W., & Stvilia, B. (2015). Web credibility assessment: Conceptualization, operationalization, variability, and models. *Journal of the Association for Information Science and Technology*, *66*(12), 2399–2414.



## References

- Choi, W., & Stvilia, B. (2022). Developing a theoretical framework for web credibility assessment—a case of social q&a sites: preliminary findings.
- Chorley, M. J., Colombo, G. B., Allen, S. M., & Whitaker, R. M. (2015). Human content filtering in twitter: The influence of metadata. *International Journal of Human-Computer Studies*, *74*, 32–40.
- Chou, W.-Y. S., Oh, A., & Klein, W. M. (2018). Addressing health-related misinformation on social media. *Jama*, *320*(23), 2417–2418.
- Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on dependable and secure computing*, *9*(6), 811–824.
- Cisneros-Velarde, P., Oliveira, D. F., & Chan, K. S. (2019). Spread and control of misinformation with heterogeneous agents. In *International workshop on complex networks* (pp. 75–83).
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.
- Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, *52*(1), 1–4.
- Cord, M., & Cunningham, P. (2008). *Machine learning techniques for multimedia: Case studies on organization and retrieval*. Springer Science & Business Media.
- Craig, D., & Strivens, E. (2016). Facing the times: A young onset dementia support group: Facebooktm style. *Australasian Journal on Ageing*, *35*(1), 48–53.
- Crawford, J. L., Guo, C., Schroeder, J., Arriaga, R. I., & Mankoff, J. (2014). Is it a question of trust? how search preferences influence forum use. In *Proceedings of the 8th international conference on pervasive computing technologies for healthcare* (pp. 118–125).
- Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion* (pp. 963–972).

## References

- Cresci, S., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2019). Cashtag piggy-backing: Uncovering spam and bot activity in stock microblogs on twitter. *ACM Transactions on the Web (TWEB)*, 13(2), 1–27.
- Creswell, J. (2003). *Research design*. Sage publications Thousand Oaks, CA.
- Creswell, J. (2014). *Research design—qualitative, quantitative & mixed methods approaches.(4: e upplagan)* sage publications. Inc.
- Creswell, J., & Clark, V. L. P. (2011). *Designing and conducting mixed methods research. 2 ed..*
- Dalmer, N. K. (2017). Questioning reliability assessments of health information on social media. *Journal of the Medical Library Association: JMLA*, 105(1), 61.
- Danielson, D. R., & Rieh, S. Y. (2007). Credibility: A multidisciplinary framework. *Annual review of information science and technology*, 41, 307–364.
- Danilovich, M. K., Tsay, J., Al-bahrani, R., Choudhary, A., & Agrawal, A. (2018). #alzheimer’s and dementia. *Topics in Geriatric Rehabilitation*, 34(1), 48–53.
- Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web* (pp. 273–274).
- Davoudi, A., Klein, A. Z., Sarker, A., & Gonzalez-Hernandez, G. (2020). Towards automatic bot detection in twitter for health-related tasks. *AMIA Summits on Translational Science Proceedings, 2020*, 136.
- De Choudhury, M., Morris, M. R., & White, R. W. (2014). Seeking and sharing health information online: comparing search engines and social media. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1365–1376).
- Del Pilar Salas-Zárate, M., Paredes-Valverde, M. A., Rodríguez-García, M. Á., Valencia-García, R., & Alor-Hernández, G. (2017). Automatic detection of satire in twitter: A psycholinguistic-based approach. *Knowledge-Based Systems*, 128, 20–33.
- de Moraes, M. B., & Gradvohl, A. L. S. (2021). A comparative study of feature selection methods for binary text streams classification. *Evolving Systems*, 12(4), 997–1013.

## References

- Dhanalakshmi, V., Kumar, A., Shivapratap, G., Soman, K., & Rajendran, S. (2009). Tamil pos tagging using linear programming. *International Journal of Recent Trends in Engineering*, 1(2), 166.
- Dizon, D. S., Graham, D., Thompson, M. A., Johnson, L. J., Johnston, C., Fisch, M. J., & Miller, R. (2012). Practical guidance: The use of social media in oncology practice. *Journal of oncology practice*, 8(5), e114–e124.
- Du, W., & Zhan, Z. (2002). Building decision tree classifier on private data. *Proc. IEEE International Conference on Privacy, Security, and Data Mining*, 1–8.
- Edwards, C., Edwards, A., Spence, P. R., & Shelton, A. K. (2014). Is that a bot running the social media feed? testing the differences in perceptions of communication quality for a human agent and a bot agent on twitter. *Computers in Human Behavior*, 33, 372–376.
- Edwards, C., Spence, P. R., Gentile, C. J., Edwards, A., & Edwards, A. (2013). How much klout do you have... a test of system generated cues on source credibility. *Computers in Human Behavior*, 29(5), A12–A16.
- Edwards, C., Stoll, B., Faculak, N., & Karman, S. (2015). Social presence on linkedin: Perceived credibility and interpersonal attractiveness based on user profile picture. *Online Journal of Communication and Media Technologies*, 5(4), 102.
- Egan, K. J., Clark, P., Deen, Z., Dutu, C. P., Wilson, G., McCann, L., ... Maguire, R. (2022). Understanding current needs and future expectations of informal caregivers for technology to support health and well-being: National survey study. *JMIR aging*, 5(1), e15413.
- Egan, K. J., Pinto-Bruno, Á. C., Bighelli, I., Berg-Weger, M., van Straten, A., Albanese, E., & Pot, A.-M. (2018). Online training and support programs designed to improve mental health and reduce burden among caregivers of people with dementia: a systematic review. *Journal of the American Medical Directors Association*, 19(3), 200–206.
- Elhadad, M. K., Li, K. F., & Gebali, F. (2020). Covid-19-fakes: A twitter (arabic/english) dataset for detecting misleading information on covid-19. In *International conference on intelligent networking and collaborative systems* (pp.

## References

- 256–268).
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological review*, *87*(3), 215.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. the MIT Press.
- Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Thomson Brooks/Cole Publishing Co.
- Eysenbach, G., & Köhler, C. (2002). How do consumers search for and appraise health information on the world wide web? qualitative study using focus groups, usability tests, and in-depth interviews. *Bmj*, *324*(7337), 573–577.
- Farina, N., Hughes, L. J., Griffiths, A. W., & Parveen, S. (2020). Adolescents' experiences and perceptions of dementia. *Aging & Mental Health*, *24*(7), 1175–1181.
- Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 french presidential election. *arXiv preprint arXiv:1707.00086*.
- Ferrara, E., Varol, O., Menczer, F., & Flammini, A. (2016). Detection of promoted social media campaigns. In *Proceedings of the international aaai conference on web and social media* (Vol. 10, pp. 563–566).
- Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in statistics* (pp. 66–70). Springer.
- Flanagin, A. J., & Metzger, M. J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New media & society*, *9*(2), 319–342.
- Flanagin, A. J., & Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, *72*(3), 137–148.
- Fogg, B. (2002). Prominence-interpretation theory: Explaining how people assess credibility. *CHI 2003: NEW HORIZONS*, 722–723.
- Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 80–87).
- Forman, J., & Damschroder, L. (2007). Qualitative content analysis. In *Empirical methods for bioethics: A primer*. Emerald Group Publishing Limited.

## References

- Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, *15*(2), 150–157.
- French, T. (2016). *Dementia and digital: Using technology to improve health and well-being for people with dementia and their carers*. United Kingdom: Tinder Foundation. Retrieved from [https://www.housinglin.org.uk/\\_assets/Resources/Housing/OtherOrganisation/dementia\\_and\\_digital.pdf](https://www.housinglin.org.uk/_assets/Resources/Housing/OtherOrganisation/dementia_and_digital.pdf)
- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367–378.
- Froehlich, T. J. (2019). The role of pseudo-cognitive authorities and self-deception in the dissemination of fake news. *Open Information Science*, *3*(1), 115–136.
- Fu, K.-W., Liang, H., Saroha, N., Tse, Z. T. H., Ip, P., & Fung, I. C.-H. (2016). How people react to zika virus outbreaks on twitter? a computational content analysis. *American journal of infection control*, *44*(12), 1700–1702.
- Gallwitz, F., & Kreil, M. (2021). The rise and fall of ‘social bot’research. *SSRN: https://ssrn.com/abstract, 3814191*.
- Gärtner, T., Flach, P., & Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. In *Learning theory and kernel machines* (pp. 129–143). Springer.
- Ghenai, A., & Mejova, Y. (2017). Catching zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on twitter. *arXiv preprint arXiv:1707.03778*. Retrieved from <https://arxiv.org/abs/1707.03778>
- Ghenai, A., & Mejova, Y. (2018). Fake cures: User-centric modeling of health misinformation in social media. *Proceedings of the ACM on human-computer interaction*, *2*(CSCW), 1–20.
- Ghenai, A., Smucker, M. D., & Clarke, C. L. (2020). A think-aloud study to understand factors affecting online health search. In *Proceedings of the 2020 conference on human information interaction and retrieval* (pp. 273–282).
- Giachanou, A., Ghanem, B., & Rosso, P. (2021). Detection of conspiracy propa-

## References

- gators using psycho-linguistic characteristics. *Journal of Information Science*, 0165551520985486.
- Gibbs, G. R. (2018). *Analyzing qualitative data* (Vol. 6). Sage.
- Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., & Crowcroft, J. (2017). Of bots and humans (on twitter). In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 349–354).
- Ginsca, A. L., Popescu, A., & Lupu, M. (2015). Credibility in information retrieval. *Foundations and Trends in Information Retrieval*, 9(5), 355–475.
- Goh, D., & Foo, S. (2007). *Social information retrieval systems: Emerging technologies and applications for searching the web effectively: Emerging technologies and applications for searching the web effectively*. IGI Global.
- Golberg, G. (2017). *When it comes to dealing with fake/bot accounts, twitter is (still) failing*. Retrieved from <https://geoffgolberg.medium.com/when-it-comes-to-dealing-with-fake-bot-accounts-twitter-is-still-failing-a79d9ece5b5d> (Accessed: Dec-2021)
- Gonçalves, R., & Dorneles, C. F. (2019). Automated expertise retrieval: A taxonomy-based survey and open issues. *ACM Computing Surveys (CSUR)*, 52(5), 1–30.
- Graham, T., Bruns, A., Zhu, G., & Campbell, R. (2020). Like a virus: The coordinated spread of coronavirus disinformation.
- Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational evaluation and policy analysis*, 11(3), 255–274.
- Gruzd, A., & Mai, P. (2020). Going viral: How a single tweet spawned a covid-19 conspiracy theory on twitter. *Big Data & Society*, 7(2), 2053951720938405.
- Guillory, J. J., & Geraci, L. (2013). Correcting erroneous inferences in memory: The role of source credibility. *Journal of Applied Research in Memory and Cognition*, 2(4), 201–209.
- Gupta, A., & Katarya, R. (2020). Social media based surveillance systems for healthcare using machine learning: A systematic review. *Journal of Biomedical Informatics*, 108, 103500.

## References

- Gupta, A., & Kumaraguru, P. (2012). Credibility ranking of tweets during high impact events. In *Proceedings of the 1st workshop on privacy and security in online social media* (pp. 2–8).
- Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014). Tweetcred: Real-time credibility assessment of content on twitter. In *International conference on social informatics* (pp. 228–243).
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157–1182.
- Hamad, E. O., Savundranayagam, M. Y., Holmes, J. D., Kinsella, E. A., & Johnson, A. M. (2016). Toward a mixed-methods research approach to content analysis in the digital age: The combined content-analysis model and its applications to health care twitter feeds. *Journal of medical Internet research*, 18(3), e5391.
- Hamidian, S., & Diab, M. (2016). Rumor identification and belief investigation on twitter. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 3–8).
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.
- Haouari, F., Hasanain, M., Suwaileh, R., & Elsayed, T. (2020). Arcov19-rumors: Arabic covid-19 twitter dataset for misinformation detection. *arXiv preprint arXiv:2010.08768*.
- Harris, R. (1997). Evaluating internet research sources. *Virtual salt*, 17(1), 1–17.
- Heigham, J., & Croker, R. (2009). *Qualitative research in applied linguistics: A practical introduction*. Springer.
- Hilligoss, B., & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *information processing & management*, 44(4), 1467–1484.
- hootsuite. (2020). *Digital 2020: Social media use spans almost half global population*. Retrieved from <https://www.hootsuite.com/newsroom/press-releases/digital-2020-social-media-use-spans-almost-half-global-population>
- Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion*.
- Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis.

## References

- Qualitative health research*, 15(9), 1277–1288.
- Huang, S. H. (2015). Supervised feature selection: A tutorial. *Artif. Intell. Res.*, 4(2), 22–37.
- Idris, I. (2016). *Python data analysis cookbook*. Packt Publishing Ltd.
- Ivanov, K. (1972). *Quality-control of information: On the concept of accuracy of information in data-banks and in management information systems* (doctoral dissertation). KTH Royal Institute of Technology.
- Jahng, M. R., & Littau, J. (2016). Interacting is believing: Interactivity, social cue, and perceptions of journalistic credibility on twitter. *Journalism & Mass Communication Quarterly*, 93(1), 38–58.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.
- Jamison, A. M., Broniatowski, D. A., & Quinn, S. C. (2019). Malicious actors on twitter: A guide for public health researchers. *American journal of public health*, 109(5), 688–692.
- Jelavić, Ž., Klemar, K. L., & Sušić, Ž. (2018). A museum programme intended for people with alzheimer’s disease and dementia. *ICOM Education 28*.
- Jeon, G. Y., & Rieh, S. Y. (2014). Answers from the crowd: how credible are strangers in social q&a? *IConference 2014 Proceedings*.
- Ji, X., Chun, S., Wei, Z., & Geller, J. (2015). Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining*, 5(1), 1–25.
- Jiang, M., Cui, P., Beutel, A., Faloutsos, C., & Yang, S. (2014). Inferring strange behavior from connectivity pattern in social networks. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 126–138).
- Jo, Y., Kim, M., & Han, K. (2019). How do humans assess the credibility on web blogs: Qualifying and verifying human factors with machine learning. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–12).
- Johnson, K. A. (2011). The effect of twitter posts on students’ perceptions of instructor credibility. *Learning, Media and Technology*, 36(1), 21–38.



## References

- Jucks, R., & Thon, F. M. (2017). Better to have many opinions than one from an expert? social validation by one trustworthy source versus the masses in online health forums. *Computers in Human Behavior*, *70*, 375–381.
- Julien, H. (2008). Content analysis. *The SAGE encyclopedia of qualitative research methods*, *1*, 120–121.
- Julien, H., & Barker, S. (2009). How high-school students find and evaluate scientific information: A basis for information literacy skills development. *Library & Information Science Research*, *31*(1), 12–17.
- Juran, J. M., et al. (1992). *Juran on quality by design: The new steps for planning quality into goods and services*. Simon and Schuster.
- Kammerer, Y., Bråten, I., Gerjets, P., & Strømsø, H. I. (2013). The role of internet-specific epistemic beliefs in laypersons' source evaluations and decisions during web search on a medical issue. *Computers in human behavior*, *29*(3), 1193–1203.
- Kang, B., Höllerer, T., & O'Donovan, J. (2015). Believe it or not? analyzing information credibility in microblogs. In *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015* (pp. 611–616).
- Kantepe, M., & Ganiz, M. C. (2017). Preprocessing framework for twitter bot detection. In *2017 international conference on computer science and engineering (ubmk)* (pp. 630–634).
- Karagöz, A. F. G. P. (2016). Credibility analysis for tweets written in turkish by a hybrid method. *Feature Engineering in Hybrid Recommender Systems*, 55.
- Kattenbeck, M., & Elsweler, D. (2019). Understanding credibility judgements for web search snippets. *Aslib Journal of Information Management*.
- Keller, T. R. (2020). To whom do politicians talk and listen?: Mapping swiss politicians' public sphere on twitter. *Computational Communication Research*, *2*(2), 175–202.
- Keller, T. R., & Klinger, U. (2019). Social bots in election campaigns: Theoretical, empirical, and methodological implications. *Political Communication*, *36*(1), 171–189.

## References

- Keshavarz, H. (2020). Evaluating credibility of social media information: Current challenges, research directions and practical criteria. *Information Discovery and Delivery*.
- Kim, H.-S., Brubaker, P., & Seo, K. (2015). Examining psychological effects of source cues and social plugins on a product review website. *Computers in Human Behavior*, *49*, 74–85.
- Kim, K. J., Park, E., & Sundar, S. S. (2013). Caregiving role in human–robot interaction: A study of the mediating effects of perceived benefit and social presence. *Computers in Human Behavior*, *29*(4), 1799–1806.
- Kirsch, S. M., Gnasa, M., & Cremers, A. B. (2006). Beyond the web: Retrieval in social information spaces. In *European conference on information retrieval* (pp. 84–95).
- Kirsch, S. M., et al. (2005). Social information retrieval. *These de Doctorat. Université de Rheinische Friedrich-Wilhelms*.
- Klawitter, E., & Hargittai, E. (2018). Shortcuts to well being? evaluating the credibility of online health information through multiple complementary heuristics. *Journal of Broadcasting & Electronic Media*, *62*(2), 251–268.
- Kramer, O. (2016). Scikit-learn. In *Machine learning for evolution strategies* (pp. 45–53). Springer.
- Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.
- Kumar, V., & Pedanekar, N. (2016). Mining shapes of expertise in online social q&a communities. In *Proceedings of the 19th acm conference on computer supported cooperative work and social computing companion* (pp. 317–320).
- Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American journal of psychology*, *113*(3), 387–404.
- Kwon, S., Cha, M., & Jung, K. (2017). Rumor detection over varying time windows. *PloS one*, *12*(1), e0168344.
- Laaksonen, C., Jalonen, H., & Paavola, J. (2014). Utilising social media for intervening and predicting future health in societies. In *International conference on well-being*

## References

- in the information society* (pp. 100–108).
- Lahuerta-Otero, E., & Cordero-Gutiérrez, R. (2016). Looking for the perfect tweet. the use of data mining techniques to find influencers on twitter. *Computers in Human Behavior*, *64*, 575–583.
- Lankes, R. D. (2007). *Trusting the internet: New approaches to credibility tools*. MacArthur Foundation Digital Media and Learning Initiative.
- Lee, E.-J., & Shin, S. Y. (2012). Are they talking to me? cognitive and affective effects of interactivity in politicians' twitter communication. *Cyberpsychology, Behavior, and Social Networking*, *15*(10), 515–520.
- Lee, J. S. (2016). Citizens' political information behaviors during elections on twitter in south korea: Information worlds of opinion leaders.
- Lee, K., Eoff, B., & Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on twitter. In *Proceedings of the international aaai conference on web and social media* (Vol. 5, pp. 185–192).
- Lee, K. M. (2004). Presence, explicated. *Communication theory*, *14*(1), 27–50.
- Li, Y., Zhang, X., & Wang, S. (2017). Fake vs. real health information in social media in china. *Proceedings of the Association for Information Science and Technology*, *54*(1), 742–743.
- Li, Y.-J., Cheung, C. M., Shen, X.-L., & Lee, M. K. (2019). Health misinformation on social media: A literature review. In *23rd pacific asia conference on information systems (pacific 2019): Secure ict platform for the 4th industrial revolution*.
- Liao, Q. V., & Fu, W.-T. (2014). Age differences in credibility judgments of on-line health information. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *21*(1), 1–23.
- Lim, Y.-s., & Lee-Won, R. J. (2017). When retweets persuade: The persuasive effects of dialogic retweeting and the role of social presence in organizations' twitter-based communication. *Telematics and informatics*, *34*(5), 422–433.
- Lin, X., Kaufmann, R., Spence, P. R., & Lachlan, K. A. (2019). Agency cues in online comments: Exploring their relationship with anonymity and frequency of helpful posts. *Southern Communication Journal*, *84*(3), 183–195.

## References

- Lin, X., & Spence, P. R. (2018). Identity on social networks as a cue: Identity, retweets, and credibility. *Communication Studies*, *69*(5), 461–482.
- Lin, X., Spence, P. R., & Lachlan, K. A. (2016). Social media and credibility indicators: The effect of influence cues. *Computers in human behavior*, *63*, 264–271.
- Lioma, C., Simonsen, J. G., & Larsen, B. (2017). Evaluation measures for relevance and credibility in ranked lists. In *Proceedings of the acm sigir international conference on theory of information retrieval* (pp. 91–98).
- Liu, J.-C., Hsu, Y.-P., Kao, P.-F., Hao, W.-R., Liu, S.-H., Lin, C.-F., ... Wu, S.-Y. (2016). Influenza vaccination reduces dementia risk in chronic kidney disease patients: A population-based cohort study. *Medicine*, *95*(9).
- Liu, Y. (2016). *Mining social media to understand consumers' health concerns and the public's opinion on controversial health topics*. (doctoral dissertation). University of Michigan.
- Low, L.-F., & Anstey, K. J. (2007). The public's perception of the plausibility of dementia risk factors is not influenced by scientific evidence. *Dementia and geriatric cognitive disorders*, *23*(3), 202–206.
- Lu, B., Fan, W., & Zhou, M. (2016). Social presence, trust, and social commerce purchase intention: An empirical research. *Computers in Human behavior*, *56*, 225–237.
- Lucassen, T., Muilwijk, R., Noordzij, M. L., & Schraagen, J. M. (2013). Topic familiarity and information skills in online credibility evaluation. *Journal of the American Society for Information Science and Technology*, *64*(2), 254–264.
- Lucassen, T., & Schraagen, J. M. (2011). Factual accuracy and trust in information: The role of expertise. *Journal of the American Society for Information Science and Technology*, *62*(7), 1232–1242.
- Lyu, J. C., Le Han, E., & Luli, G. K. (2021). Covid-19 vaccine-related discussion on twitter: topic modeling and sentiment analysis. *Journal of medical Internet research*, *23*(6), e24435.
- Ma, J., Gao, W., Wei, Z., Lu, Y., & Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th*

## References

- acm international on conference on information and knowledge management* (pp. 1751–1754).
- Ma, J., Gao, W., & Wong, K.-F. (2018). Rumor detection on twitter with tree-structured recursive neural networks..
- Ma, T. J., & Atkin, D. (2017). User generated content and credibility evaluation of online health information: A meta analytic study. *Telematics and Informatics*, *34*(5), 472–486.
- Mackie, B. R., Mitchell, M., & Marshall, A. P. (2019). Patient and family members' perceptions of family participation in care on acute care wards. *Scandinavian journal of caring sciences*, *33*(2), 359–370.
- Martínez-Pérez, B., de la Torre-Díez, I., Bargiela-Flórez, B., López-Coronado, M., & Rodrigues, J. J. (2015). Content analysis of neurodegenerative and mental diseases social groups. *Health informatics journal*, *21*(4), 267–283.
- Mazza, M., Cresci, S., Avvenuti, M., Quattrocioni, W., & Tesconi, M. (2019). Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th acm conference on web science* (pp. 183–192).
- McCorriston, J., Jurgens, D., & Ruths, D. (2015). Organizations are users too: Characterizing and detecting the presence of organizations on twitter. In *Proceedings of the international aaai conference on web and social media* (Vol. 9, pp. 650–653).
- McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement. *Communications Monographs*, *66*(1), 90–103.
- McGinnies, E., & Ward, C. D. (1980). Better liked than right: Trustworthiness and expertise as factors in credibility. *Personality and Social Psychology Bulletin*, *6*(3), 467–472.
- Mei, Y., Zhong, Y., & Yang, J. (2015). Finding and analyzing principal features for measuring user influence on twitter. In *2015 ieee first international conference on big data computing service and applications* (pp. 478–486).
- Meinert, J., Aker, A., & Krämer, N. (2019). The impact of twitter features on credibility ratings-an explorative examination combining psychological measurements and feature based selection methods. In *Proceedings of the 52nd hawaii international*

## References

- conference on system sciences.*
- Metag, J. (2016). Content analysis in climate change communication. In *Oxford research encyclopedia of climate science.*
- Metzger, M. J. (2007). Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American society for information science and technology*, 58(13), 2078–2091.
- Metzger, M. J., & Flanagin, A. J. (2015). Psychological approaches to credibility assessment online. *The handbook of the psychology of communication technology*, 445–466.
- Metzler, D., & Bruce Croft, W. (2007). Linear feature-based models for information retrieval. *Information Retrieval*, 10(3), 257–274.
- Mierzecka, A., Wasilewski, J., & Kisilowska, M. (2019). Cognitive authority, emotions and information quality evaluations.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.
- Molina, M. D., Sundar, S. S., Le, T., & Lee, D. (2021). “fake news” is not simply false information: A concept explication and taxonomy of online content. *American behavioral scientist*, 65(2), 180–212.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2013). *Solutions manual to accompany introduction to linear regression analysis.* John Wiley & Sons.
- Morgan, D. L. (1993). Qualitative content analysis: A guide to paths not taken. *Qualitative health research*, 3(1), 112–121.
- Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012). Tweeting is believing? understanding microblog credibility perceptions. In *Proceedings of the acm 2012 conference on computer supported cooperative work* (pp. 441–450).
- Mosemghvdlishvili, L., & Jansz, J. (2013). Framing and praising allah on youtube: Exploring user-created videos about islam and the motivations for producing

## References

- them. *New Media & Society*, 15(4), 482–500.
- Muntinga, T., & Taylor, G. (2018). Information-seeking strategies in medicine queries: A clinical eye-tracking study with gaze-cued retrospective think-aloud protocol. *International Journal of Human–Computer Interaction*, 34(6), 506–518.
- Nagel, A. K., Loetscher, T., Smith, A. E., & Keage, H. A. (2021). What do the public really know about dementia and its risk factors? *Dementia*, 20(7), 2424–2440.
- Neal, D. M. (2010). The conundrum of providing authoritative online consumer health information: Current research and implications for information professionals. *Bulletin of the American Society for Information Science and Technology*, 36(4), 33–37.
- Neal, D. M., & McKenzie, P. J. (2011). Putting the pieces together: Endometriosis blogs, cognitive authority, and collaborative information behavior. *Journal of the Medical Library Association: JMLA*, 99(2), 127.
- Neuendorf, K. A. (2017). *The content analysis guidebook*. sage.
- Ng, L. H. X., & Carley, K. M. (2021). “the coronavirus is a bioweapon”: Classifying coronavirus stories on fact-checking sites. *Computational and Mathematical Organization Theory*, 27(2), 179–194.
- Oltulu, P., Mannan, A. A. S. R., & Gardner, J. M. (2018). Effective use of twitter and facebook in pathology practice. *Human pathology*, 73, 128–143.
- ONS. (2019, Aug). *Deaths registered in england and wales: 2018*. Office for National Statistics. Retrieved from <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregistrationssummarytables/2018>
- Onwuegbuzie, A. J., Johnson, B., & Turner, L. (2007). Toward a definition of mixed methods research. *Journal of mixed methods research*, 1(2), 112–133.
- Overbey, L. A., Ek, B., Pinzhoffer, K., & Williams, B. (2019). Using common enemy graphs to identify communities of coordinated social media activity. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation* (pp. 92–102).
- O’reilly, M., & Parker, N. (2013). ‘unsatisfactory saturation’: a critical exploration of

## References

- the notion of saturated sample sizes in qualitative research. *Qualitative research*, 13(2), 190–197.
- Padilla, J.-L., & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. In *Understanding and investigating response processes in validation research* (pp. 211–228). Springer.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web*. (Tech. Rep.). Stanford InfoLab.
- Park, S., Oh, H.-K., Park, G., Suh, B., Bae, W. K., Kim, J. W., . . . Kang, S.-B. (2016). The source and credibility of colorectal cancer information on twitter. *Medicine*, 95(7).
- Pasi, G., & Viviani, M. (2020). Information credibility in the social web: Contexts, approaches, and open issues. *arXiv preprint arXiv:2001.09473*.
- Patro, J., Baruah, S., Gupta, V., Choudhury, M., Goyal, P., & Mukherjee, A. (2019). Characterizing the spread of exaggerated health news content over social media. In *Proceedings of the 30th acm conference on hypertext and social media* (pp. 279–280).
- Pei-Chi, L., & Ee-Peng, L. (2018). On learning psycholinguistics tools for english-based creole languages using social media data. In *2018 ieee international conference on big data (big data)* (pp. 751–760).
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of liwc2015* (Tech. Rep.).
- Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.
- Petrocchi, M., & Viviani, M. (2022). Romcir 2022: Overview of the 2nd workshop on reducing online misinformation through credible information retrieval. In *European conference on information retrieval* (pp. 566–571).
- Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In *Communication and persuasion* (pp. 1–24). Springer.
- Peute, L. W., de Keizer, N. F., & Jaspers, M. W. (2015). The value of retrospective and concurrent think aloud in formative usability testing of a physician data query



## References

- tool. *Journal of biomedical informatics*, 55, 1–10.
- Pluviano, S., Della Sala, S., & Watt, C. (2020). The effects of source expertise and trustworthiness on recollection: The case of vaccine misinformation. *Cognitive Processing*, 21(3), 321–330.
- Qazi, U., Imran, M., & Offi, F. (2020). Geocov19: A dataset of hundreds of millions of multilingual covid-19 tweets with location information. *SIGSPATIAL Special*, 12(1), 6–15.
- Qazvinian, V., Rosengren, E., Radev, D., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 1589–1599).
- Qureshi, K. A., Malick, R. A. S., & Sabih, M. (2021). Social media and microblogs credibility: Identification, theory driven framework, and recommendation. *IEEE Access*, 9, 137744–137781.
- Rani, S., Gill, N. S., & Gulia, P. (2021). Analyzing impact of number of features on efficiency of hybrid model of lexicon and stack based ensemble classifier for twitter sentiment analysis using weka tool. *Indonesian Journal of Electrical Engineering and Computer Science*, 22(2), 1041–1051.
- Rath, B., Gao, W., & Srivastava, J. (2019). Evaluating vulnerability to fake news in social networks: A community health assessment model. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 432–435).
- Rauchfleisch, A., & Kaiser, J. (2020). The false positive problem of automatic bot detection in social science research. *PloS one*, 15(10), e0241045.
- Resende, J., Durelli, V. H., Moraes, I., Silva, N., Dias, D. R., & Rocha, L. (2020). An evaluation of low-quality content detection strategies: Which attributes are still relevant, which are not? In *International conference on computational science and its applications* (pp. 572–585).
- Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the web. *Journal of the American society for information science and technology*, 53(2), 145–161.

## References

- Rieh, S. Y. (2010). Credibility and cognitive authority of information. In *Encyclopedia of library and information sciences* (pp. 1337–1344).
- Rieh, S. Y. (2014). Credibility assessment of online information in context. *Journal of information science theory and practice: JISTaP*, 2(3), 6–17.
- Rieh, S. Y., & Belkin, N. (2000). Interaction on the web: Scholars' judgement of information quality and cognitive authority. In *Proceedings of the annual meeting-american society for information science* (Vol. 37, pp. 25–38).
- Rieh, S. Y., Kim, Y.-M., Yang, J. Y., & St. Jean, B. (2010). A diary study of credibility assessment in everyday life information activities on the web: Preliminary findings. *Proceedings of the American Society for Information Science and Technology*, 47(1), 1–10.
- Roberts, J. S., McLaughlin, S. J., & Connell, C. M. (2014). Public beliefs and knowledge about risk and protective factors for alzheimer's disease. *Alzheimer's & Dementia*, 10, S381–S389.
- Robillard, J. M., Johnson, T. W., Hennessey, C., Beattie, B. L., & Illes, J. (2013). Aging 2.0: Health information about dementia on twitter. *PLoS One*, 8(7), e69861.
- Rodriquez, J. (2013). Narrating dementia: Self and community in an online forum. *Qualitative Health Research*, 23(9), 1215–1227.
- Roesslein, J. (2009). tweepy documentation. *Online*] <http://tweepy.readthedocs.io/en/v3>, 5.
- Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? an agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, 28(4), 394–412.
- Sabbeh, S. F., & Baatwah, S. Y. (2018). Arabic news credibility on twitter : An enhanced model using hybrid features. *journal of theoretical & applied information technology*, 96(8).
- Saha Roy, S., et al. (2020). " how good are they?"-a state of the effectiveness of anti-phishing tools on twitter (Unpublished doctoral dissertation). University of Texas at Arlington.

## References

- Salih, A. A., & Abdulrazaq, M. B. (2019). Combining best features selection using three classifiers in intrusion detection system. In *2019 international conference on advanced science and engineering (icoase)* (pp. 94–99).
- Sandelowski, M. (2000). Combining qualitative and quantitative sampling, data collection, and analysis techniques in mixed-method studies. *Research in nursing & health*, *23*(3), 246–255.
- Sandim, H., Azevedo, D., da Silva, A. P. C., & Moro, M. M. (2018). The role of social capital in information diffusion over twitter: A study case over brazilian posts. In *Bidu-posters vldb*.
- Santini, R. M., Salles, D., Tucci, G., Ferreira, F., & Graef, F. (2020). Making up audience: Media bots and the falsification of the public sphere. *Communication Studies*, *71*(3), 466–487.
- Sayed, H., Dafoulas, G., & Saleeb, N. (2018). Social network sites and methodological practices. *International Journal of Advanced Technology and Engineering Exploration*, *5*(38), 1–10.
- Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2020). Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th acm international conference on information & knowledge management* (pp. 2725–2732).
- Sbaffi, L., Rowley, J., et al. (2017). Trust and credibility in web-based health information: A review and agenda for future research. *Journal of medical Internet research*, *19*(6), e7579.
- Schulz, K., Rauenbusch, J., Fillies, J., Rutenburg, L., Karvelas, D., & Rehm, G. (2022). User experience design for automatic credibility assessment of news content about covid-19. *arXiv preprint arXiv:2204.13943*.
- SDRC. (2022). *Scottish dementia research consortium annual report 2021/22*. Retrieved from <https://www.sdrc.scot/wp-content/uploads/2022/05/SDRC-Report-2022.pdf>
- Selbæk, G. (2021). Dementia risk: Time matters. *The Lancet Public Health*, *6*(2), e85–e86.

## References

- Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In *Emerging technology in modelling and graphics* (pp. 99–111). Springer.
- Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, *9*(1), 1–9.
- Shariff, S. M. (2020). A review on credibility perception of online information. In *2020 14th international conference on ubiquitous information management and communication (imcom)* (pp. 1–7).
- Shariff, S. M., Zhang, X., & Sanderson, M. (2017). On the credibility perception of news on twitter: Readers, topics and features. *Computers in Human Behavior*, *75*, 785–796.
- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, *19*(1), 22–36.
- Shu, S., & Woo, B. K. (2021). Use of technology and social media in dementia care: Current and future directions. *World Journal of Psychiatry*, *11*(4), 109.
- Sicilia, R., Giudice, S. L., Pei, Y., Pechenizkiy, M., & Soda, P. (2018). Twitter rumour detection in the health domain. *Expert Systems with Applications*, *110*, 33–40.
- Sicilia, R., Merone, M., Valenti, R., Cordelli, E., D’Antoni, F., De Ruvo, V., . . . Soda, P. (2018). Cross-topic rumour detection in the health domain. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (bIBM)* (pp. 2056–2063).
- Sillence, E., Briggs, P., Fishwick, L., & Harris, P. (2004). Trust and mistrust of online health sites. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 663–670).
- Silverman, D. (2015). *Interpreting qualitative data*. Sage.
- Snelson, C. L. (2016). Qualitative and mixed methods social media research: A review of the literature. *International Journal of Qualitative Methods*, *15*(1), 1609406915624574.

## References

- Sommariva, S., Vamos, C., Mantzarlis, A., ào, L. U.-L., & Martinez Tyson, D. (2018). Spreading the (fake) news: Exploring health messages on social media and the implications for health professionals using a case study. *American journal of health education, 49*(4), 246–255.
- Son, J., Lee, J., Oh, O., Lee, H. K., & Woo, J. (2020). Using a heuristic-systematic model to assess the twitter user profile’s impact on disaster tweet credibility. *International Journal of Information Management, 54*, 102176.
- Soucy, P., & Mineau, G. W. (2005). Beyond tfidf weighting for text categorization in the vector space model. In *Ijcai* (Vol. 5, pp. 1130–1135).
- Spence, P. R., Edwards, A., Edwards, C., & Jin, X. (2019). ‘the bot predicted rain, grab an umbrella’: Few perceived differences in communication quality of a weather twitterbot versus professional and amateur meteorologists. *Behaviour & Information Technology, 38*(1), 101–109.
- Statista. (2022, Apr). *Global penetration social media 2021*. Retrieved from <https://www.statista.com/statistics/274773/global-penetration-of-selected-social-media-sites/>
- Steiner, V., Pierce, L. L., & Salvador, D. (2016). Information needs of family caregivers of people with dementia. *Rehabilitation Nursing, 41*(3), 162–169.
- St Louis, C., & Zorlu, G. (2012). Can twitter predict disease outbreaks? *Bmj, 344*.
- Su, Q., Wan, M., Liu, X., Huang, C.-R., et al. (2020). Motivations, methods and metrics of misinformation detection: An nlp perspective. *Natural Language Processing Research, 1*(1-2), 1–13.
- Sugiyama, M. (2015). *Introduction to statistical machine learning*. Morgan Kaufmann.
- Sun, Y., Zhang, Y., Gwizdka, J., & Trace, C. B. (2019). Consumer evaluation of the quality of online health information: Systematic literature review of relevant criteria and indicators. *Journal of medical Internet research, 21*(5), e12522.
- Sundar, S. S. (2008). *The main model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative Cambridge, MA.
- Talbot, C. V., O’Dwyer, S. T., Clare, L., Heaton, J., & Anderson, J. (2020). How

## References

- people with dementia use twitter: A qualitative analysis. *Computers in Human Behavior*, *102*, 112–119.
- Talbot, C. V., O'Dwyer, S., Clare, L., Heaton, J., & Anderson, J. (2020). Identifying people with dementia on twitter. *Dementia*, *19*(4), 965–974.
- Talbot, C. V., O'Dwyer, S. T., Clare, L., & Heaton, J. (2021). The use of twitter by people with young-onset dementia: A qualitative analysis of narratives and identity formation in the age of social media. *Dementia*, *20*(7), 2542–2557.
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54.
- Tawfik, A. A., Gill, A., Hogan, M., S York, C., & Keene, C. W. (2019). How novices use expert case libraries for problem solving. *Technology, Knowledge and Learning*, *24*(1), 23–40.
- Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, *27*(2), 237–246.
- Thomas, P., Kazai, G., White, R., & Craswell, N. (2022). The crowd is made of people: Observations from large-scale crowd labelling. In *Acm sigir conference on human information interaction and retrieval* (pp. 25–35).
- Timberg, C., & Dwoskin, E. (2019, Dec). *Twitter is sweeping out fake accounts like never before, putting user growth at risk*. Retrieved from <https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/>
- Tombros, A., Ruthven, I., & Jose, J. M. (2005). How users assess web pages for information seeking. *Journal of the American society for Information Science and Technology*, *56*(4), 327–344.
- Tseng, S., & Fogg, B. (1999). Credibility and computing technology. *Communications of the ACM*, *42*(5), 39–44.
- Tshimula, J. M., Chikhaoui, B., & Wang, S. (2022). Discovering affinity relationships between personality types. *arXiv preprint arXiv:2202.10437*.
- Twitter. (2016). *An update on our efforts to combat violent extremism*. Retrieved

## References

- from [https://blog.twitter.com/en\\_us/a/2016/an-update-on-our-efforts-to-combat-violent-extremism](https://blog.twitter.com/en_us/a/2016/an-update-on-our-efforts-to-combat-violent-extremism) (Accessed: Dec-2021)
- Twitter. (2017a). *Automation rules*. Retrieved from <https://help.twitter.com/en/rules-and-policies/twitter-automation> (Accessed: Dec-2021)
- Twitter. (2017b). *Report violations*. Retrieved from <https://help.twitter.com/en/rules-and-policies/twitter-report-violation> (Accessed: Dec-2021)
- Unkel, J., & Haas, A. (2017). The effects of credibility cues on the selection of search engine results. *Journal of the Association for Information Science and Technology*, 68(8), 1850–1862.
- Vanderstoep, S. W., & Johnson, D. D. (2008). *Research methods for everyday life: Blending qualitative and quantitative approaches* (Vol. 32). John Wiley & Sons.
- Van Someren, M., Barnard, Y. F., & Sandberg, J. (1994). The think aloud method: A practical approach to modelling cognitive. *London: AcademicPress*, 11.
- Varol, O., Ferrara, E., Davis, C., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the international aaai conference on web and social media* (Vol. 11).
- Veronese, N., Demurtas, J., Smith, L., Michel, J. P., Barbagallo, M., Bolzetta, F., . . . Maggi, S. (2022). Influenza vaccination reduces dementia risk: A systematic review and meta-analysis. *Ageing Research Reviews*, 73, 101534.
- Viehbeck, S. M., Pettecrew, M., & Cummins, S. (2015). Old myths, new myths: Challenging myths in public health. *American journal of public health*, 105(4), 665–669.
- Volkova, S., Shaffer, K., Jang, J. Y., & Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 647–653).
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151.
- Vraga, E. K., & Bode, L. (2018). I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information, Communica-*

## References

- tion & Society*, 21(10), 1337–1353.
- Wagle, V., Kaur, K., Kamat, P., Patil, S., & Kotecha, K. (2021). Explainable ai for multimodal credibility analysis: Case study of online beauty health (mis)-information. *IEEE Access*, 9, 127985–128022.
- Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4), 5–33.
- Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, 240, 112552.
- Washha, M. (2018). *Information quality in online social media and big data collection: An example of twitter spam detection* (doctoral dissertation). Université de Toulouse, Université Toulouse III-Paul Sabatier.
- Washha, M., Qaroush, A., Mezghani, M., & Sèdes, F. (2017). Information quality in social networks: Predicting spammy naming patterns for retrieving twitter spam accounts. In *International conference on enterprise information systems* (Vol. 2, pp. 610–622).
- Waszak, P. M., Kasprzycka-Waszak, W., & Kubanek, A. (2018). The spread of medical fake news in social media—the pilot quantitative study. *Health policy and technology*, 7(2), 115–118.
- Wathen, C. N., & Burkell, J. (2002). Believe it or not: Factors influencing credibility on the web. *Journal of the American society for information science and technology*, 53(2), 134–144.
- Weir, G. R. (2007). The posit text profiling toolset. In *Proceedings of the 12th conference of pan-pacific association of applied linguistics*.
- Weir, G. R., Dos Santos, E., Cartwright, B., & Frank, R. (2016). Positing the problem: Enhancing classification of extremist web content through textual analysis. In *2016 ieee international conference on cybercrime and computer forensic (icccf)* (pp. 1–3).
- Weir, G. R., & Ozasa, T. (2010). Learning from analysis of japanese efl texts. *Educational Perspectives*, 43, 56–66.



## References

- Weitzel, L., de Oliveira, J. P. M., & Quaresma, P. (2014). Measuring the reputation in user-generated-content systems based on health information. *Procedia Computer Science*, *29*, 364–378.
- Westerman, D., Spence, P. R., & Van Der Heide, B. (2012). A social network as information: The effect of system generated reports of connectedness on credibility on twitter. *Computers in Human Behavior*, *28*(1), 199–206.
- WHO. (2021, Sep). *Dementia: Key facts*. World Health Organization. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/dementia>
- Wilson, P. (1983). Second-hand knowledge: An inquiry into cognitive authority.
- Wu, K., Yang, S., & Zhu, K. Q. (2015). False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st International Conference on Data Engineering* (pp. 651–662).
- Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: Definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, *21*(2), 80–90.
- Xu, J., & Li, H. (2007). Adarank: A boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 391–398).
- Xu, Q. (2013). Social recommendation, source credibility, and recency: Effects of news cues in a social bookmarking website. *Journalism & Mass Communication Quarterly*, *90*(4), 757–775.
- Yamamoto, Y., & Tanaka, K. (2011). Enhancing credibility judgment of web search results. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1235–1244).
- Yang, J., Counts, S., Morris, M. R., & Hoff, A. (2013). Microblog credibility perceptions: comparing the USA and China. In *Proceedings of the 2013 conference on computer supported cooperative work* (pp. 575–586).
- Yang, K.-C., Ferrara, E., & Menczer, F. (2022). Botometer 101: Social bot practicum for computational social scientists. *arXiv preprint arXiv:2201.01608*.
- Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019).

## References

- Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1), 48–61.
- Yang, K.-C., Varol, O., Hui, P.-M., & Menczer, F. (2020). Scalable and generalizable social bot detection through data selection. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 1096–1103).
- Yang, L., Qiu, M., Gottipati, S., Zhu, F., Jiang, J., Sun, H., & Chen, Z. (2013). Cqarank: Jointly model topics and expertise in community question answering. In *Proceedings of the 22nd acm international conference on information & knowledge management* (pp. 99–108).
- Young, S. D., Mercer, N., Weiss, R. E., Torrone, E. A., & Aral, S. O. (2018). Using social media as a tool to predict syphilis. *Preventive medicine*, 109, 58–61.
- Yue, D., Yu, G., Shen, D., & Yu, X. (2014). A weighted aggregation rule in crowdsourcing systems for high result accuracy. In *2014 ieee 12th international conference on dependable, autonomic and secure computing* (pp. 265–270).
- Zannettou, S., Sirivianos, M., Blackburn, J., & Kourtellis, N. (2019). The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality (JDIQ)*, 11(3), 1–37.
- Zhang, H., & Su, J. (2004). Naive bayesian classifiers for ranking. In *European conference on machine learning* (pp. 501–512).
- Zhang, X., & Zhu, R. (2021). How source-level and message-level factors influence journalists' social media visibility during a public health crisis. *Journalism*, 14648849211023153.
- Zhang, Y., Shah, D., Foley, J., Abhishek, A., Lukito, J., Suk, J., ... Garlough, C. (2019). Whose lives matter? mass shootings and social media discourses of sympathy and policy, 2012–2014. *Journal of Computer-Mediated Communication*, 24(4), 182–202.
- Zhang, Z., & Ahmed, W. (2019). A comparison of information sharing behaviours across 379 health conditions on twitter. *International journal of public health*, 64(3), 431–440.
- Zhao, H., Xu, X., Song, Y., Lee, D. L., Chen, Z., & Gao, H. (2018). Ranking users in

## References

- social networks with higher-order structures. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).
- Zhao, Y., Da, J., & Yan, J. (2021). Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. *information processing & management*, 58(1), 102390.
- Zhao, Y., & Zhang, J. (2017). Consumer health information seeking in social media: a literature review. *Health Information & Libraries Journal*, 34(4), 268–283.
- Zheng, Z., Cai, Y., & Li, Y. (2015). Oversampling method for imbalanced classification. *Computing and Informatics*, 34(5), 1017–1037.

# Appendices

# Appendix A

## List of LIWC features

Category	Examples	Category	Examples
Word Count (WC)		<b>Cognitive processes</b>	Cause, know, ought
<b>Summary Language Variables</b>		Insight	Think, how
Analytical thinking (Analytic)		Causation	Because, effect
Clout		Discrepancy	Should, would
Authentic		Tentative	Maybe, perhaps
Emotional tone		Certainly	Always, never
Words/sentence (WPS)		Differentiation	Hasn't, but, else
Words > 6 letters (Sixltr)		Perceptual processes	Look, heard, feeling
Dictionary words (Dic)		See	View, saw, seen
Linguistic Dimensions		Hear	Listen, hearing
Total function words	It, to, no, very	Feel	Feels, touch
Total pronouns	I, them, itself	<b>Biological processes</b>	Eat, blood, pain
Personal pronouns	I, them, her	Body	Check, hands, spit
1st pers singular	I, me, mine	Health	Clinic, flu, pill
1st pers plural	We, us, our	Sexual	Horny, love, incest
2nd person	You, your, thou	Ingestion	Dish, eat, pizza
3rd pers singular	She, her, him	<b>Drives</b>	
3rd pers plural	They, their, they'd	Affiliation	Ally, friend, social
Impersonal pronouns	It, it's, those	Achievement	Win, success, better
Article	A, an, the	Power	Superior, bully
Prepositions	To, with, above	Reward	Take, prize, benefit
Auxiliary verbs	Am, will, have	Risk	Danger, doubt
Common Adverbs	Very, really	<b>Time orientations</b>	
Conjunctions	And, but, whereas	Past focus	Ago, did, talked
Negations	Not, not, never	Present	focus Toady, is, now
<b>Other Grammar</b>		Future focus	Will, may, soon
Common verbs	Eat, come, carry	Relativity	Area, bend, exit
Common adjectives	Free, happy, king	Motion	Arrive, car, go
Comparisons	Greater, best, after	Space	Down, in, thin
Interrogatives	How, when, what	Time	End, until, season
Numbers	Second, thousands	<b>Personal concerns</b>	
Qualifiers	Few, many, much	Work	Job, majors, Xerox
<b>Psychological Processes</b>		Leisure	Cook, chat, movie
Affective processes	Happy, cried	Home	Kitchen, landlord
Positive emotion	Love, nice, sweet	Money	Audit, case, owe
Negative emotion	Hurt, ugly, nasty	Religion	Altar, church
Anxiety	Worried, fearful	Death	Bury, coffin, kill
Anger	Hate, kill, annoyed	<b>Informal language</b>	
Sadness	Crying, grief, sad	Swear words	Fuck, damn, shit
Social process	Mate, talk, they	Netspeak	Btw, lol, thx
Family	Daughter, dad, aunt	Assent	Agree, OK, yes
Friends	Buddy, neighbour	Nonfluencies	Er, hm, umm
Female reference	Girl, her, mom	Fillers	I mean, you know
Male reference	Boy, his, dad		

source: (Pennebaker et al., 2015)

## Appendix B

### Example of posit run command

```
root@test-VirtualBox:/usr/bin/posit# posit_all.sh TweetsBook2
posit_all.sh: command not found
root@test-VirtualBox:/usr/bin/posit# positd.sh TweetsBook2

Running Posit on each text item...

:::::::::::::::::::: Processing Item 1 ::::::::::::::::::::::

Converting to sentences..
Total number of sentences: 7
Counting total number of words
Tokenizing TweetsBook2/Book1.txt
Extracting unique tokens...
Counting token frequencies...
Counting word types...
Number of characters: 640

Removing temporary files...
Tagging input file... (may take some time)
Invoking POS tagger...
loading the models from the directory "/usr/bin/posit/model_wsje02-21/" ...loadi
g /usr/bin/posit/model_wsje02-21/model.la...done
done
```

## Appendix C

### Example of Posit output file

---

```
@RELATION NEWS
```

```
@ATTRIBUTE classification {NEGATIVE, POSITIVE}
@ATTRIBUTE total_words NUMERIC
@ATTRIBUTE total_unique_words NUMERIC
@ATTRIBUTE ttr NUMERIC
@ATTRIBUTE number_of_sentences NUMERIC
@ATTRIBUTE asl NUMERIC
@ATTRIBUTE number_of_chars NUMERIC
@ATTRIBUTE awl NUMERIC
@ATTRIBUTE noun_types NUMERIC
@ATTRIBUTE verb_types NUMERIC
@ATTRIBUTE adjective_types NUMERIC
@ATTRIBUTE preposition_types NUMERIC
@ATTRIBUTE possessive_types NUMERIC
@ATTRIBUTE personal_types NUMERIC
@ATTRIBUTE determiner_types NUMERIC
@ATTRIBUTE adverb_types NUMERIC
@ATTRIBUTE particle_types NUMERIC
@ATTRIBUTE interjection_types NUMERIC
@ATTRIBUTE verbs NUMERIC
@ATTRIBUTE nouns NUMERIC
@ATTRIBUTE preposition NUMERIC
@ATTRIBUTE possessive NUMERIC
@ATTRIBUTE personal NUMERIC
@ATTRIBUTE particles NUMERIC
@ATTRIBUTE interjections NUMERIC
@ATTRIBUTE determiners NUMERIC
@ATTRIBUTE adverbs NUMERIC
@ATTRIBUTE adjectives NUMERIC
```

## Appendix D

### Posit features

No	Features	No	Features
1	Total words (tokens)	15	Possessive pronoun types
2	Total unique words (types)	16	Particle types
3	Type/Token Ratio (TTR)	17	Interjection types
4	Number of sentences	18	Nouns
5	Average sentence length (ASL)	19	Verbs
6	Number of characters	20	Prepositions
7	Average word length	21	adjectives
8	Noun types	22	Determiners
9	Verb types	23	Adverbs
10	Adjective types	24	Personal pronouns
11	Adverb types	25	Possessive pronouns
12	Preposition types	26	Particles
13	Personal pronoun types	27	interjections
14	Determiner types		



# Appendix E

## Botometer versions

Botometer Versions	#Features	Training Datasets Names	Datasets Descriptions
V1 (Davis et al., 2016)	1,150	caverlee	<b>caverlee:</b> dataset consists of bots enticed by honeypot accounts and confirmed human accounts (K. Lee, Eoff, & Caverlee, 2011)
V2 (Varol et al., 2017)	1,150	caverlee, varol-icwsm	<b>varol-icwsm:</b> manually labeled bot and humans accounts selected by Botometer score grades (Varol et al., 2017)
V3 (K.-C. Yang et al., 2019)	1,209	caverlee, varol-icwsm, cresci-17, pornbots, vendor-purchased, botometer-feedback, celebrity, political bots	<b>cresci-17:</b> Spam bots (conventional spambots, social) and humans (Cresci et al., 2017) <b>Pornbots:</b> dataset consisting of bot groups that share scam sites. shared by Andy Patel (github.com/r0zetta/pronbot2) used in (K.-C. Yang et al., 2019). <b>vendor-purchased:</b> fake followers bought by the CNetS team and researchers from various companies. <b>botometer-feedback:</b> physically annotating identified accounts reported by Botometer users. <b>celebrity:</b> dataset consisting of accounts collected from celebrities. <b>political-bots</b> dataset of politically oriented groups bots shared by Twitter user @josh emerson.
V4(Sayyadharikandeh et al., 2020)	1,200	caverlee, varol-icwsm, cresci-17, pornbots, vendor-purchased, botometer-feedback, celebrity, political-bots, gilani-17, cresci-rtbust, cresci-stock, botwiki, astroturf, midterm-2018, kaiser-1, kaiser-2, kaiser-3, combined-test	<b>gilani-17:</b> Physically labeled bots and humans selected by accounts grouped into four popularity groups based on the number of followers (Gilani, Farahbakhsh, Tyson, Wang, & Crowcroft, 2017) <b>cresci-rtbust:</b> all Italian retweets between 17–30 June 2018, collected and manually labelled into an almost balanced set of human and bot accounts (Mazza, Cresci, Avvenuti, Quattrociocchi, & Tesconi, 2019) <b>cresci-stock:</b> dataset of accounts with similar timelines during five months in 2017, from tweets with selected cashtags(Cresci et al., 2017) <b>botwiki:</b> dataset based on self-identified bot accounts from the botwiki.org archive. <b>midterm-2018:</b> dataset based on political tweets during the 2018 U.S. midterm elections (K.-C. Yang, Varol, Hui, & Menczer, 2020). <b>astroturf:</b> a new dataset that includes hyper-active political bots following and/or systematically deleting trains content (Sayyadharikandeh et al., 2020) <b>kaiser-1 ,kaiser-2 ,kaiser-3:</b> datasets of American and German politicians, containing manually annotated German language bots and accounts listed in the botwiki dataset (Rauchfleisch & Kaiser, 2020)

source:<https://botometer.osome.iu.edu/bot-repository/datasets.html>

## Appendix F

# Example of Botometer API response

```
{
  "scap": {
    "english": 0.0018818614025648,
    "universal": 0.5557322218336633
  },
  "display_scores": {
    "english": {
      "astroturf": 0.0,
      "fake_follower": 4.1,
      "financial": 1.5,
      "other": 4.7,
      "overall": 4.7,
      "self_declared": 3.2,
      "spammer": 2.8
    },
    "universal": {
      "astroturf": 0.3,
      "fake_follower": 3.2,
      "financial": 1.6,
      "other": 3.8,
      "overall": 3.8,
      "self_declared": 3.7,
      "spammer": 2.3
    }
  },
  "raw_scores": {
    "english": {
      "astroturf": 0.0,
      "fake_follower": 0.81,
      "financial": 0.3,
      "other": 0.94,
      "overall": 0.94,
      "self_declared": 0.63,
      "spammer": 0.57
    },
    "universal": {
      "astroturf": 0.06,
      "fake_follower": 0.64,
      "financial": 0.3133333333333333,
      "other": 0.76,
      "overall": 0.76,
      "self_declared": 0.74,
      "spammer": 0.47
    }
  },
  "user": {
    "majority_lang": "en",
    "user_data": {
      "id_str": "11330",
      "screen_name": "test_screen_name"
    }
  }
}
```

# Appendix G

## Ethics approval

6/23/22, 10:09 AM

CIS Ethics Approval System – Computer and Information Sciences – local.cis



**Computer and Information  
Sciences - local.cis**

departmental information for staff and students

[Home](#) | [Ethics](#) | [Events](#) | [Safety](#) | [Systems Support](#) | [Teaching](#) | [Utilities](#)

[Browse: Home](#) / [Utilities](#) / CIS Ethics Approval System

### CIS Ethics Approval System

You are *Fatimah Alhayari* (Research Student - 201790662)

[Return to Main](#)

Application ID: 1232

**Title of research:**  
Assessment of Dementia-related Information on Twitter

**Summary of research (short overview of the background and aims of this study):**  
This research aims to explore the type of dementia information searched by users on Twitter and their perception of the dementia related tweets credibility. The researcher intend to make interviews for collecting information from caregivers of people with dementia to understand their opinions and experiences with dementia related information on Twitter.

Note: Ethical Approval was granted for the last application for the same research, which was supposed to be conducted face to face (Application NO was 985) but was not conducted. Slight changes have been made in the research protocol to carry it out online.

**How will participants be recruited?**

Participants will be invited to participate from the viewpoint of a caregiver, or partner who cares for a person with dementia in a one-on-one interview. The participants will be recruited by posting an invitation link on social media platforms (i.e. Twitter or Facebook). The invitation link is a registration form that displays a brief about the study and criteria that people need to meet to be eligible for participation, but full details will be on the information sheet. Participants will be required to carefully read the information sheet and consent form before deciding whether to participate or not. If they decide to participate, they will enter their name, contact information and the time options that suit them for conducting the interviews. Registration form as in the following link:  
[https://strathsci.qualtrics.com/jfe/form/SV\\_eRueePu77i0P](https://strathsci.qualtrics.com/jfe/form/SV_eRueePu77i0P)

We will request from some UK dementia organizations that interested parties, either caregivers or partners of people with dementia circulate the interview invitation to potential participants for the study. Suggested networks are:

TIDE-together in dementia everyday

NDCAN-Alzheimer Scotland's National Dementia Carers Action Network.

My supervisor has a list of key contacts at the above-mentioned organizations. We will contact them as soon as I have the ethics form approved, to request the possibility to circulate the invitation link.

**What will the participants be told about the proposed research study?** Either upload or include a copy of the briefing notes issued to participants. In particular this should include details of yourself, the context of the study and an overview of the data that you plan to collect, your supervisor, and contact details for the Departmental Ethics Committee.

PDF File: [View document](#)

Information Sheet Uploaded

**How will consent be demonstrated?** Either upload or include here a copy of the consent form/instructions issued to participants. It is particularly important that you make the rights of the participants to freely withdraw from the study at any point (if they begin to feel stressed for example), nor feel under any pressure or obligation to complete the study, answer any particular question, or undertake any particular task. Their rights regarding associated data collected should also be made explicit.

PDF File: [View document](#)

Consent Form Uploaded- is part of the registration

**What will participants be expected to do?** Either upload or include a copy of the instructions issued to participants along with a copy of or link to the survey, interview script or task description you intend to carry out. Please also confirm (where appropriate) that your supervisor has seen and approved both your planned study and this associated ethics application.

PDF File: [View document](#)

PDF File: [View document](#)

If the participant agrees to participate, the researcher will arrange the interview time with the participant by email. During the interview, the researcher will ask the participant to fill in the questionnaire and then will ask the participants to perform a task. Once the participant completed the task, the researcher will start a semi-

<https://local.cis.strath.ac.uk/wp/extras/ethics/?view=1232>

1/2

## Appendix G. Ethics approval

6/23/22, 10:09 AM

CIS Ethics Approval System – Computer and Information Sciences – local.cis

structured interview by using open-ended questions regarding the participant's opinion and experience.

Questionnaire as in the link : [https://strathsci.qualtrics.com/jfe/form/SV\\_e5UI#CsuDf9Jz](https://strathsci.qualtrics.com/jfe/form/SV_e5UI#CsuDf9Jz)  
The task description and the questions had been uploaded.

I confirm that my supervisor has seen and approved the forms associated ethics application.

**What data will be collected and how will it be captured and stored? In particular indicate how adherence to the Data Protection Act and the General Data Protection Regulation (GDPR) will be guaranteed and how participant confidentiality will be handled.**

The participant data, audio recordings, and transcribed interview will be securely stored on the University of Strathclyde's secure cloud file storage service, with access only available to the research team in accordance with the General Data Protection Regulation (GDPR) provisions. For more information about GDPR visit <https://www.strath.ac.uk/dataprotection/gdprfaq/>

The identifiable information provided by participants i.e. participant name will be recorded anonymously, and will not be released by the researcher.

**How will the data be processed? (e.g. analysed, reported, visualised, integrated with other data, etc.) Please pay particular attention to describing how personal or sensitive data will be handled and how GDPR regulations will be met.**

The interview transcripts will be analysed by using content analysis software (such as Nvivo) in order to find if there are specific themes collected from the participants.  
The final result will give the researcher insight about to what extent caregivers use Twitter to find dementia information. Also, if the users consider specific criteria to assess the information they find.

**How and when will data be disposed of? Either upload a copy of your data management plan or describe how data will be disposed.**

**PDF File:** None.

To protect participants, the following steps will be taken with regards to anonymity and confidentiality of information:  
Records of the audio recording will be destroyed at the end of the project. The participant data, transcribed interviews and consent forms will be kept for four years after the researcher's PhD study period is completed, and then they will be destroyed.

<https://local.cis.strath.ac.uk/wp/extras/ethics/?view=1232>

2/2

# Appendix H

## Consent form

7/16/2020

Online Survey Software | Qualtrics Survey Solutions



### [Participant Information Sheet](#)

Below is the Informed Consent Agreement for this study. Please review each individual statement below and provide your initials for each statement you agree before signing the form.

- I confirm that I have read and understood the Participant Information Sheet for the above project and the researcher has answered any queries to my satisfaction.
- I confirm that I have read and understood the Privacy Notice for Participants in Research Projects and understand how the information provided by me, will be used and what will happen to it (i.e. how it will be stored and for how long).
- I consent to be audio recorded on Zoom as part of the interview.
- I consent for anonymised data which do not contain my identity information to be made available for research purposes.
- I understand that my participation is voluntary and that I am free to withdraw from the project at any time, up to the point of interview completion, without having to give a reason and without any consequences.

### Signature:

SIGN HERE

clear

### Date (mm/dd/yyyy):

[https://strathsci.qualtrics.com/jfe/form/SV\\_eRulePuf710Pj](https://strathsci.qualtrics.com/jfe/form/SV_eRulePuf710Pj)

1/2

# Appendix I

## Information sheet



### Participant Information Sheet for Participants

**Title of the research** Assessment of Dementia-related Information on Twitter

**Dear Participants,**

You are invited to participate in a research study designed to assess your thoughts about dementia information on Twitter. You will get a £15 purchase voucher from Amazon for your participation. Before you decide to do so, please take time to read the following information. If you have any other questions about the research, please ask the researcher.

**1. What is the purpose of this research?**

This research aims to explore the type of dementia information searched by users and their opinions toward dementia-related information on Twitter.

**2. Why have you been invited to take part?**

You are invited to take part in a research study because you are caregiver/partner caring for a person with dementia, residing in the UK, and use Twitter for finding or posting information related to dementia.

**3. Do I have to take part?**

It is up to you to decide whether or not to take part. If you decide to take part, you will be required to read this sheet carefully and be confident that you understand its contents. You will be asked to submit an online consent form as well.

Your participation in this research is voluntary and you are free to withdraw at any time up to the end of your interview without giving a reason. You also have the right to withdraw retrospectively any consent given, and to require that any data gathered from you be destroyed.

**4. What will happen to me if I take part? What do I have to do?**

- You will fill in the registration form with your name, email and three options of the best times that suit you to attend an interview on Zoom. The researcher then will send you a confirmation email within 24 hours with the interview date/time and invitation details (link with password).
- During the interview, the researcher will send you a short web-based questionnaire about standard demographics, your roles in relation to the person with dementia and your social media usage. This part takes approximately 3-5 minutes to answer.
- After that, you will be given a short-simulated task scenario which will be explained after you complete the survey. Then, the researcher will ask you open-ended questions in relation to your experience and opinion toward the information shown during the task. You will have approximately 5-10 minutes to perform the task and 15-20 minutes to answer the questions.

**5. Will I be asked about the person I take care of?**

There are no direct questions related to the person with dementia will be asked except the stage he/she is in.

The place of useful learning

The University of Strathclyde is a charitable body, registered in Scotland, number SC015263

## Appendix I. Information sheet



### **6. Will I be recorded, and how will the recorded media be used?**

The interview will be audio recorded and transcribed with your permission.

### **7. What are the potential risks to you in taking part?**

Participating in the task is not anticipated to cause the participants any disadvantage or discomfort.

### **8. What are the possible benefits of taking part?**

Your contribution in this interview is valuable for research, and the findings will assist researchers and stakeholders in understanding the role of Twitter in dementia information. You will also receive £15 voucher for Amazon as a thank you for your participation.

### **9. How and when will I receive the voucher?**

After the interview is over, you will receive the voucher via email.

### **10. Will my taking part in this project be kept confidential?**

All the information that we collect about you during the course of the research will be kept strictly confidential and will only be accessible to members of the research team. You will not be able to be identified in any reports or publications. There will be no way to identify the participants from the information they provided. You will be identified by ID number (e.g. P1). Your contact details will never be used in this research so you cannot be recognised from it.

### **11. What is the legal basis for processing my personal data?**

The data will be recorded anonymously, i.e. your name or any identifiable information will not be associated with them.

### **12. Where will the information be stored and how long will it be kept for?**

To protect participants, the following steps will be taken with regards to anonymity and confidentiality of information:

The participant data, audio recordings, and transcribed interview will be securely stored on the University of Strathclyde's secure cloud file storage service, with access only available to the research team in accordance with the General Data Protection Regulation (GDPR) provisions. For more information about GDPR visit <https://www.strath.ac.uk/dataprotection/gdprfaq/>

Records of the audio recording will be destroyed at the end of the project. The participant data, transcribed interviews and consent forms will be kept for four years after the researcher's PhD study period is completed, and then they will be destroyed.

### **13. What will happen to the results of the research project?**

Results of the research will be published and included in the researcher's PhD thesis. You will not be identified in any report or publication. If you wish to be given a copy of any reports resulting from the research, please ask the researcher.

### **14. If I wish to change any information provided during the study, how should I go about it?**

The interviewee will be given four weeks to communicate to the researcher any notes, concerns or modifications. Once these four weeks are over, it will be assumed that the interviewee agrees with all provided information.

The place of useful learning

The University of Strathclyde is a charitable body, registered in Scotland, number SC015263

## Appendix I. Information sheet



### 15. Who should I contact for further information?

Researcher: Fatimah Alhayan      Email: [fatimah.alhayan@strath.ac.uk](mailto:fatimah.alhayan@strath.ac.uk)

Primary Supervisor: Dr. Diane Pennington      Email: [diane.pennington@strath.ac.uk](mailto:diane.pennington@strath.ac.uk)

### 16. Who has ethically reviewed the project?

This investigation was granted ethical approval by the CIS Departmental Ethics Committee, Department of Computer and Information Sciences. If you have any questions/concerns, during or after the investigation, or wish to contact an independent person to whom any questions may be directed or further information may be sought from, please contact: [ethics@cis.strath.ac.uk](mailto:ethics@cis.strath.ac.uk)



## Appendix J

# High-low level search keywords

High Level Search Keywords	
Category	Keywords
Professional key-words	'dementia specialist', 'dementia researcher', 'special interest in dementia', 'dementia consultant', 'Dementia's disease researcher', 'at Dementia's Research', 'research in Dementia', 'Geriatrician', 'at Alzheimer's Research', 'Alzheimer Researcher', 'Alzheimer's disease researcher', 'Neuropsychologist', 'Gerontologist', 'Psychologist', 'deep into Alzheimer's disease', 'interest in ageing', 'interest in neurodegenerative', 'Cognitive health specialist', 'Medical doctor', 'Occupational Therapist', '#researcher', 'Clinical psychologist', 'Senior Lecturer', 'Neurologist', 'Neuroscientist', 'Physician', 'Biomedical scientist', 'Professor in', 'Psychiatrist', 'pathologist', 'Rehabilitation Consultant', 'Medical Teacher'
organisation key-words	'The Alzheimer Society', 'Alzheimer's Association', 'Alzheimer's Disease International (ADI)', 'The International Psychogeriatric Association (IPA)', 'ian Association', 'The official', 'Our mission:', 'Dementia Helpline', 'dementia organisation', 'our vision', 'we', 'call us', 'Tel', 'join us', 'like us on', 'follow us on', '#Helpline', 'Official Twitter page', 'Official account', 'Non-profit organisation', 'mission is', 'A premier provider', 'institution', 'provides', 'aims to', 'Dementia Forum X', 'charity', 'is a forum', 'Our', 'Founded in', 'consulting firm', 'is a', 'Institute of', 'Official Twitter account', 'a growing community', 'Aging Company', 'is the foundation', 'placement company', 'Email us at', 'center for'
Low Level Search Keywords	
Category	Keywords
Home	'Care-providers', 'Home Care Assistance', 'home caregiving services', 'home care', 'health services', 'Leading non-profit', 'homecare services', 'the best care', 'state-of-the-art Memory Care', 'Inc.', '24/7'
Promoters	'your campaign promoted', 'Get Promotion For Your' Media', 'Sign up to our', 'breaking news', 'is an online magazine', 'Follow us for news', 'Follow us for healthcare news', 'Journal of', 'The Journal of', 'Latest medical news', 'Daily, peer reviewed medical news', 'relevant medical news', 'medical news from', 'free CMEs', 'medical news', 'Track the latest', 'news site for', 'the latest medication news', 'fully open access journal', 'Sign up for health tweets'

# Appendix K

## User categorisation codebook

Category	Sub-category	Description of Qualifying User Account
Individuals-Professionals	Medical Professionals (IP-MP)	Individual users who include professional medical titles (as recognized by health practitioner registrations boards) in their biographical descriptions, e.g. doctor, registered nurse, nurse, physician, neurologist etc. or academic titles, e.g. professor of clinical neuropsychology, professor of integrative medicine, etc.) Examples of terms and/or phrases indicate medical titles you may need in their descriptions including, but not limited to: [Neurologist] [Neuro-surgeon] [Neuro-psychologist] [organisational Psychologist] [Social geriatrician] [Occupational Therapist] [Rehabilitation Consultant] [Mental health specialist] [Nursing home doctor][Biomedical Scientist] [Speech pathologist] or combinations of the above.
Individuals-Other	Caregiver (IO-C)	Formal/informal caregivers who provide care to person with dementia, regardless of whether he/she has medical qualifications or an occupation relating to the field of dementia/Alzheimer's disease.
	Health Activist (IO-HA)	Individual users who are dementia/Alzheimer's/mental health advocates or who are involved in active campaigning with the purpose of bringing about human or social change in the field of healthcare.
	Artist (IO-A)	Individual users notable for their fame in art, such as music, photography, or visual arts.
	Marketer (IO-M)	Individual users who specialize in marketing to promote their own products, books or equipment or work on the behalf of a company/organisation to promote products, books, equipment, etc.
	Author (IO-AU)	Individual users who are expert writers and publish written material in works such as books, newspapers, magazines, etc.
	Others (IO-LP)	Individual users who do not belong in the above categories.
Entities	General organisations (E-G)	These include government/public organisations, private organisations, non-profit organisations, interest groups, or charities that provide emotional support, activities, research, arrange seminars and develop communities.
	Care Providers (E-OCP)	Entities including profit or non-profit home care-providers or providers of services specifically for people with dementia and/or their caregivers and families. It may include agency or web directory help to find senior care-providers. Bio-descriptors may include phrases such as home care assistance, care-giver services, carer-services, nursing services, caregiver training, private duty home care, mobility assistance, memory care, rehabilitation, health and wellbeing services, music therapy etc.
	Promoters (E-P)	Promoters include technology and product development companies related directly to healthcare (e.g., devices, pharmaceuticals, biotechnologies). They also include marketing companies providing services or products not related directly to healthcare (e.g., law services, food, furniture).
	Media (E-MN)	Media includes electronic media such as news channels (BBC, CNN), print media such as newspapers (New York Times), research media (journal articles, research papers etc.), websites or social media profiles (Face-book, Instagram) to provide tips and information related to health.
	Books and Apps	<b>Books (E-B):</b> An account for book publishers, tweeting about collections of books or a specific published book. <b>Dementia App (E-AD):</b> An account for a software program/app/-tool /game/system that is specifically designed to serve people with dementia or Alzheimer's disease, their families and caregivers. <b>Health App (E-AH):</b> An account for a software program/application/tool that is designed to increase general health and well-being.
Empty and Unknown	Unknown	Unknown includes places or events (e.g., conferences).
	Empty	The Empty category refers to profiles without descriptions.

## Appendix L

### Features selected by Anova

Features Selected by ANOVA Method			
No	Feature	Score	Frequency
1	money	644.93496	2
2	Dic	642.75527	2
3	Clout	629.90947	3
4	awl	566.35033	2
5	WC	565.04396	3
6	relativ	534.28937	2
7	Analytic	517.07969	2
8	relig	494.87957	2
9	function	492.3392	2
10	number_of_chars	477.71379	3
11	Colon	468.06965	2
12	leisure	452.64873	2
13	home	423.97226	3
14	prep	399.97683	1
15	friend	352.04086	1
16	motion	334.72817	1
17	number_of_sentences	309.77435	2
18	particles	301.11864	1
19	WPS	297.24687	2
20	pronoun	297.15201	1
21	swear	288.95721	3
22	verbs	275.38443	1
23	total_unique_words	257.72512	3
24	noun_types	241.87904	2
25	nouns	235.87869	1
26	Period	233.44937	1
27	SemiC	228.36591	1
28	space	227.54714	1
29	Sixltr	218.46623	2
30	URL	212.59725	3

## Appendix M

### Features selected by REF

Features Selected by Recursive Feature Elimination Method			
No	Feature	Rank	Frequency
1	total_unique_words	1	3
2	number_of_sentences	1	2
3	number_of_chars	1	3
4	noun_types	1	2
5	verb_types	1	1
6	preposition_types	1	1
7	article	1	2
8	WC	1	3
9	Clout	1	3
10	negate	1	2
11	adj	1	2
12	anger	1	2
13	cause	1	1
14	percept	1	1
15	see	1	1
16	hear	1	1
17	feel	1	1
18	bio	1	1
19	drives	1	1
20	tentat	1	1
21	Authentic	1	2
22	time	1	1
23	home	1	3
24	death	1	1
25	swear	1	3
26	filler	1	1
27	Tone	1	1
28	QMark	1	1
29	cogproc	1	2
30	URL	1	3

## Appendix N

### Features selected by RF

Features Selected by Random Forest Method			
No	Feature	Score	Frequency
1	total_words	0.048687	1
2	Clout	0.045916	3
3	home	0.035692	3
4	Dic	0.031569	2
5	Sixltr	0.030395	2
6	awl	0.030091	2
7	relativ	0.029352	2
8	WC	0.024487	3
9	function	0.021606	2
10	money	0.019045	2
11	adj	0.017237	2
12	negate	0.01684	2
13	total_unique_words	0.016649	3
14	number_of_chars	0.0158	3
15	relig	0.015462	2
16	URL	0.014329	3
17	Analytic	0.013608	2
18	listed_count	0.013578	1
19	leisure	0.013108	2
20	differ	0.012459	1
21	anger	0.012343	2
22	article	0.012312	2
23	WPS	0.012191	2
24	cogproc	0.011782	2
25	risk	0.011758	1
26	statuses_count	0.010394	1
27	swear	0.009801	3
28	Colon	0.009771	2
29	Authentic	0.009142	2
30	friends_count	0.009064	1


# Appendix O

## Example of high bot score profile

... · 11 Sep 2020

More evidence that proactive healthy steps promote a healthy brain!

Multi-focus program cuts lifestyle risk for dementia  
[mcknights.com/news/clinical-...](#) #Alzheimers #FCKAlzheimers #No2Alzheimers



mcknights.com  
Multi-focus program cuts lifestyle risk for dementia  
Seniors educated on the benefits of a health lifestyle, Mediterranean diet, physical activity and ...

---

... · 24 Jul 2020


More research on factors around a healthy brain - Read in good (brain) health!  
Study reveals factors that keep the oldest adults cognitively sharp  
[mcknights.com/news/clinical-...](#) #Alzheimers #FCKAlzheimers #No2Alzheimers

1

---

... · 22 Jun 2020

More great things to nosh on to support a healthy brain!  
Can grapes halve a person's dementia risk? [mcknights.com/news/clinical-...](#)  
#Alzheimers #FCKAlzheimers #No2Alzheimers



mcknights.com  
Can grapes halve a person's dementia risk?  
People who eat a handful of grapes twice a day might cut their Alzheimer's disease risk virtually in half, a new study suggests.

# Appendix P

## Questionnaire

7/16/2020

Online Survey Software | Qualtrics Survey Solutions



**1- Your Gender :**

- Male
- Female
- Prefer not to answer

**2- Your Age:**

- More than 51 years
- 36-50 years
- 21-35 years
- Less than 21 years
- Prefer not to answer

**3- Your Educational Level:**

- High school or equivalent
- College
- Undergraduate
- Postgraduate

**4- Your relationship to the person living with dementia:**

- Formal Carer (i.e. nurse)
- Informal Carer (i.e. family member, friend)
- Other ("Please Specify:")

**5- Which stage of the disease is the person with dementia currently in?**

- Early stage

[https://strathsci.qualtrics.com/jfe/form/SV\\_e5UHCsuDI6I9Jz](https://strathsci.qualtrics.com/jfe/form/SV_e5UHCsuDI6I9Jz)

1/4

## Appendix P. Questionnaire

7/16/2020

Online Survey Software | Qualtrics Survey Solutions

- Mid stage
- Severe midstage
- Late stage
- Not sure

6- How often do you usually use social media to read information related to dementia?

- Never
- Rarely
- Sometimes
- Often
- Always

7- How often do you usually use social media to post information related to dementia?

- Never
- Rarely
- Sometimes
- Often
- Always

8-

	How often do you read information related to dementia from these platforms?					How often do you post information related to dementia from these platforms?		
	Never	Rarely	Sometimes	Often	Always	Never	Rarely	Sometimes
Twitter	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Facebook	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
YouTube	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Blogs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Other ("Please Specify:") <input type="text"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

9- How frequently do you use Twitter?

- Hourly
- Daily

[https://strathsci.qualtrics.com/jfe/form/SV\\_e5UHICsuD1519Jz](https://strathsci.qualtrics.com/jfe/form/SV_e5UHICsuD1519Jz)

2/4



## Appendix P. Questionnaire

7/16/2020

Online Survey Software | Qualtrics Survey Solutions

- Weekly
- Monthly
- Not sure

### 10- How long have you used Twitter?

- Less than 1 year
- 1-2 years
- 3 to 4 years
- More than 4 years

### 11- Which of the following account types do you usually follow, or check to find information related to dementia? And would you please give a description or examples of these accounts? (You can choose more than one answer)

- Individuals
- Professionals
- Organizations
- News
- Home Care Providers
- Books
- Apps
- Companies provide services or products
- Other ("Please Specify")

### 12- What type(s) of information from each of these accounts do you usually look for? (You can choose more than one answer)

- Support Groups/Carers Communities
- Research based Information (Articles)

[https://strathsci.qualtrics.com/jfe/form/SV\\_e5UHCsuDl5l5Jz](https://strathsci.qualtrics.com/jfe/form/SV_e5UHCsuDl5l5Jz)

3/4

7/16/2020

Online Survey Software | Qualtrics Survey Solutions

- Care Services
- Events
- Medication
- Other ("Please Specify")

Back

Next

Powered by Qualtrics 

# Appendix Q

## Task session homepage

Restart Survey [Place Bookmark](#) Tools ▾

Mobile view off

**Task:**  
Your partner has been diagnosed with dementia recently. You would like to help by finding out what is generally recommended for people in his /her situation. Six types of Twitter source (users) links are shown below. There are two users from different categories. Explore the users in each category and select which of these options you think will be a reliable source for the task.

**Instructions :**

- 1- Open both user links in each category.
- 2- Select the user(s) you think will be a reliable source for the task. You can select "None", if you prefer neither.
- 3- Justify your selection for each in the text entry below the user link.
- 4- Rate the credibility of each user 1-7, least to best.
- 5- You are encouraged to think aloud while you are exploring the profiles.

*I will leave you on your own while you are completing the task.*

# Appendix Q. Task session homepage

Restart Survey [Place Bookmark](#) Mobile view of Tools

**Category 1**

<a href="https://twitter.com/User1">https://twitter.com/User1</a>	<a href="https://twitter.com/User2">https://twitter.com/User2</a>	None
---	---	------

◀ ▶

User 1							User 2						
1	2	3	4	5	6	7	1	2	3	4	5	6	7
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Category 2**

<a href="https://twitter.com/User1">https://twitter.com/User1</a>	<a href="https://twitter.com/User2">https://twitter.com/User2</a>	None
---	---	------

◀ ▶

User 1							User 2						
1	2	3	4	5	6	7	1	2	3	4	5	6	7
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Category 3**

<a href="https://twitter.com/User1">https://twitter.com/User1</a>	<a href="https://twitter.com/User2">https://twitter.com/User2</a>	None
---	---	------

◀ ▶

User 1							User 2						
1	2	3	4	5	6	7	1	2	3	4	5	6	7
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

◀ ▶