
# The Users' Behaviour in Audio Streaming Services: Investigations in the Music and Podcast Domains

Francesco Meggetto

Computer and Information Sciences

University of Strathclyde

Thesis submitted for the degree of *Doctor of Philosophy*

April 2024

Glasgow

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



Signed: Francesco Meggetto

Date: 3rd October 2023

# Abstract

In recent years, online audio streaming services (e.g., Amazon Music and Spotify) have witnessed an increase in popularity due to content digitisation. These platforms, now offering on-demand music, personalised playlists, and recently, podcasts, have significantly transformed users' behaviour and their interactions with these platforms [1]. Podcasts, defined as spoken documents that can be represented by their transcriptions [2, 3], are swiftly becoming a central medium for online information seeking activities. Due to their considerable demand, streaming services have expanded their catalogues to include podcasts alongside music [4, 5]. Thus, there is an important research need for effective, cross-domain, and multi-modal, information access tools and methods that can guide users through these vast content libraries by aligning with their preferences and needs.

However, despite the research relevance, understanding, modelling, and predicting how users interact with content on such streaming services remains under-researched [3, 6]. This thesis aims to address this gap by delving into the nuances of users' behaviour in order to improve our overall understanding. This is motivated by the invaluable stream of information that an accurate representation of the users' behaviour can provide to the underlying recommendation process. In particular, the focus of this thesis is the intricate relationship between understanding users' behaviours, predicting these, and developing novel user-centric interfaces that are informed by these findings. This research is performed across both the music and podcast domains, aiming to unravel new facets of users' behaviour and thus inform novel user modelling and recommendation techniques.

The first part of the thesis focuses on understanding and predicting users' behaviour

in the music domain. In particular, it presents extensive investigations into users' skipping behavior during listening sessions. Chapter 3 introduces a novel approach to identify fine-grained session skipping behaviors. Four major session skipping patterns are identified through extensive evaluation, namely the *listener*, *listen-then-skip*, *skip-then-listen*, and *skipper*. A subsequent analysis of the differences among these patterns under varying listening contexts is also presented. With a deeper understanding of the users' music skipping behaviour, Chapter 4 investigates the utility of users' historical data for the task of sequentially predicting users' skipping behaviour. To this end, the applicability and effectiveness of Deep Reinforcement Learning (DRL) for this task is demonstrated. An in-depth post-hoc and ablation analysis indicates that users' behaviour features are the most discriminative of how the proposed DRL model predicts music skips. Content and contextual features are reported to have a lesser effect.

The second part of the thesis delves into the podcast domain. Chapter 5 introduces *Podify*, the first web-based podcast streaming platform specifically designed for academic research. Resembling existing streaming services, *Podify* supports academic research in the podcast domain, specifically in the under-researched areas of search and user behavioural analysis. Chapter 6 and Chapter 7 present, respectively, the methodology of a user study conducted through *Podify* and a discussion on the impact of text-based components, such as captions and full-text transcriptions, on how users assess the relevance of podcast content to their information needs. This is motivated by their well-established multidimensional role for improved information accessibility [7,8], and the alignment with the principles of Universal Design [9, 10], which support the ability to cater to diverse audiences and learning styles [11, 12]. By combining qualitative (i.e., the participants' reported relevance judgements of podcasts) and quantitative (i.e., listening activity) data, the importance of these textual components in enabling users to better assess the relevance of podcast content is shown.

# Contents

# List of Figures

List of Figures

# List of Tables

List of Tables

# Preface/Acknowledgements

Embarking on this research journey was a significant challenge, made possible through the unwavering support of my supervisors, Dr. Yashar Moshfeghi, Professor Crawford Revie, and Dr. John Levine. I am deeply thankful for your patience, knowledge, encouragement, and your consistent belief in my abilities. In moments filled with doubt, you helped me to go through in this adventure. My gratitude also extends to the other academics, who undoubtedly provided the required support to submit this thesis. Thank you, Dr. Murray Wood.

Zuzana, I simply could not have done it without you. You are just too good to be true. Your presence, surrounded with your patience, guidance, and advice, has really made this possible. Words cannot describe how thankful I am to have you in my life. You have always been there since the first day. Always. I love you and I cannot await what the future holds for us! Let's make this last forever, because everything is better when you are around.

I dedicate this thesis to the rest of my family: Giovanni, Pina, Federico, and Giulia. Your unconditional love and constant support have been my foundation throughout this journey. Even during periods of intense research, whenever I needed somebody, you were always there to support me. Grazie di cuore a tutti voi.

Finally, I would also like to acknowledge all the friends that helped me in countless ways during this endeavour. Each one of you holds a special place in my heart, and I am beyond thankful for our friendship and your help in achieving this accomplishment. Every moment shared, and every laugh has played a key part in reaching this stage. Thank you Amine, Ali, Carlos, Jack, Jim, and all the others, which are too many to name. You really are a part of of this, and I will never forget you.

# Glossary

**ALPR** Adversarial Learning-based Podcast Representation. 23

**API** Application Programming Interface. 21, 23, 88

**ASR** Automatic Speech Recognition. 6, 22, 92, 111, 122, 123, 132, 142

**AWS** Amazon Web Services. 93, 94

**BA** Baseline. xi, xii, 97, 98, 99, 103, 107, 108, 115, 122, 123, 125, 126, 127, 128, 129, 130, 132, 133, 143

**BCSM** Between Cluster Scatter Matrix. 40

**CH** Calinski-Harabasz Index. x, 39, 40, 42, 43, 44

**CI** Confidence Interval. xiv, xv, 76, 82, 182

**CIS** Computer & Information Sciences Department. 98, 106

**CSV** Comma-Separated Values. 92, 93, 106

**CVI** Cluster Validity Index. 39, 40, 42, 53, 139

**DB** Davies-Bouldin Index. x, 40, 41, 42, 43, 44

**DDQN** Double Deep Q-Network. 61, 182

**Di** Difficult. xii, xv, 97, 100, 101, 102, 103, 108, 129, 130, 131, 132, 133, 134, 138, 143, 144

**DQN** Deep Q-Network. xi, xiii, xiv, xv, 20, 59, 61, 62, 64, 65, 66, 72, 73, 74, 76, 77, 78, 79, 80, 81, 82, 181, 182, 183

**DRL** Deep Reinforcement Learning. iii, 5, 6, 8, 15, 20, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 68, 74, 76, 77, 78, 79, 81, 83, 84, 137, 141, 142, 144, 145, 146, 181

**DRQN** Deep Recurrent Q-Network. 181, 182

**Ea** Easy. xii, xv, 97, 100, 101, 102, 103, 108, 129, 130, 132, 133, 134, 138, 143

**EN** Enriched. xi, xii, 97, 98, 99, 103, 105, 107, 108, 115, 122, 123, 125, 126, 127, 128, 129, 130, 132, 133, 143

**FPA** First Prediction Accuracy. xiv, xv, 70, 76, 77, 82, 83, 141, 182

**GDPR** General Data Protection Regulation. 93, 106

**GeMAPS** Geneva Minimalistic Acoustic Parameter Set. 23

**GPU** Graphics Processing Unit. 77, 84

**GRU** Gated Recurrent Unit. 72, 181, 182

**IN** Information Need. 4, 9, 22, 27, 28, 43, 96, 97, 100, 109, 111, 130, 131, 132, 138, 143, 147

**INE** Inertia Index. x, 40, 41, 42, 43, 44

**IO** Information Overload. 130, 133, 134, 144, 148

**IP** Internet Protocol. 94, 106

**IR** Information Retrieval. 6, 7, 21, 22, 28, 111, 112, 132, 134, 142, 143

**JSON** JavaScript Object Notation. 93

**KI** Known-Item. xii, xv, 96, 97, 98, 100, 101, 102, 103, 120, 129, 130, 131, 132, 133, 134, 138, 143, 144

**LSTM** Long Short Term Memory. 71, 72, 181, 182

**MAA** Mean Average Accuracy. xiv, xv, 70, 76, 77, 82, 83, 141, 182

**MC** Monte Carlo. 62

**MDP** Markov Decision Process. 59, 64, 65

**MFCC** Mel-frequency Cepstral Coefficients. 23

**MRS** Music Recommender System. 6, 15, 16, 17, 56, 57, 58, 83, 84, 141, 145, 146

**MSSD** Music Streaming Sessions Dataset. 19, 33, 37, 40, 53, 58, 67, 70, 71, 74, 75, 77, 81, 137, 139, 141, 142, 144, 145

**nDCG** Normalized Discounted Cumulative Gain. 101

**NIST** National Institute of Standards and Technology. 100

**NLP** Natural Language Processing. 21, 147

**NN** Neural Network. 60

**PC** Principal Component. 39

**PCA** Principal Component Analysis. 36, 39, 42

**PLP** Perceptual Linear Predictions. 23

**POMDP** Partially Observable Markov Decision Process. 64, 65

**PRS** Podcast Recommender System. 26, 27

**RE** Refinding. 98, 100

**RL** Reinforcement Learning. 20, 59, 60, 62, 63, 64

**RNN** Recurrent Neural Network. 20, 62, 71, 75, 76, 77, 78, 181

**RQ** Research Question. 28, 33, 34, 38, 40, 42, 43, 46, 48, 50, 58, 76, 79, 83, 86, 96, 103, 112, 122, 128, 133, 136, 138, 139, 141, 142, 143

**RS** Recommender System. 14, 16, 18, 26, 27, 28, 55, 56

**RSS** Really Simple Syndication. 6, 22, 23, 88, 93, 101, 104, 111

**SHAP** Shapley Additive Explanations. xi, 59, 74, 77, 78, 79, 80, 137

**TD** Temporal Difference. 62, 63, 66

**TO** Topical. xii, xv, 96, 97, 98, 100, 101, 102, 103, 117, 129, 130, 131, 132, 133, 134, 138, 143, 144

**TREC** Text REtrieval Conference. 6, 7, 13, 22, 23, 24, 30, 86, 96, 98, 100, 103, 105, 108, 111, 113, 114, 128, 132, 142

**UE** User Engagement. xii, 7, 8, 13, 16, 19, 22, 27, 29, 30, 50, 54, 104, 105, 109, 123, 125, 126, 127, 128, 133, 140

**UI** User Interface. xi, xii, 6, 7, 9, 10, 26, 28, 29, 86, 87, 89, 90, 91, 96, 97, 98, 107, 110, 111, 112, 115, 123, 124, 125, 128, 131, 132, 134, 138, 142, 143, 146, 147

**UKRI** UK Research and Innovation. 107

**WCSM** Within Cluster Scatter Matrix. 40

**WCSS** Within Cluster Sum-of-Squares Criterion. 41

**YAMNet** Yet Another MobileNet. 23

Glossary

# Part I

# Introduction and Background

# Chapter 1

# Introduction

This chapter introduces the context and underlying motivations for the research conducted in this thesis (Section 1.1). This is followed by the thesis statement in Section 1.2. Subsequently, the contributions to knowledge of this research are outlined (Section 1.3). Finally, the chapter presents the outline and structure for the rest of the thesis (Section 1.4), along with a list of publications derived from this research (Section 1.5).

## 1.1 Motivation

In the dynamic and continually evolving digital landscape, online audio streaming platforms such as Apple Music, Amazon Music, Pandora, and Spotify are witnessing an unprecedented surge in popularity. This is due to content digitisation and, for example, the introduction of on-demand music and the automatic generation of personalised playlists. Recently, we have also observed the integration of the high-demanded media of podcasts into such platforms. Overall, this has fundamentally revolutionised the ways users engage and interact with these digital platforms [1]. Podcasts are spoken documents that are representable through transcriptions of their content [2, 3], and they are experiencing increasing attention as a valuable medium for online information seeking activities. Recognising this increasing demand, audio streaming platforms have started to extend their catalogues to include podcasts, thereby offering both the music and podcast mediums [4, 5].

In the current digital age, music and podcasts are integral components of audio streaming platforms, each offering distinct, yet interconnected, user experiences. Typically, podcasts demand focused listening (i.e., listeners pay close attention to the content) and are primarily consumed during specific times, such as weekday mornings. On the other, music is frequently consumed in the background [3, 5]. Compared to music, with an average duration of typically a few minutes, the duration of podcast episodes can range from a few minutes to several hours. This necessitates a more considerable investment of time. Despite these differences, both mediums consist of audio content and exhibit some similarities. This includes a popularity bias problem, which permeates across existing streaming platforms [3]. Furthermore, they have an overlapping functional use for numerous people [5] and with entertainment being a key consumption goal within podcasts, partly similar to other multimedia items such as music and movies [3].

This significant growth in content availability prompts the necessity for more advanced, cross-domain, and multi-modal information access tools and methods. These are vital to aid users in navigating the vast amount of streaming content that is available. Thus, they play a pivotal role in helping users discover content tailored and aligned to their preferences and information needs (INs). However, despite the significance and relevance of this research area, comprehensively understanding, modelling, and predicting users' interactions and their behaviour within these platforms still remain significantly under-researched [3, 6]. Overall, these facets underscore the challenges in designing interfaces and recommendation systems that cater effectively to the nuances of user interaction with each medium. This PhD thesis aims to address this research gap by delving into the nuances of users' behaviour. The insights presented in this thesis can be leveraged to create more user-centric and fine-grained understandings of users' behaviour. This is an important step towards improving our understanding of how users interact with the streamed content. This is motivated by the invaluable stream of information that an accurate representation of the users' behaviour can provide to the underlying streaming platform (e.g., to the recommendation process). Specifically, the focus of this thesis is on the intricate relationship between

understanding users' behaviours, predicting these, and developing novel user-centric interfaces that are informed by these findings. An investigation of their interplay and resulting impact is performed in the music and podcast domains. In particular, in this thesis, I present comprehensive explorations into the patterns of users' behaviour on streaming platforms.

The first part of the thesis revolves around the music domain as it investigates the users' skipping behaviour from a large real world dataset of music streaming listening sessions (i.e., Spotify). Understanding users' skipping behaviour is an under-explored domain [6, 13, 14]. It is a challenging problem due to its noisy nature: a skip may suggest a negative interaction, but a user may also skip a song that they like because they recently heard it elsewhere. Previous work that analysed such skipping behaviour revealed universal behaviours in skipping across songs, with geography, audio fluctuations or musical events affecting how people skip music [15–17]. Recently, the effectiveness of deep learning models has also been explored for the task of predicting the users' sequential skipping behaviour in song listening sessions [18–24]. While they made a significant contribution towards this direction, their process is usually seen as an independent and static procedure. They may not account for the dynamic nature of the users' behaviour, and do not intuitively optimise for the long-term potential of user satisfaction and engagement [25–30]. The users' shifting interests and behaviour make it hard to learn a generalisable model to tailor to a user's specific needs at any given time; it is a case where Deep Reinforcement Learning (DRL) is required due to its capabilities for continuous learning and adaption [29–31]. Therefore, in this thesis, I tackle the task of identifying and categorising different behaviours during entire listening sessions with regards to the users' session-based skipping activity. To this end, I propose an effective data transformation and clustering-based approach. With a richer and more in-depth understanding of how people skip music, I then aim to understand how a DRL-based model predicts music skips. By comprehensively analysing the utility of users' historical data, I analyse the impact and effect of various factors in the classification task of predicting the users' music skipping behaviour. These factors include the users' behaviour (e.g., the user action that leads to the current playback to start),

listening content (i.e., the listened song), and contextual (e.g., the hour of the day) features. I propose a novel approach that leverages and adapts DRL for this classification task. This is to most closely reflect how a DRL-based Music Recommender System (MRS) could learn to detect music skips.

In the latter part of this thesis, I shift the focus to the podcast domain. Despite the significant growth and widespread recognition of this medium, this domain remains largely under-explored [3]. Unified platforms, incorporating both music and podcasts, present challenges and opportunities, requiring robust search systems to aggregate diverse content into a user-friendly interface (UI) [4, 32]. The coexistence of diverse media within a single platform raises questions about the optimal design of audio-focused information access systems. Specifically, there is a need for dedicated research to understand user behaviour and optimise podcast streaming platforms. This can be achieved by considering the unique characteristics of podcasts and the concept of relevance in this context [4, 33]. A critical challenge in podcast information retrieval (IR) is finding specific information within episodes. This issue was the focus of the 2020 and 2021 Text REtrieval Conference (TREC) Podcast Track[1], which attracted numerous submissions to its tasks of retrieval of fixed two-minute segments and episode summarisation [34]. The track was also associated with the Spotify Podcast Dataset, comprising over 100,000 episodes with audio files, transcriptions, metadata, and Really Simple Syndication (RSS) feeds [35]. These transcriptions, auto-generated through Automatic Speech Recognition (ASR) systems, allow content-based search and user navigation but pose challenges to standard IR methods because of their length and errors [36,37]. Therefore, there is a need to segment podcast episodes into, for example, fixed two-minute chunks (i.e. segments) [34]. Further, transcriptions serve as a powerful tool for bridging gaps and enhancing understanding across diverse audiences and domains. They facilitate access to content for the hearing-impaired community, aligning with principles of Universal Design [9, 10], and enhancing language learning and comprehension through theories such as Dual Coding [38–40] and the Cognitive Theory of Multimedia Learning [41]. This multi-modal approach caters to diverse learning

[1] https://trecpodcasts.github.io/

styles and aids in the retention of complex information [11,12]. In the realms of research and data analysis, transcriptions enable effective pattern recognition, thematic analysis, and support grounded theory methodologies [42, 43]. Searchable text enhances IR capabilities and aligns with the Information Foraging Theory, aiding in navigation and helping users to find information efficiently [7, 8]. These facets and issues motivate my investigation into the impact of incorporating text-based components, such as captions and full-text transcripts, into the UI of a podcast streaming platform. To this end, I release the first web-based podcast streaming platform (*Podify*) that is specifically designed to support academic research, with an emphasis on the under-researched areas of search and user behavioural analysis. Designed to closely resemblance existing streaming services, *Podify* offers a high-level of familiarity to users. By incorporating these textual components into the UI of *Podify*, my user study aims to determine whether this incorporation improves the user experience and whether it affects how users assess the relevance of podcast segments (i.e., a two-minute snippet of a podcast episode, with this concept originating from the TREC segment retrieval task).

Through its comprehensive exploration of users' behaviour, this thesis aims to provide valuable insights that can inform future development and design of more effective recommendation procedures in the rapidly evolving landscape of online audio streaming services.

## 1.2 Thesis Statement

The surge in popularity of both online music and podcasts, and their co-existence on the same streaming platforms, necessitates a deeper understanding of users' behaviours in order to be able to optimise their engagement with the streamed content. This PhD thesis delves into the intricacies of users' behaviour in these domains. It aims to unravel new facets of users' behaviour that could be leveraged to enhance the level of user engagement (UE). Specifically, UE will be measured through the analysis of the users' interaction patterns such as skipping behaviour in music streaming sessions, the integration of textual components in podcast platforms, and their respective impacts on

content relevance assessment. Success will be gauged by the ability to accurately predict music skips using DRL, the exploration of differences in skipping behaviours based on contextual factors (e.g., time of day, type of playlist, account type), and improved user experience and engagement metrics in podcast consumption by including captions and full-text transcriptions. The statement that this thesis aims to investigate, given the convergence and co-existence of music and podcast content on shared platforms, is: "*can I uncover and comprehend behavioural patterns that might enhance UE within music and podcast streaming?*". This investigation will be structured around key research questions focusing on session-level skipping behaviours, the effect of contextual variables on these behaviours, and the role of captions and transcripts in enhancing podcast user experience and relevance assessment.

## 1.3 Contributions

This section outlines the main contributions of this PhD thesis. This research delves into the nuances of users' behaviour in the music and podcast domains. Specifically, I focus on the intricate relationships between understanding and predicting users' behaviour and developing novel user-centric interfaces. The aim is to collect findings that can inform the development of user modelling, recommendation, and personalisation techniques. I believe these findings can improve the level of UE with the streamed content. The key contributions of this thesis are:

- Music Domain:
  - I conducted an extensive investigation into the users' music skipping behavior during listening sessions, which led to the identification and categorisation of four types of session skipping behaviour: the *listener*, *listen-then-skip*, *skip-then-listen*, and *skipper*.
  - I comprehensively analysed the utility of users' historical data in the classification task of predicting the users' music skipping behaviour. To this end, I proposed a novel approach that leverages and adapts DRL for this task.
- Podcast Domain:

- I designed a methodology that integrates a user study with the development and release of *Podify*. Developed to closely resemble existing streaming services, this web-based platform is specifically designed for academic research. *Podify* automatically logs all user interactions, which can be easily exported for subsequent analysis. It reduces the overhead researchers face when conducting user studies in the podcast domain.
- I investigated the effects of incorporating textual components, such as captions and full-text transcripts, into the *Podify*'s UI. The findings highlight the positive influence of these components on the users' process of accurately identifying the relevance of podcast content to their IN.

## 1.4 Thesis Layout

This thesis is organised into the following parts and corresponding chapters.

**PART I: Introduction and Background**

**Chapter 1** sets the context for the thesis, detailing its outline, aims, and contributions to the field.

**Chapter 2** provides a comprehensive foundation for the thesis, starting with a review of online audio streaming. Then, it delves into the music medium, the importance of user behaviour, and an analysis of music skipping behaviour. Finally, it reports an in-depth and comprehensive exploration of the podcast medium. This includes its recent integration into music streaming services and its distinctive characteristics and properties.

**PART II: The Music Skipping Behaviour**

**Chapter 3** aims to analyse and provide a deeper understanding of how users skip

music. The users' music skipping behaviour is investigated during entire listening sessions. A data transformation and clustering-based approach to identify and categorise skipping types is proposed.

**Chapter 4** delves deeper into understanding the music skipping behaviour. This chapter focuses on the task of predicting users' skipping behavior during music streaming sessions. Further, an analysis of the impact and effect of users' behaviour (e.g., the user action that leads to the current playback to start), listening content (i.e., the listened song), and contextual (e.g., the hour of the day) features in this task is proposed.

**PART III: The Podcast**

**Chapter 5** presents *Podify*, the web-based platform that serves as the foundation for the methodology and study described in the subsequent chapters. This chapter delves into the platform's features, UI, search functionality, catalogue creation procedure, user behaviour collection mechanisms, and technical implementation details.

**Chapter 6** outlines the methodology underpinning the analyses detailed in Chapter 7. It discusses the experimental design, procedure, and the participants' recruitment.

**Chapter 7** aims to evaluate the benefits of incorporating textual components, such as captions and full-text transcripts, into the *Podify*'s UI. It investigates their impact on user experience and the users' process of assessing the relevance of podcast content. Further, it examines the participants' engagement levels and the multi-faceted values added by these textual features.

**PART IV: Conclusions and Future Work**

**Chapter 8** concludes the thesis, highlighting the achieved objectives and limitations. It also discusses future research directions.

**APPENDIX**. The thesis includes two appendices. Appendix A complements the analysis conducted in Chapter 4. Appendix B includes the participant overview and information sheets, consent forms, task execution sheets, and questionnaires for the methodology described in Chapter 6 and the corresponding analyses of Chapter 7.

## 1.5 Publications

The research material presented in this thesis has been submitted and published to various peer-reviewed venues during the course of this PhD programme:

- Meggetto, F., Moshfeghi, Y. and Jones, R., 2021, September. On Building a Podcast Collection with User Interactions. In *Workshop on Podcast Recommendations, part of the 15th ACM Conference on Recommender Systems (RecSys)* (PodRecs '21) [44].
  The content of this paper is discussed in Chapter 2.

- Francesco Meggetto, Crawford Revie, John Levine, and Yashar Moshfeghi. 2021. On Skipping Behaviour Types in Music Streaming Sessions. *In Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (CIKM '21). Association for Computing Machinery, New York, NY, USA, 3333–3337. https://doi.org/10.1145/3459637.3482123 [14].
  The content of this paper is discussed in Chapter 3.

- Francesco Meggetto, Crawford Revie, John Levine, and Yashar Moshfeghi. 2023. Why People Skip Music? On Predicting Music Skips using Deep Reinforcement Learning. *In Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (CHIIR '23). Association for Computing Machinery, New York, NY, USA, 95–106. https://doi.org/10.1145/3576840.3578312 [45].
  The content of this paper is discussed in Chapter 4.

- Francesco Meggetto and Yashar Moshfeghi. 2023. Podify: A Podcast Streaming Platform with Automatic Logging of User Behaviour for Academic Research. *In*

*Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '23). Association for Computing Machinery, New York, NY, USA, 3215–3219. https://doi.org/10.1145/3539618.3591824 [2].

The content of this paper is discussed in Chapter 5.

# Chapter 2

# Background and Motivation

## 2.1 Introduction

This chapter provides the background knowledge for the various concepts and methodologies utilised throughout this thesis. It describes both the research and application context that underpins the work presented in Part II and III.

An introduction to online audio streaming services is provided in Section 2.2. Then, a detailed exploration of the music medium and the critical role of user behaviour is discussed in Section 2.3. In Section 2.4, I delve into the specific user behaviour of music skipping, by discussing its relevance, previous analyses, and state-of-the-art approaches for its prediction. Last, Section 2.5 introduces the domain of podcasts. This section offers a comprehensive overview of this medium, its inherent characteristics, the 2020 and 2021 TREC Podcast Track, and the latest studies on user consumption and behaviour. Further, I also delve into complex domains of podcast recommendation, the multi-modal nature of this medium, and the UE aspect.

## 2.2 Online Audio Streaming Services

Online audio streaming services are platforms that offer users vast libraries of music tracks, albums, playlists, and most recently, podcasts [4, 46]. These platforms have started a transformative shift from the traditional media consumption methods, which primarily relied on physical copies such as CDs, to digital access [47]. Nowadays, these

streaming platforms ensure that users have seamless access to their preferred music and podcasts at any time, and within a few clicks.

Emerging towards the end of the 20th century, these platforms have significantly redefined the music accessibility landscape. Precursors such as Napster started the digitisation of music in the late 1990s and early 2000s, albeit with multiple concerns, especially those related to copyright [48–51]. Notwithstanding this, they laid the foundation for a pivotal evolution in the music sector [51, 52]. By the latter part of the 2000s, revolutionary platforms such as Spotify, Apple Music, and SoundCloud were created [53]. From being digital music repositories, these platforms started revolutionising consumption behaviours. This resulted in a transition from traditional music ownership to a more adaptable, subscription-driven model [54]. Instead of buying individual tracks, users could browse vast libraries and stream any song (or podcast episode), either by paying a periodic fee (i.e., monthly or yearly) or freely with the insertion of ads [55].

There are several factors that contributed to the success of these streaming platforms. First, their ubiquitous accessibility allowed users to no longer be restricted from device storage limits (such as in MP3 players) and the need to purchase individual tracks in, for example, CDs. Additionally, these platforms feature catalogues that are continuously updated, promptly featuring the latest releases. Finally, these platforms introduced personalised user experiences. These are distinguishing features that most significantly impacted how users listen to music [56]. For example, Spotify's "Discover Weekly" [57] leverages advanced algorithms to suggest new music tailored to individual listening preferences and needs [58]. Overall, all these factors contributed to the global success of these platforms.

Besides hosting music, audio streaming platforms have also recently started integrating the podcast medium in their catalogues [4, 5]. Podcasts have expanded the user base and global impact of these platforms with their potential for attracting diverse audience groups [59]. The success of these streaming platforms is primarily because of their sophisticated recommendation procedures, commonly referred to as Recommender Systems (RSs) [60]. RSs aspire to tackle the problem of providing the users the

support they need to access these large collections of items and find songs or podcasts that match their interests and needs [61, 62]. These systems, by closely analysing user interactions (such as track skips or time spent on a song), curate personalised playlists and suggest content to users [63]. While traditional approaches such as collaborative filtering (where content is recommended based on similar user tastes) and content-based filtering (based on a user's historical interactions) remain relevant [61, 64, 65], recent research has also seen a significant effort towards integrating black-box based approaches such as Deep Learning [66] and DRL [67,68]. This is because they overcome the obstacles of conventional models (i.e., they can effectively capture non-linear and non-trivial user-item relationships) and achieve higher recommendation quality [66].

Overall, the past years have witnessed online music streaming services (e.g., Spotify) to achieve substantial growth [50]. Their rise in popularity, paired with the rise of digital music distribution and the ubiquitous availability of music, led to the emergence of new listening paradigms. Nowadays, these streaming platforms offer a listening ecosystem that is characterised by personal and tailored user experiences [69].

## 2.3 The Role of User Behaviour

Music recommender systems (MRSs) have information filtering algorithms at their core [63]. These algorithms are designed to curate relevant music content for users from extensive catalogs [63, 70]. Crucial to their effectiveness is the system's ability to leverage and understand user interactions, since they provide insights into the users' multifaceted behaviours [65, 71]. These interactions, often categorised as implicit and explicit feedback, offer unique perspectives on users' preferences [71–73].

***Explicit Feedback***. A manifestation of direct user input and interest in items, explicit feedback encompasses actions such as ratings, reviews, and expressive signals such as like or dislike [74–76]. Such feedback captures a user's intent, such as adding an album to favourites or liking a song [77]. Despite explicit feedback provides invaluable and accurate insights into user's preferences, by capturing both positive and negative preferences, it is usually scarce and rare. Thus, it is difficult to obtain sufficient and representative feedback from a population of users. This can be partially explained by

the considerable cognitive effort in its collection [74, 78].

***Implicit Feedback***. Explicit rating data, and especially in the music domain, is relatively scarce in today's systems. Even when available, it tends to be sparse. Therefore, modelling implicit feedback is becoming of acquired importance. This type of feedback is not provided directly by the user [79]. Instead, it is inferred from user actions and behaviours such as playback frequency, listening duration, or song skips [63, 74, 80–82]. Therefore, an implicit feedback system must rely on the application of domain-dependent tools and methodologies for capturing and interpreting this type of feedback. While abundant and domain-specific, implicit feedback only captures positive interactions, and it is prone to noise [83, 84]. This makes the analysis of such signals a challenging task. For instance, a track replayed multiple times might indicate a positive user's preference, but a skipped song does not necessarily imply a negative preference [73, 84].

Explicit and implicit feedback provide different degrees of expressivity of the user's preferences. In order to build effective RSs, an integration and comprehensive understanding of both feedback types is required. However, such systems face inherent challenges, including data sparsity and the cold-start problem (recommendations for items with no prior interactions) [63, 85]. In the realm of audio streaming, user behaviours such as song skipping hold the potential to offer profound understanding into the user's interests, preferences, and needs [14, 45, 84]. Its modelling and understanding during music listening sessions plays a crucial role in understanding users' behaviour [14]. The skips are often the only information available to the underlying MRS, and therefore they are used as a proxy to infer music preference [84]. By understanding the depth of these behavioural patterns and their nuances, MRSs can refine their algorithms and thus yield higher personalisation and UE.

Overall, integrating implicit and explicit feedback provides a way for MRSs to elicit user preferences. This is achieved by leveraging multiple facets of user behaviours, ranging from direct ratings to nuanced song skips. By accurately leveraging these patterns, MRSs can offer a more personalised and user-centric listening experience.

## 2.4 The Music Skipping Behaviour

A successful MRS needs to meet the users' various requirements at any given time [86–88]. Thus, user modelling is a key element. A line of research has tried to untangle the relationship between personality and the users' musical preferences [89–91]. Volokhin and Agichtein [92] introduced the concept of music listening intents and showed that intent is distinct from context (user's activity). A different, and arguably complementary, research direction is trying to understand and model how users interact with the underlying platform. With explicit rating data relatively scarce and rare in today's systems, modelling implicit feedback is becoming of acquired importance. This is a long-standing and under-researched problem of online streaming services [6]. An example of these interactions is the skips between songs. The skipping is a signal that can measure users' satisfaction, dissatisfaction or lack of interest, and engagement with the platform [14, 45]. Its modelling and understanding during music listening sessions plays a crucial role in understanding users' behaviour [14]. The skips are often the only information available to the underlying MRS, and therefore they are used as a proxy to infer music preference [84]. For example, in a lean-back formulation, the case of automatic playlists or radio streaming, the user interaction is minimised. Users are presented with a single song at a time. The MRS needs to rely almost entirely on implicit feedback signals such as the skipping or scrubbing (i.e., seeking forward and backward by moving the cursor [93]) to predict satisfaction and engagement [29, 30].

Recent research in music skipping behaviour can be categorised as belonging to one of three main categories: its relevance as an implicit feedback signal (Section 2.4.1), the analyses aimed at providing a deeper understanding of this behaviour (Section 2.4.2), and finally its prediction (Section 2.4.3).

### 2.4.1 Research Relevance

Research revolving around skipping behaviour on online platforms spans mainly across ads on social media platforms [94–96] and music [13, 15–17, 97]. The latter, however, has been largely under-researched. Modelling and, most importantly, understanding this

skipping behaviour in music listening sessions arguably play a crucial role in better understanding and defining user behaviour in modern streaming services. For instance, the skipping signal has already been used as a measure in heuristic-based playlist generation systems [98,99], user satisfaction [87,88,100], relevance [101], and as counterfactual estimators in RSs [102]. However, despite being abundant in quantity, they are noisy in nature [83]. A skipped track does not necessarily imply a negative preference. Implicit feedback is notoriously difficult to interpret as absolute relevance judgements due to multiple biases such as first impression and trust [76,103]. Their interpretation is made difficult due to the prevailing presence of false-positive interactions, which may not reflect the true user satisfaction [88], therefore they are a noisy measure of user's preferences. Finally, implicit feedback typically also consists of only the "positive-data", meaning that the negative feedback is missing in the available data collections [104,105].

### 2.4.2 Analysis

In a preliminary analysis of skip profiles [13], at an individual song level, it was noted that a quarter of all streamed songs are skipped within the first couple of seconds, and only half of all songs are listened to in their entirety. Montecchio et al. [15] identified a connection between skip behaviour and musical structure. They show that users are more likely to skip a song directly after a change of musical sections. The skip identity of a song is both very specific to the song as well as stable across time and geographical region. Such skip profile also follows a universal U-shaped pattern, with spikes in skipping rate at the beginning and end of the playback. In subsequent work by Donier [16], the idea of skips being, for the most part, reactions to salient musical events is further reinforced and confirmed. Wen et al. [88] show that post-click feedback (such as skips) is pervasive across domains. In their analysis, they show that more than half of all clicks in music and short videos lead to potential user dissatisfaction. Moreover, they note how the skipping behaviour manifests different patterns for music and short videos. Ng and Mehrotra [17] demonstrate that fluctuations in audio features are common in music streaming sessions and relate to the skipping behaviour of users. These fluctuations are also later studied in the work by Heggli et al. [106], where they

further identify that the users' musical preferences vary in terms of time of the day and day type. Fazelnia et al. [107] recently proposed a variational autoencoder-based model to process slow-moving and fast-moving features. In their analysis, they found a high pairwise correlation between total plays and skips, meaning that the skipping behaviour is found to be correlated with more plays overall. This finding strongly suggests that the skipping behaviour is a measure of UE with the streaming music, and further validates its usefulness to better understand the users' current preferences. In a controlled user study, Taylor and Dean [97] find that people who usually listen to songs in their entirety (users were asked in advance for this information), show higher listening duration than those who do not.

In contrast to prior works, in Chapter 3 I propose an in-depth analysis on users' skipping behaviour during entire listening sessions, with the aim of finding and characterising different types of session skipping behaviour. To this end, I propose an effective data transformation and clustering based-approach in order to identify different behaviours during entire listening sessions. This analysis is performed on the entire training set of the real-world Music Streaming Sessions Dataset (MSSD) [6].

### 2.4.3 Prediction

While numerous datasets have advanced music information retrieval and recommendation research, such as the Yahoo! Music Dataset [82], Million Playlist Dataset [108], and LFM-1b [109], they have not specifically focused on the intricacies of user behaviour, particularly in the context of music skipping. In 2019, Spotify identified music skip prediction as an important challenge. To encourage research in this under-developed field and to explore approaches that could alleviate this problem, they released the MSSD [6] and the *Sequential Skip Prediction Challenge* [1]. The challenge focused on predicting whether individual tracks encountered in a listening session will be skipped or not. To respond to this challenge, several deep-neural networks [18–24] and supervised learning [110] models were proposed. Afchar and Hennequin [111] proposed using interpretable deep neural networks for skip interpretation via feature attribution.

[1] https://www.aicrowd.com/challenges/spotify-sequential-skip-prediction-challenge

Whilst neural networks, and in particular Recurrent Neural Networks (RNNs), have been shown to effectively model sequential data, they consider the procedure as a static process. They do not intuitively provide a mechanism for the long-term optimisation of user satisfaction and engagement, continuous learning, and the modelling of the dynamic nature of the user's behaviour [25–27, 29, 30]. Therefore, it is a case where DRL is required, an investigation and application of which has never been explored before. A research gap that Chapter 4 aims to address.

### DRL for Classification

The *Sequential Skip Prediction Challenge* is a binary classification task. Despite receiving limited attention to date, DRL has been shown to be suitable and effective in classification tasks. It can assist classifiers in learning advantageous features [112, 113] and select high-quality instances from noisy data [114]. Wiering et al. [115] demonstrate that Reinforcement Learning (RL) is indeed suitable for classification. Their model slightly outperforms existing classifiers, but training time and extra computational requirements are major drawbacks. With the recent advances in the field, a body of research is showing the superiority of DRL-based approaches in general classification tasks [113, 114, 116–118]. In particular, the authors in [113, 116] show that a Vanilla Deep Q-Network (DQN) [119] approach is superior and more robust to state-of-the-art algorithms.

In particular, I explore, for the first time, the applicability of DRL in the task of sequentially predicting users' music skipping behaviour. This is motivated by the limitations of existing approaches and the advantages of DRL. By comprehensively analysing users' historical data, I study its utility and effect in my approach to this task. The work presented in Chapter 4 is the first step in understanding how people skip music from a DRL-based model perspective.

## 2.5 The Rise of Podcasts

Podcasts have seen an unprecedented rise in popularity in recent years. Initially referred to as "audioblogs" [120], these spoken documents, representable via their speech transcripts [2], have become a popular medium for online information seeking activities. Originating from the combination of the word "iPod" and "broadcast", podcasts are an online episodic series of digital audio files. From 2014 to 2020, the average monthly podcast listeners increased three-fold, reaching 15 million per month [121, 122]. With music streaming services such as Amazon Music, Apple Music, and Spotify, this has resulted in including both music and podcasts on a unified platform [4, 5]. Given their rise in popularity, they have become an integral part of people's listening habits. For example, as of 2023, the number of active podcasts has exceeded one million, with over 30 million episodes in over 100 languages [123]. The soaring popularity of podcasts is clear, with 75% of the United States population recognising the term "podcasting", 55% having listened to a podcast at least once, and 37% being monthly listeners [123]. Despite their long history and availability, they have only recently attracted significant research interest.

Traditional speech datasets have primarily focused on clean, structured audio from formal settings (e.g., TIMIT [124] or broadcast news corpora [125–127]), lacking the spontaneous, diverse nature of podcasts. In 2020, Spotify identified the podcast as an important research domain and released the Spotify Podcast Dataset [35]. This dataset uniquely fills this gap by offering a comprehensive collection of unscripted, conversational podcast content not represented in prior available datasets (e.g., the Stuttering Events in Podcasts (SEP-28k) [128], a podcast dataset dedicated to detecting stuttering events in speech). The Spotify Podcast Dataset is a large corpus of over 100,000 episodes, each comprising an audio file, automatically transcribed text via Google's Cloud Speech-to-Text Application Programming Interfaces (APIs), and associated metadata. Although the dataset's great applicability to various tasks in fields such as speech and audio processing, natural language processing (NLP), IR, and computational linguistics, it is unsuitable for those where logged user behaviour is re-

quired [44]. This is the case of analyses of user INs, their characteristics and behaviour, relevance, search, recommendation, and personalisation systems.

Recent research in podcasts is multifaceted, and it can be categorised as follows. Section 2.5.1 outlines the properties of podcasts. Sections 2.5.2 and 2.5.3 outline the TREC Podcast Track and previous research on user consumption and behaviour, respectively. Section 2.5.4 elaborates on previous work related to podcast recommendations, while Section 2.5.5 discusses the multi-modal nature of this medium. Finally, Section 2.5.6 concludes with a discussion on UE in the context of podcasts.

### 2.5.1 Properties

Podcasts are typically distributed as audio streams or files, commonly through RSS feeds. The RSS standard for podcasts includes various metadata fields [129]. However, this metadata often suffers from noise, making it inadequate and ill-defined. For instance, the quality and breadth of episode descriptions differ significantly, and category labels frequently prove unreliable because of incentives for creators to over-categorise to achieve higher exposure [3, 130, 131]. Contrary to other spoken documents (e.g., news [132] or TED Talks [133, 134]), podcasts exhibit distinctive characteristics. Their typical duration (averaging between half an hour to an hour), conversational style (unscripted or impromptu discourse), speaker count, format (e.g., interviews, monologues, debates), and extra content such as advertisements [135] contribute to the complexity of their analysis [3, 136].

To enable content-based search and indexing through conventional IR techniques, a valuable approach is to adopt a complete textual representation in the form of a transcript [3, 36, 137]. ASR systems are used to derive textual representations from audio streams. Nonetheless, these systems pose significant challenges due to the extended lengths of podcasts [37] (and thus necessitating segmentation), and the errors introduced by the ASR system (an 18% error was reported for the Spotify Podcast Dataset) [35]. Moreover, transcripts ignore the paralinguistic features of spoken language [138]. Martikainen et al. [138] proposed a clustering-based approach to group podcast episodes by their audio-based stylistic content.

To overcome the limitations of transcripts, representations of podcasts could be augmented with acoustic features such as Mel-frequency cepstral coefficients (MFCCs) [139], Perceptual Linear Predictions (PLPs) [140], and Adversarial Learning-based Podcast Representation (ALPRs) [141]. While these methods prove effective and can utilise unlabelled data, they are not interpretable with regard to downstream applications [3].

### 2.5.2 TREC Podcast Track

To foster research in this domain, Spotify released the English Podcast Dataset [35] in 2020. This dataset was later expanded to include the Portuguese language [142] and pre-computed audio features [143] such as the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and the Yet Another MobileNet (YAMNet) features. The dataset release coincided with the TREC Podcast Track [34], which was held in 2020 and 2021. The track focused on two shared tasks: segment retrieval and summarisation.

#### Search: Spoken Passage Retrieval

High-quality search of topical content of podcast episodes is challenging. Existing search engines primarily rely on indexing available metadata and textual descriptions of shows and episodes [35,144]. However, these descriptions often fail to capture the full breadth of content within the episodes. The segment retrieval task addresses this gap by aiming to identify relevant segments, from podcast episodes, based on a variety of search queries, ranging from specific phrases to broader topics. To facilitate this, text transcripts were automatically generated using Google's Cloud Speech-to-Text APIs from the complete audio files, which are part of the dataset. These transcripts provide detailed word-level time alignments, speaker diarisation, casing, and punctuation. This enhances the granularity at which content can be searched and retrieved. Additionally, the dataset encompasses metadata such as episode names, descriptions, publisher details, duration, and RSS headers, enriching the context for search and retrieval tasks. A segment was defined as a two-minute chunk starting at the minute mark (e.g., 120 - 139.9), allowing retrieval systems to index content with precise time offsets. Finally, a curated set of search information needs, called topics, was also released. These topics,

created following those used by the TREC [145], encompass keyword queries and descriptions of the user information needs. The needs can be one of three types, namely topical (general information about the topic), refinding (searching for a specific episode the user heard before), and known-item (finding something that is known to exist but under an unknown name) [137]. The evaluation of segment relevant to these topics is based on human judgements, with a gold standard data for assessing the performance.

### Summarisation

Conversely, the summarisation task involved the generation of a concise text snippet that accurately conveys the content of a podcast episode by considering its audio and transcription. Automated document summarisation is the task of condensing an input text into a shorter form that preserves most of the salient information. This task underscores the complexities inherent in processing automatically transcribed documents, which may contain speech recognition errors, conversational nuances, and significant longer length compared to typical summarisation inputs [35].

While creator-generated episode descriptions offer a starting point for summary generation, their variability in quality and intent poses additional challenges, reflecting the diverse motivations and genres present in podcast content [3, 130, 131]. Therefore, in training summarisation models, these descriptions could be considered as reference summaries. To derive a set of gold labelled data, a subset of episodes were manually annotated based on the outputs of different baseline systems. The annotators were asked to assess a summary's quality on a Perfect/Excellent/Good/Fair/Bad (PEGFB) scale, after reading the full transcript and/or listening to some of the audio if needed [35].

In response to these tasks, various approaches were proposed for both segment retrieval [146–149] and summarisation [33, 150–159]. A recent analysis of the podcast summaries conducted by Rezapour et al. [160] suggested that high-quality summaries tend to incorporate proper nouns, determiners, and adverbs. This is by limiting the use of verbs. Additionally, these summaries often contain more segments that are repeated from the input information.

### 2.5.3 User Consumption and Behaviour

Despite recent investigations into podcast consumption and listener behaviour, this research area remains largely under-explored. Tsagkias et al. [161–163] suggested four indicators, namely podcast content, creator, context, and technical execution, to characterise the quality, credibility, and prediction of podcast preferences. With the medium's rising popularity, and the ever-growing number of shows and episodes, it is crucial to devise novel approaches that can leverage this vast amount of podcast content and understand how users interact with this medium.

Despite some recent research efforts [5, 164–170], the absence of podcast streaming platforms for research and datasets containing logged user behaviour [44] has hindered progress in this field. This limitation is also relevant to the Spotify Podcast Dataset [44]. To address this challenge, in Chapter 5 I release *Podify*, a podcast streaming platform that automatically logs all user behaviour, and that is specifically designed for academic research. In particular, it aims to reduce the overhead that researchers face when conducting user studies in this domain.

Previous work aimed at understanding podcast usage patterns revealed motivations akin to those found in music consumption, such as relaxation and entertainment [164, 165], and for education purposes [166, 167]. A user study by Edison Research [170] endorsed this, reporting that over 70% of participants listen to podcasts for either entertainment or educational purposes. Furthermore, music is one of the most popular topics among podcast listeners. Integrating podcast listening into music streaming platforms could potentially influence and change the original way users engage with music in terms of listening time, duration, and frequency [5]. Li et al. [5] revealed that users tend to listen to podcasts during weekday mornings, with music listening happening during the evenings, nights or weekends. Hashemi et al. [171] conducted a comprehensive log analysis study contrasting podcast and music search behaviour. In their study, they found that the search effort for podcast retrieval is relatively higher than for music, with distinct differences in user consumption behaviours across these two mediums.

Podcasts have also been identified as effective educational tools, improving students'

performance [166] and offering easier access to knowledge and intellectually challenging content [167]. However, much of the prior research has focused on podcast usage in higher education, where students were required to use podcasts as part of their curriculum [168,169]. In addition, podcasts appear to be more favoured by technologically inclined demographics, such as men and young adults with a high income [164]. Chadha et al. [164] suggest that this difference may stem from men's presumed higher propensity to adopt new technologies. In line with this, Edison Research [170] also revealed that men represented the larger share of the podcast audience (54%). In Chapter 7, I extend the current understanding of podcast consumption by being the first to explore the concept of relevance in this domain. By analysing both explicit and implicit feedback, and by leveraging the recently proposed *Podify* streaming platform (see Chapter 5), I investigate the users' process of assessing the relevance of podcast content (i.e., segments). To this end, I consider factors such as search intent, task complexity, and system variations, including the integration of the text modality in the *Podify*'s UI.

### 2.5.4 Podcast Recommendation

RSs have become indispensable in predicting and offering personalised content that aligns with users' taste and preferences [62, 172, 173]. By modelling users' previous behavioural data, these systems generate tailored and personalised recommendations [122,174]. Although there is an established interest in developing RSs specific for podcasts, methodologies and approaches addressing the unique challenges and opportunities in this area remain limited [175].

Classified as speech recommenders, podcast recommender systems (PRSs) have distinctive characteristics [176]. Friends and family recommendations remain among the top three methods for discovering podcasts [170]. While research in podcast recommendations is relatively scarce, there have been some recent works in this direction. Benton et al. [177] introduced trajectory-based podcast recommendation, and they recognised the sequential nature of podcast consumption. Aziz et al. [178] focused on the underserving issue of existing PRSs, proposing the use of semantic information through knowledge graphs to enhance podcast discovery. Yang et al. [179] examined the impact

of recommendations on user content choices when related to intentions (such as subscriptions and listening). They highlight the importance of these systems, but also how they implicitly alter the users' online behaviour. Finally, McDonald et al. [180] recently explored the content exploration problem and optimising podcast recommendations for the long-term of UE.

Other recent approaches aimed to exploit textual information extracted from the audio or through conversational interfaces [181, 182]. Yang et al. [182] identified distinct user interaction patterns when recommendations are delivered via voice interfaces. They observed users tend to explore less and prefer higher-ranked items. This underscores the importance of offering diverse, personalised, and dynamically evolving top-ranked recommendations. Incorporating multiple engagement signals and of various relevance degrees is essential for enhancing PRSs [183]. Each podcast recommendation should align with specific user goals and needs. However, the problem of effectively matching new users with relevant content (i.e., the cold-start problem) remains a substantial challenge and in need of further exploration. Nazari et al. [172] delved into the efficacy of cross-domain recommendations for cold-start users using their music preferences. In [122], goal-focused consumption in RSs is examined, with the identification of low-involvement (i.e., alleviating boredom) and high-involvement (e.g., learning something new) goals. Huber et al. [184] tackled the demand for explainable recommendations. They found that highlighting content differences in podcasts through labels and summaries allows listeners to better differentiate between episodes, increasing their awareness of podcast diversity and improving the overall experience.

Current podcast information access tools, including RSs, do not adequately leverage user preferences. This gap results in listeners struggling to discover suitable podcasts for their needs [3]. In Chapter 6, I introduce a top@10 recommendation experimental design. By systematically controlling the ranking and relevance within these recommendations (i.e., "good" or "bad" recommendations), I examine the complexity of the task and how they are perceived by the participants in Chapter 7. My approach represents the first exploration of how users' assessment of podcast content relevance varies based on the recommendation's alignment with their INs.

### 2.5.5 The Multi-Modal Nature

Information access tools such as search engines and RSs play an integral role in podcast discovery and engagement [3]. However, effectively capturing the multi-modal nature of podcasts presents a significant challenge for the existing approaches [3]. Podcasts introduce unique challenges within the spoken domain due to their distinctive properties (see Section 2.5.1), such as noisy speech transcription and sentence segmentation [185]. Directly leveraging the audio signal to circumvent transcription errors is thus of considerable importance [3]. Additionally, research has also investigated the video modality of podcasts, by examining its impact on education [186, 187], and peer learning and project quality [188].

Incorporating podcasts into existing platforms has introduced new challenges for online audio streaming services, including those related to UI design [5]. A significant open research question (RQ) is determining the type and amount of information needed to guide informed decisions by users in selecting podcasts to listen to [4]. Captions, a representation of the podcast's textual modality, have emerged as an effective modality for enhancing understanding across diverse audiences [189]. Transcriptions, for instance, serve as a bridge, facilitating access to content for the hearing-impaired community and aligning with principles of Universal Design [9, 10]. They also support diverse learning styles and aid retention through theories such as Dual Coding [38, 39] and the Cognitive Theory of Multimedia Learning [41]. Further, transcriptions enable effective pattern recognition and enhance IR capabilities, aligning with the Information Foraging Theory [7, 8]. This is in line with a recent study on the TikTok platform by Mudra and Kitsa [189]. They found that videos with subtitles, akin to podcast captions, enhance information perception, attracting a broader audience and accommodating various accessibility needs [190].

Therefore, in Chapter 7, I perform an extensive investigation of the effects of incorporating the textual modality into the *Podify*'s UI. Specifically, I explore the captions and full-text transcriptions components. This analysis aims to investigate the influence of these text-based components on the users' capability to accurately identify the relevance of podcast content to their IN.

### 2.5.6 User Engagement (UE) in Podcasts

UE refers to the quality of a user's experience with an online application, encompassing the positive aspects and the inclination to use the application more frequently and for longer durations [191]. It is a multifaceted concept that includes the users' emotional, cognitive, and behavioural experiences with a technological resource, both in the short and long term [191–194]. O'Brien and Toms [195] proposed a model characterising key dimensions of UE: focused attention, aesthetics, perceived usability, endurability, novelty, and involvement. These dimensions provide a comprehensive and holistic understanding of UE across the emotional, cognitive, and behavioural facets [196]. Given its multifaceted nature, UE is usually quantified using both subjective measures (e.g., a user's self-reported perception) and objective measures (e.g., the number of mouse clicks required to complete a task) measures [192].

In the context of podcasts, Holtz et al. [197] identified an "engagement-diversity trade-off". While personalised recommendations enhance the UE, they also affect the diversity of the consumed content. Nazari et al. [183] noted the dependency of podcast types on UE patterns and the potential to tailor each podcast to specific user goals and needs. Chan-Olmsted and Wang [198], in their study of United States podcast users from the perspectives of motivation and usage, found that affective and entertainment-related motives significantly influence consumption, whereas cognitive and information-related motives lead to higher engagement with the podcast content and host (i.e. subscriptions). Additionally, an in-depth study by García-Marín [199] using semi-structured interviews identified 13 factors that determine UE. These factors can be classified into three groups, namely medium-centered (e.g., genres and formats), user-centered (e.g., perceived relevance of their participation), and podcaster-centered (e.g., tone).

In Chapter 7, I employ the subjective measures proposed by O'Brien and Toms [195] to evaluate the influence of incorporating text-based components into the *Podify*'s UI on the perceived levels of UE. Additionally, I analyse objective measures such as navigation (e.g., page changes) and listening interactions (i.e., play/pause actions and scrubbing) to provide a comprehensive assessment of UE within the *Podify* platform.

## 2.6 Chapter Summary

This chapter presented a comprehensive overview of the background knowledge, key research themes, and methodologies within the context of this thesis.

First, Section 2.2 presented an overview of online streaming services by discussing their development and current prominence. Section 2.3 then delved into an extensive exploration of the music medium, emphasising the pivotal role of user behaviour in the evolution and current research conducted within modern streaming services. In Section 2.4, we provided an extensive examination and review of music skipping behaviour. The relevance of this behaviour, previous studies conducted on the topic, and the state-of-the-art approaches employed for its prediction were discussed to provide a deeper understanding of this behaviour. Finally, Section 2.5 offered an in-depth overview of the podcast domain, recognising its growing popularity and subsequent integration into the offerings of existing music streaming services. This section also presented a comprehensive overview of podcasts' distinctive and unique characteristics, the TREC Podcast Track, and relevant recent studies on user consumption and behaviour. This also includes the intricate aspects of recommendation, multi-modal nature of this medium, and the importance of UE.

Overall, this chapter establishes the foundation for the research conducted in this thesis. Through an exploration of online audio streaming services, music behaviour, and the podcast domain, it provides the background knowledge and context to understand the subsequent discussions and findings presented in the later chapters. By highlighting the key connections between these areas and their complexities, this chapter provided an understanding of the complex landscape of online audio streaming services and the research relevance of understanding users' behaviour.

# Part II

# The Music Skipping Behaviour

# Chapter 3

# On Skipping Behaviour Types in Music Streaming Sessions

## 3.1 Introduction

As discussed in Chapter 2, Section 2.3, the ability to skip songs is a core feature in modern online audio streaming services. Its introduction has led to a new music listening paradigm and has changed the way users interact with the underlying services. Listening behaviours such as skipping and scrubbing (i.e. seeking forward and backward by moving the cursor) have recently gained research significance [13, 93]. This is because they are implicit feedback signals that can be considered as measures of users' satisfaction (dissatisfaction or lack of interest), in turn affecting the users' engagement with the platforms [14, 29, 30]. However, the modelling and understanding of these behaviours, in particular the skipping signal, in music listening sessions remain an under-researched domain [6, 13, 14, 45]. Gaining a deeper and more comprehensive understanding of this behavior is crucial as it provides valuable insights for enhancing user modeling techniques and improving the performance of underlying recommendation models.

Existing prior research has primarily focused on analysing the skip patterns as a function of the time at which this takes place within a song (skip profile) [15, 16]. In contrast to prior works, this chapter presents a comprehensive analysis of users'

skipping behaviour during entire listening sessions. The aim of this work is to obtain deeper insights and understanding into how users skip music.

In this chapter, I propose an effective data transformation and clustering-based approach to identify and categorise different types of skipping behaviour. This analysis is conducted by considering various listening contextual factors such as day type, time of the day, playlist type, account type (premium or free subscription plan), and shuffle listening mode. Additionally, I provide a thorough evaluation of the clustering performance and the skipping type identification process.

### 3.1.1 Research Motivation

Investigating music skipping behaviour provides insights into users' satisfaction and engagement with online streaming services [15]. While the users' interactions are associated with patterns which suggest preferences, disinterest, or the desire for variety, they are also considered being multifaceted, dynamic, complex and context-dependent (e.g., [200]). Previous research has acknowledged contextual information such as the day of the week [200] or playlist type [201] affect user's behaviour. However, given the rapid evolution of music streaming platforms, most of the current research considering contextual variables is outdated [98]. Additionally, the existing studies often fall short of examining the specific influence of context on music skipping behaviour. This work, therefore, seeks to fill this research gap by offering a more comprehensive understanding of how listeners skip music depending on their context. To provide contemporary insight, this research will include contextual features available in the recently released real-word MSSD [6] provided by Spotify.

### 3.1.2 Research Questions

To better understand the variations in the distribution of skipping types and the interplay with listening context information, I investigate the following five RQs:

- **RQ-3.1:** What are the main types of skipping behaviour that we can identify at a session level?

- **RQ-3.2:** For those types, how does weekday/weekend affect their overall distribution?
- **RQ-3.3:** Furthermore, do different times of the day, i.e. morning, afternoon, evening, and night, affect users' skipping interaction with the streaming service?
- **RQ-3.4:** In what ways do playlist types (e.g., personalised playlist, radio, etc.) affect users' skipping behaviour?
- **RQ-3.5:** Finally, how is the users' skipping behaviour influenced by account type (premium or free subscription plan) and when listening is performed in a shuffle mode?

I investigate my RQs by analysing a large real-world music streaming dataset (i.e., Spotify) and on sessions of varying length, thereby providing generalisation to my findings.

### 3.1.3 Contributions

The contributions of this chapter are four-fold:

- I perform an extensive investigation on users' skipping behaviour during entire listening sessions.
- I propose an approach that can identify any number of fine-grained session skipping behaviours.
- By extensive evaluation, I identify four to be the optimal number of main session skipping behaviours.
- Finally, I investigate the influence that various listening contexts have on the distribution of these four patterns.

This chapter is organised as follows. First, I provide an overview of my approach in Section 3.2. Then, the settings of the analysis are outlined in Section 3.3. The results are presented in Section 3.4, followed by a summary and discussion of the chapter's main findings in Section 3.5.

## 3.2 Approach

In this section, I present my approach to analyse the users' skipping activity and identify the music skipping types in listening sessions.

### 3.2.1 Session Skipping Pattern Extraction

In order to analyse and identify skipping types, I first need to extract the skipping activity across entire listening sessions. I refer to such patterns as the *session skipping patterns*. The skip features available in the selected dataset are *skip_1*, *skip_2*, *skip_3*, and *not_skipped*. They are predefined features provided as part of the dataset's original feature set and they represent defined thresholds that respectively indicate whether the track has been played very briefly, briefly, mostly, or in full. In Table 3.1, a summary of the interplay among these features at record-level is presented, together with an integer transformation (ID) for every observed pattern. This is motivated by the easing of result reporting in later sections. It is important to note that a rarely occurring pattern, namely "$False, True, False, False$", is disregarded in my analysis since I believe it to be a logging error. Additionally, sessions with missing values are also excluded from my analysis. The exclusion of these sessions is motivated by the abundance of complete listening sessions (i.e., without missing values) that are available for analysis. Therefore, to avoid the introduction of potential noise in my analysis through imputation of missing data, these sessions are excluded.

With a scalar representation of the skipping activity for every record of a session, I can now construct a session-level vector representation. This is the ordered sequence by session position of all individual song skipping activities that form a session. Thus, for a session of length 20, a vector of 20 elements captures the *session skipping pattern*. For example, "$1, 1, 2, 1, 5, 3, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 2, 5, 3, 5$", indicates that in this session a user played very very briefly the first two songs, followed by very briefly and very very briefly for the 3rd and 4th songs respectively, before the 5th song was played in full.

Table 3.1: Summary of skip patterns and their corresponding translation in terms of for how long the current track was played. ID is an integer value associated with each pattern, used in the construction of session skipping patterns.

| skip_1 | skip_2 | skip_3 | not_skipped | Listening Length | ID |
|---|---|---|---|---|---|
| True | True | True | False | Very Very Briefly | 1 |
| False | True | True | False | Very Briefly | 2 |
| False | False | True | False | Briefly | 3 |
| False | False | False | False | Most | 4 |
| False | False | False | True | All | 5 |

### 3.2.2 Session Skipping Type Identification

Given a set of session skipping patterns, I apply Principal Component Analysis (PCA) [202] to their vector representations. This is motivated by the fact that the input space contains noisy patterns as well as that by reducing the dimensionality of such large sets I decrease the computational requirements. Subsequently, the transformed data, i.e. the PCA components, are used as the input space for k-means clustering [203], which uses Euclidean distance to define cluster centroids. In evaluating the clustering algorithms for these large sets, alternatives, such as hierarchical clustering [204] and DBSCAN [205], were also explored. However, based on a local evaluation, the k-means was identified as the most suitable clustering algorithm, given its computational efficiency and effectiveness in identifying skipping patterns. I also adopted the k-means++ initialisation scheme, known to augment speed and quality of the clusters [206].

In this section, my proposed approach to identify skipping types is presented. The skip features in the selected dataset are analysed, and the skipping activity is transformed to a session-level vector representation (i.e., *session skipping pattern*). To reduce the noise and dimensionality of the large number of *session skipping patterns*, PCA is applied, followed by k-means clustering, to identify the skipping types.

## 3.3 Analytical Settings

In this section, I describe the settings of the analysis that support my proposed approach for the identification of different behaviours during entire listening sessions with regards to users' session-based skipping activity.

### 3.3.1 Dataset

The analyses are conducted using the publicly available MSSD [6] provided by Spotify. The MSSD comprises of approximately 150 million logged streaming sessions, spanning a period of 66 days from July 15th to September 18th 2018. Each day comprises of ten logs, where each log includes streaming listening sessions uniformly sampled at random throughout the day. Sessions are defined as sequences of songs or tracks that users have listened to, ranging from 10 to 20 records per session (one record per song). The dataset encompasses various types of contextual information about the stream (e.g., the playlist type) and interaction history with the platform (e.g., scrubbing, which is the number of seek forward/back within the track). Additionally, despite track titles not being available, descriptive audio features and metadata are provided to describe the tracks (e.g., acousticness, valence, and year of release). The available implicit feedback is captured through actions such as song skipping, pauses during playback, and user-initiated actions that lead to the start or end of a particular track. It is important to note that the MSSD does not include user identification, demographic information, or geographical data. Therefore, it is not possible to determine if two sessions belong to the same user or different users. Finally, I perform my analysis on the complete training set. The testing set is not used since most of the metadata as well as the skipping attributes, are missing.

### 3.3.2 Conditions of Interest

After removal of sessions with unrecognised skipping activity, as described in Section 3.2.1, of the resulting 125 million listening sessions, the majority (47.7%) corresponds to sessions of length 20. The remainder are distributed in decreasing order from sessions of

length 10 (8.8%) to those of length 19 (2.8%). Given this, I perform my primary analysis on long sessions (length 20). However, to show the adaptability and extendibility of my approach to sessions of varying length, I also generalise my results by including data from medium (length 15) and short (length 10) sessions. Finally, further analysis is also performed on all session lengths, to further validate the validity and generalisability of my results.

To answer my RQs, the following contextual scenarios are formulated:

- **Weekdays and Weekends.** The 66 days of training data can be grouped into 9 full weeks, hence 45 weekdays and 18 weekend days. The 10th Week is only partially available, with only 3 available days, i.e. 16th, 17th, and 18th of September. For a fair comparison those dates are therefore ignored when answering **RQ-3.2**.

- **Time of The Day.** I define five time windows, where numbers in brackets correspond to the hour range: night (0-5), morning (6-11), afternoon (12-17), evening (18-23), all (0-23). If a session spans across two hours, I round and consider the whole session as either part of start or end hour.

- **Playlist Type.** The playlist type is defined as the type of playlist that the playback occurred within. It includes editorial playlist, user collection, catalog, radio, charts, and personalized playlist. As this can be subject to change throughout a session (playlist switch), for the scope of this condition (**RQ-3.4**) I restrict my analysis only to sessions with a single playlist type.

- **Account Type.** The account type refers to the type of subscription plan that the current user is currently subscribed to. It distinguishes between the premium and free subscription plans.

- **Shuffle Mode.** The use of shuffle listening mode can vary throughout a session, as users may enable or disable it at any time. For the purposes of this study, a session is considered being a high-shuffle session if at least 80% of its tracks were played in shuffle mode. Conversely, a session is considered being a low-shuffle session if at most 20% of its tracks were played in shuffle mode.

- **All.** This final scenario refers to an analysis that is performed on all available days and their corresponding listening sessions. In this scenario, all listening sessions are considered, with no exclusions, such as the ones based on weekdays and weekends, time of the day, playlist type, account type, or shuffle mode.

### 3.3.3 Procedure

#### PCA Components Selection

A common, but subjective, cut-off point of the total explained variability by PCA is 70% [207]. In order to find the optimal variance to retain, I performed local evaluation on the number of PCA components. Increasing the retained variance and/or applying clustering on the raw data yields similar results. By decreasing the variance, only the outliers are less represented in the PC space. Since my goal is to find the main behavioural types, the difference is minimal given that the highly relevant patterns are well described in the Principal Component (PC) space. In order to retain 70% of the explained variance, the results in this chapter are based on using 5 PCA components for sessions with a length in the range $[10-12]$, 6 for the range $[13-17]$, and 7 for $[18-20]$.

#### Clustering Performance Evaluation

The overall aim of clustering is to find a partitioning scheme that best fits the underlying data. One of the most important issues related to evaluating the performance of the resulting clusters is cluster analysis and validity. An "optimal" clustering scheme is defined as delivering an outcome by running a clustering algorithm that best fits the inherent partitions of the dataset [208].

Cluster Validity Indices (CVIs) are used for both estimating the quality of a clustering algorithm and for determining the optimal number of clusters in the data [209]. However, existing measures can be affected by various data characteristics. In the work by Liu et al. [210], it is noted that the optimal number of clusters can be greatly affected by various factors. For example, the noise in the data is shown to affect the Calinski-Harabasz (CH) Index [211], with the clusters' proximity affecting the Davies-

Bouldin (DB) Index [212]. Each individual CVI is designed to capture a specific aspect of the partitioning scheme that attempts to indicate how adequate it is. On the other hand, this inevitably results in other aspects to be inadequately represented or ignored altogether [213]. Thus, no CVI can *a-priori* be assumed to be better than its alternatives [214]. Despite being an active research domain for many years and for a variety of disciplines, several important RQs remain open as to how best to assess the quality and validity of a clustering procedure [208].

In this work, I adopt the CH and DB indexes, together with the traditional Elbow Method on the Inertia (INE) values (i.e., the sum of squared distances of samples to their closest cluster centre) and self-judgement to arbitrarily decide on the optimal number of clusters for the MSSD dataset. The Silhouette Coefficient [215], although recognised as one of the most widely adopted CVIs, is not used in this work due to the large population size of the clusters and related computation impracticability.

***CH Index***, also known as the Variance Ration Criterion, is the ratio of the sum of between cluster scatter matrix (BCSM) and of within cluster scatter matrix (WCSM) for all clusters. This index is highly dependent on a separation measure between clusters and a measure for compactness of clusters based on distance. By maximising the BCSM and minimising the WCSM, I obtain well separated and compact clusters. For a dataset $E$ of size $n_E$ and $k$ clusters, the *CH* score (where a higher score indicates better defined clusters) is defined as:

$$CH_k = \frac{BCSM}{k-1} \times \frac{n_E - k}{WCSM} \tag{3.1}$$

The *BCSM* is based on the distance between clusters and is defined as:

$$BCSM = \sum_{i=1}^{k} n_i \times d(z_i, z_{tot})^2 \tag{3.2}$$

where $z_i$ and $n_i$ are respectively the center and number of points in $c_i$. The *WCSM* is defined as:

$$WCSM = \sum_{i=1}^{k} \sum_{x \in c_i} d(x, z_i)^2 \tag{3.3}$$

where $x$ is a data point belonging to cluster $c_i$.

***DB Index*** can be used to evaluate a partitioning scheme by computing the similarity between clusters. It is a measure that compares the distance between clusters with their respective sizes. The index identifies clusters which are far from each other and compact, and is defined as (with a value closer to zero indicating a better partition):

$$DB_k = \frac{1}{k}\sum_{i=1}^{k} \max_{j=1,\ldots,k,i\neq j} \left\{ \frac{diam(c_i) + diam(c_j)}{d(z_i, z_j)} \right\} \tag{3.4}$$

with the diameter of a cluster defined as:

$$diam(c_i) = \sqrt{\frac{1}{n_i}\sum_{x\in c_i} d(x, z_i)^2} \tag{3.5}$$

where $z_i$ and $n_i$ are respectively the centroid and the number of points in $c_i$.

***INE Index*** is another measure that can be used to evaluate the quality of a partitioning scheme by how internally coherent the clusters are. It is oftentimes used in conjunction with the Elbow method in order to find the optimal number of clusters to partition a population. The inertia, or within-cluster sum-of-squares criterion (WCSS), is defined as:

$$WCSS_k = \sum_{i=0}^{n} \min_{z_j\in C}(\parallel x_i - z_j \parallel^2) \tag{3.6}$$

where $z_j$ is the centroid of the cluster.

### Comparison of Identified Types

To allow comparison of the identified types for specific session length and different conditions, I perform cluster matching so as to pair highly similar clusters from one analysis (e.g. morning) to another (e.g. afternoon). For every cluster, I group all of its session skipping patterns and produce a single averaged vector representation. That is, every position in the session is the average for that position of all session skipping patterns. I call this single averaged vector representation the *average skipping session* for that cluster. Clusters matching is then performed via pairwise Euclidean distance. Each average skipping session from one analysis is matched to its closest cluster in another analysis. I find Euclidean distance to be the most suitable metric for this task.

Other metrics, such as cosine similarity, generated inconsistencies and mismatches when applied on the average skipping sessions, even when the averages were computed on the PCA components rather than session positions.

Local evaluation on the number of clusters (in the range 2 to 40) and on different runs suggested high similarity and stability of the results. Hence, the reporting of results is for seed number 0. For long, medium, and short sessions, the highest observed Euclidean distances among seeds for all reported results are of 0.2259, 0.2363, and 1.1756 respectively. For complete implementation details and for reproducibility of my work, I refer the readers to the publicly available code at https://github.com/NeuraSearch/Spotify-Session-Skipping-Behaviour.

## 3.4 Results

In this section, I first provide an empirical analysis on the clustering performance and how I empirically found four to be the optimal number of clusters. To explore my **RQ-3.1** further, I then investigate the properties of these four dominant skipping types at a session level and for long, medium, and short sessions. Finally, an analysis and discussion on the effect that various listening context information have on their distributions (**RQ-3.2-5**) is reported.

### 3.4.1 Analysis on Clustering Performance

In order to identify the optimal number of clusters to partition the population, which depends on the selected condition of interest (Section 3.3.2), I perform an empirical study on the clustering process from a CVIs perspective.

In Figure 3.1 I report the CH, DB, and INE values for the "All" analysis and on all available session lengths. Notwithstanding an order of magnitude higher in the CH and INE values for sessions of length 20, I observe similar patterns across all indices, session lengths, and number of clusters. This suggests high stability and similarity of the clustering process, which is independent of session length and number of clusters. As reported in Section 3.3.2, the majority of listening sessions, 47.7%, correspond to

sessions of length 20. Therefore, I note the order of magnitude higher in the CH and INE scores for sessions of length 20 to be due to the differences in size of the individual sets of session skipping patterns.

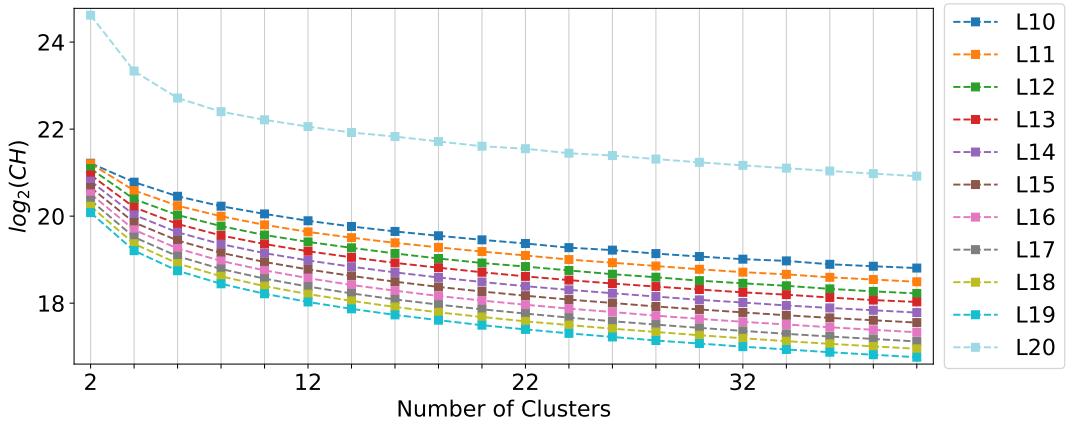
In a closer investigation, we observe the CH index suggests an optimal solution to be achieved with a low number of clusters. A $k$ of 2 yields the highest score, hence the most well separated and compact clusters. A value of 4 also yields a relatively high score, with all subsequent values that are considered being sub-optimal solutions. In the case of the DB index, where a value closer to zero indicates a better partitioning, We note how again 2 or 4 appear to be optimal. Interestingly, we can observe a common intersection point among most session lengths at $k = 4$. It is the point of lowest variance on the index values, and this further proves the stability of my proposed approach. Finally, a traditional elbow method on the IN values would suggest an optimal partitioning scheme to be achieved with $k < 8$.

Overall, these empirical results on clustering performance strongly suggest that an optimal partitioning scheme is to be achieved with a small number of clusters. All indices seem to agree on $k = 4$ to be among the best options. In the next section, I discuss the identification and classification of these four clusters (skipping types), and then further evaluate the performance from a human self-judgement point of view.

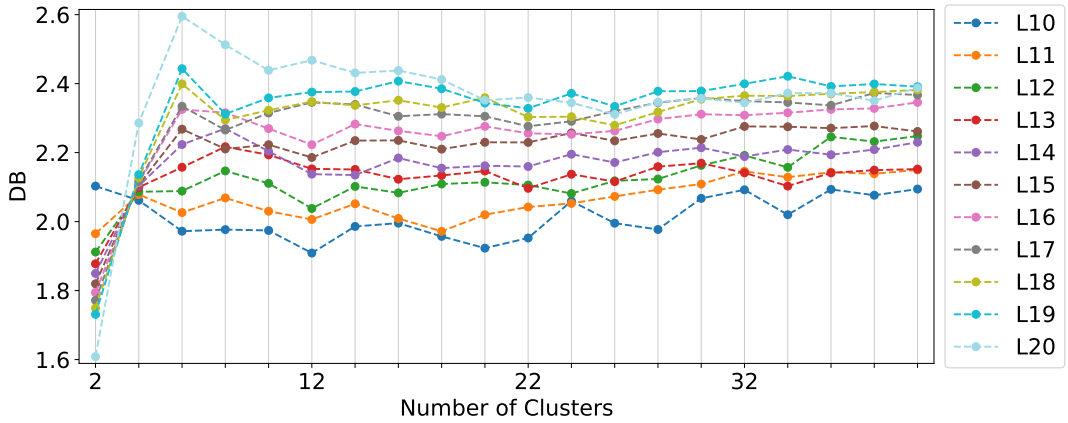
### 3.4.2 Types Identification

In Figure 3.2, the average skipping sessions and their corresponding type name (i.e., *listener*, *listen-then-skip*, *skip-then-listen*, *skipper*) for a "All" analysis is reported. As can be observed, the four identified types appear to be consistent across sessions of different length. This suggests that such dominant behaviours have no strong relation to the absolute length of a session and are thus generalisable. Additionally, a closer examination reveals that in fact there are two main distinctive behaviours: *listener* and *listen-then-skip*. The remaining two types, i.e. *skipper* and *skip-then-listen*, can be seen as their respective complimentary behaviour (addressing **RQ-3.1**).

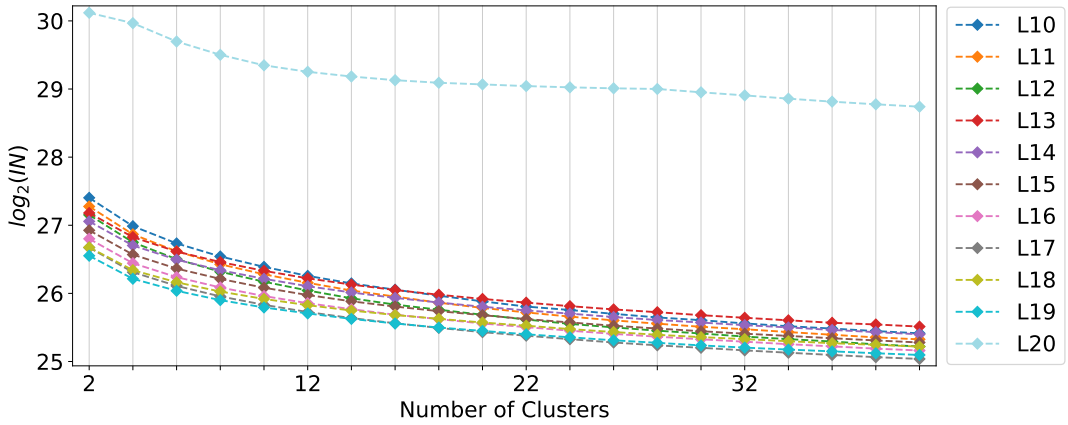
Intuitively, categorising all listening sessions in four types yields a high variance, as shown in Figure 3.2. This is because each type agglomerates sessions that are




(a) Calinski-Harabasz (CH)

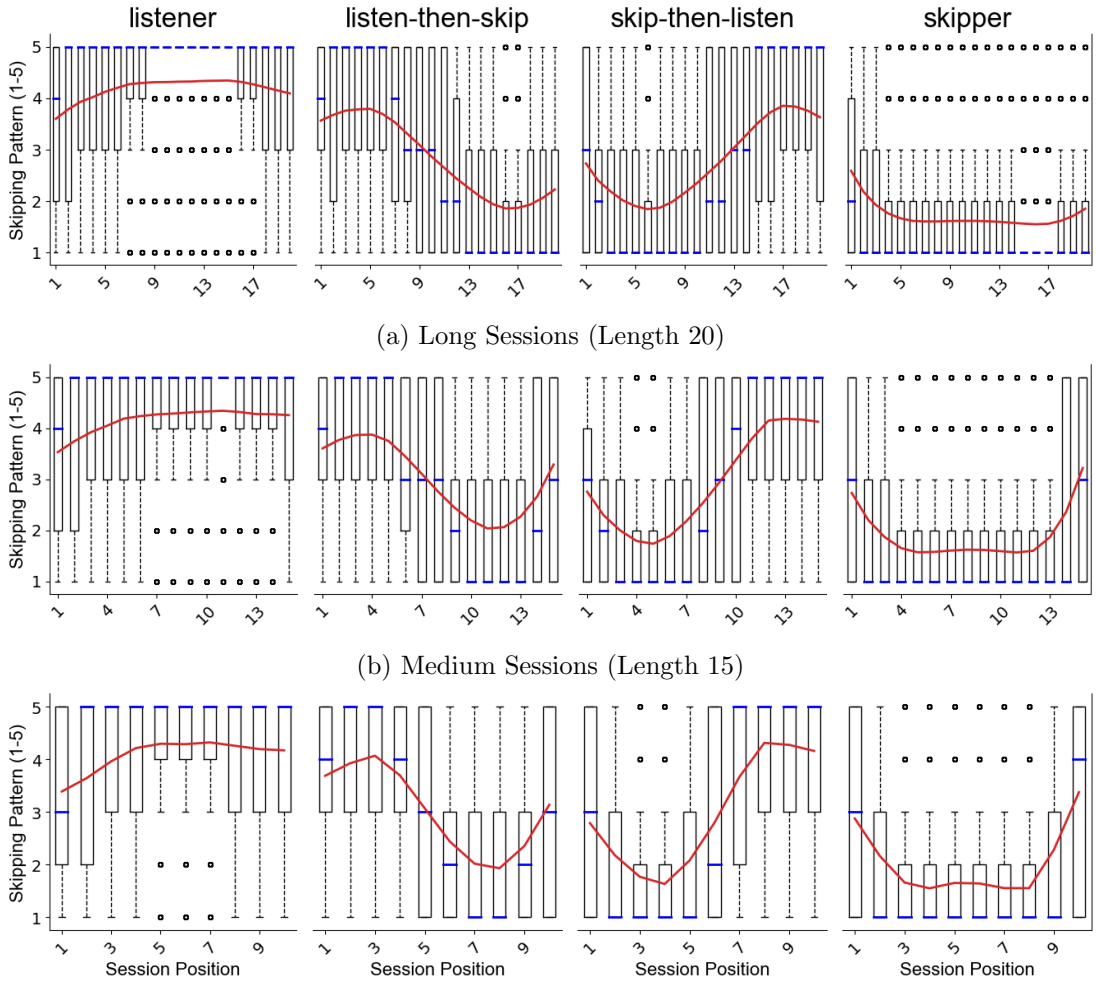



(b) Davies-Bouldin (DB)




(c) Inertia (INE)

Figure 3.1: CH (3.1a), DB (3.1b), and INE (3.1c) values for a "All" analysis (see Section 3.3.2) and all available session lengths. The x-axis is the number of clusters ([2..40] with a step size of 2). The y-axis represents the value for each index, with the CH and INE indexes being transformed using the $log_2$ to ease the presentation.




(a) Long Sessions (Length 20)

(b) Medium Sessions (Length 15)

(c) Short Sessions (Length 10)

Figure 3.2: Box plots of the identified skipping types for different lengths and for a "All" analysis (see Section 3.3.2). The x-axis is the range of session positions $[1..n]$, where $n$ is 20, 15, or 10, depending on the selected length. The y-axis represents the skipping patterns (ID 1-5 in Table 3.1). The red line indicates the average skipping session.

considered similar but yet contain a noisy sequence of session skipping patterns. For a *skipper* type, an example is a pattern that contains all ID1s except for a middle ID5. These listening sessions can arguably be a different, more refined and fine-grained, behaviour and not necessarily a member of *skipper*. Those branches of sub-behaviours with lower variance emerge with a higher number of clusters. This phenomenon can be clearly observed in Figure 3.3, which reports an "All" analysis conducted with a partitioning scheme of 20 clusters. We can note how some of these refined types are the transposed representation of other types. That is, the ID5 shift of the previous example can happen in any of the 20 positions. This shifting on the x-axis can be clearly observed in the sequence of clusters #2, #8, #6, #15, #16, #10, and #13. Therefore, this can result in a likewise number of different behaviours, where the resulting bell shape is shifted along the session's position. Ultimately, however, they can arguably represent a single type of skipping behaviour which is position-independent. This is represented by cluster #13, in the case of $k = 20$ of Figure 3.3, or by the *skipper* type in $k = 4$ of Figure 3.2. The former has inherently a lower variance than the latter. This is because, as $k$ increases, the noisy sequences of session skipping patterns that have been included in the main dominant skipping types are extracted and filtered out. These sessions are grouped together to then form new and more refined skipping behaviours, hence the total variance of all individual types is decreased as $k$ increases. Finally, by refining types, there also emerges the behaviour type that consists of the continuous alternation of skipping and no skipping. This yields an average skipping session similar to a cosine function (#5 and #12), which is believed to be capturing the noisy sequences of the middle skipping patterns (ID2-4).

However, as I perform refining, I note divergence in common behaviour across analyses. Whereas with a low number of types, the clusters matching has high accuracy, with a higher number, the comparisons might be performed on different behaviours. This is due to the identification of ungeneralisable behaviours that tend to differ across analyses. Since the goal of this chapter, as stated in **RQ-3.1**, is to identify the main generalisable behaviours and their variations under different listening context information rather than their dissimilarities, in the next section I report results for the analysis

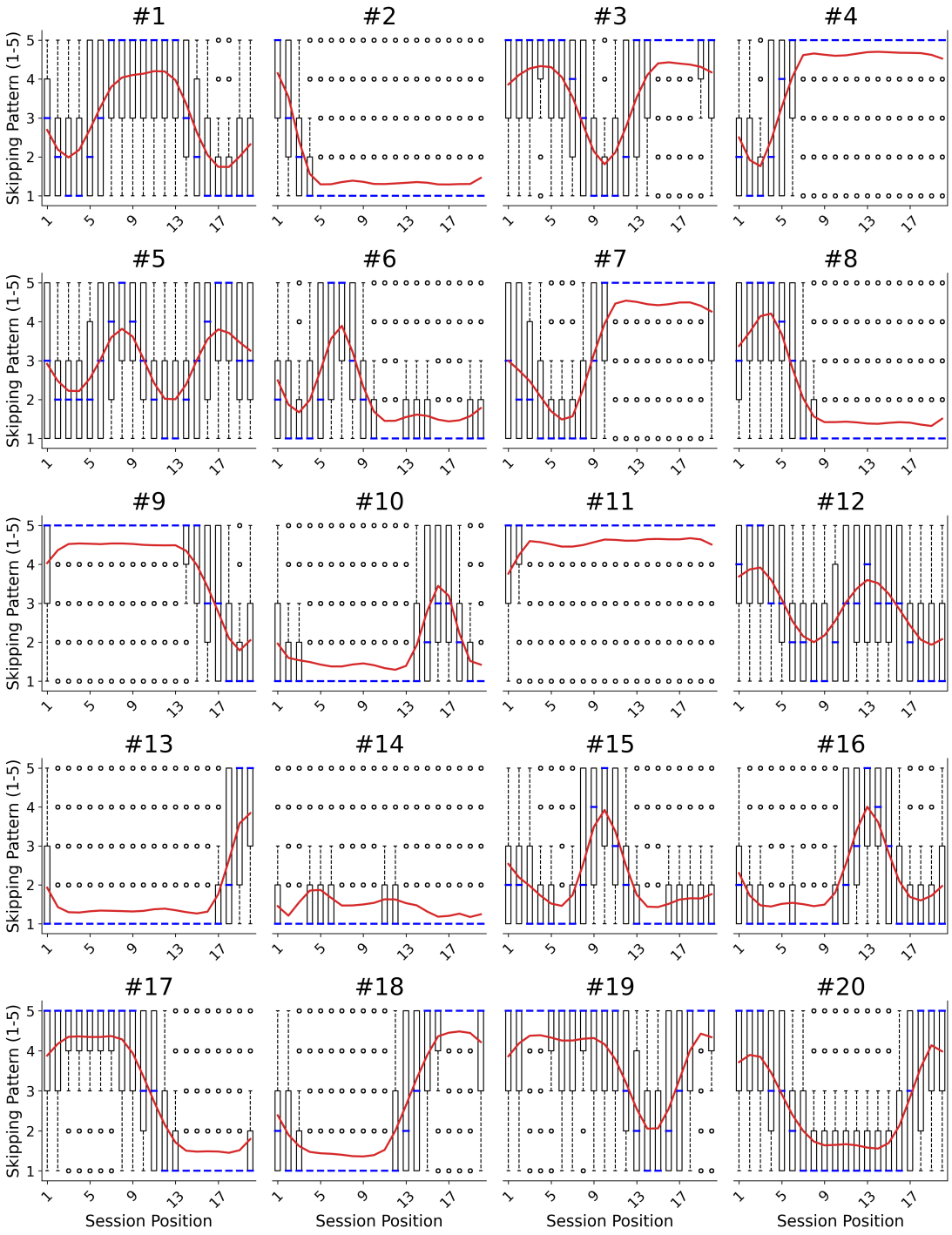



Figure 3.3: Box plots of the skipping types for a "All" analysis (see Section 3.3.2), on long sessions, and with 20 clusters. The x-axis is the range of session positions [1..20]. The y-axis represents the skipping patterns (ID 1-5 in Table 3.1). The red line indicates the average skipping session.

on the four types previously outlined.

### 3.4.3 Distribution Differences

#### Proportions on Individual Session Lengths

We observe from Figure 3.4 that there are no marked differences in distribution across weekdays and weekends (addressing **RQ-3.2**). However, given Spotify's global user distribution and cultural differences around the definition of the "weekend", it may be unsurprising that no differences were observed. Further inclusion of geographical and time-of-day variables may be required to properly evaluate any "weekend" effect.

Differences emerge when I narrow down my analysis to the different time of the day. Morning users have the lowest *skipper* activity. In contrast, such behaviour is always at its highest in the evening. This suggests that the amount of *skipper* activity increases as we progress through the day. I hypothesise this to be due to morning users paying higher attention to their daily tasks rather than the songs currently playing. In contrast, evening users may have more spare time, thus allowing them to pay closer attention to the streamed content (addressing **RQ-3.3**).

To address **RQ-3.4**, we can immediately note how the *skipper* activity is consistently lower for catalog and personalised playlists. This steep decrease in comparison with other types is highly prevalent in long and medium sessions but less pronounced for short sessions. It would also appear to be the case that the user collection, radio, and charts categories are associated with the highest degree of skipping activity regardless of session length. These findings partially meet my expectations. Catalog users have higher control over music selection than charts. Surprisingly, charts have consistently been one of the highest *skipper* activity, despite being a collection of the most popular titles of the moment. This suggests that music popularity impacts skipping behaviour. Last, for the two under-represented types, namely *listen-then-skip* and *skip-then-listen*, we note how the former has always a higher activity, independently of the absolute session length. This suggests that it is more likely for users to listen to the first half and skip the second rather than the opposite. In the case of the radio, such under-represented types are the lowest of all playlist types. This suggests that radio

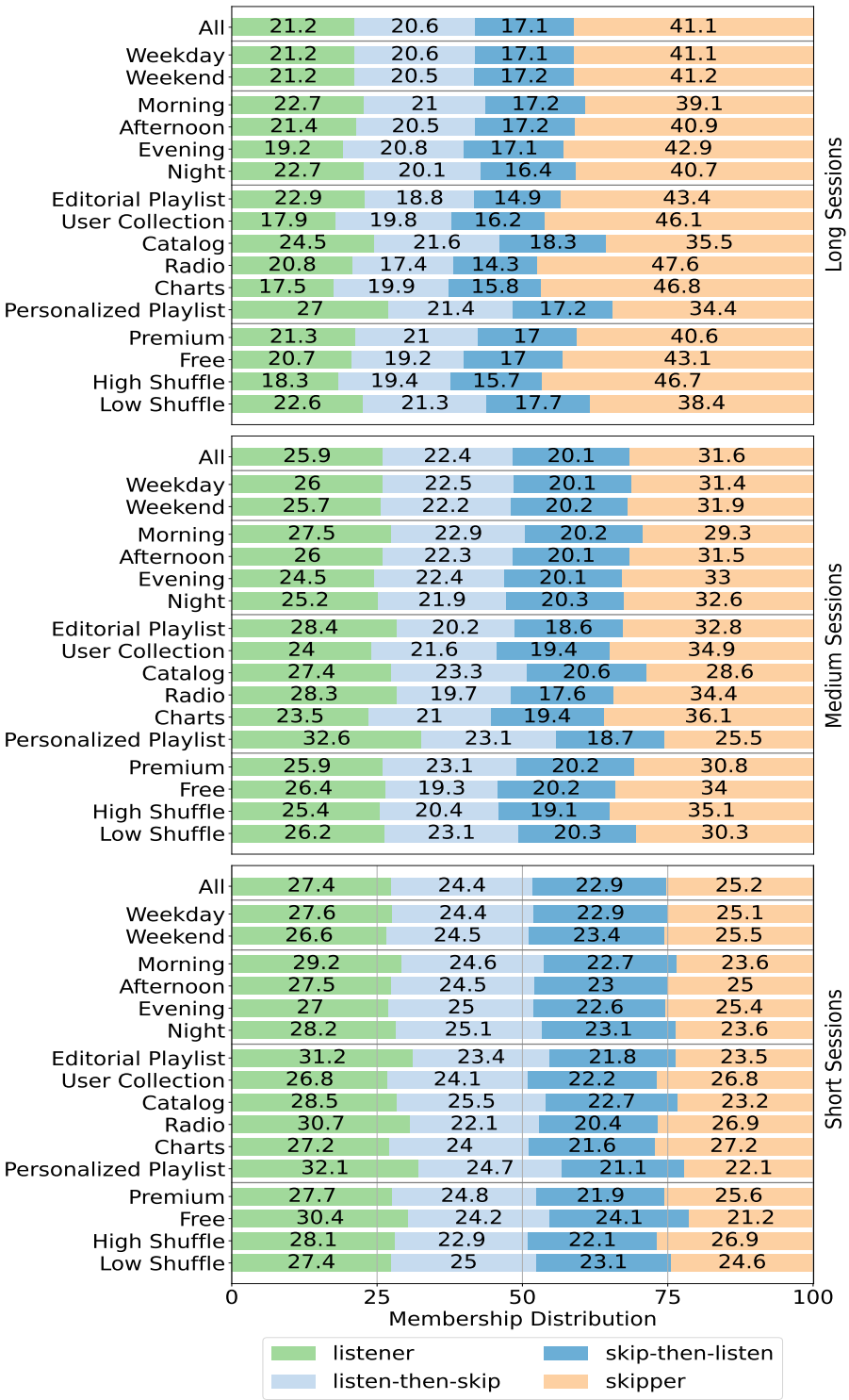



Figure 3.4: Distribution of types under the different analytical settings described in Section 3.3.2 and for long (*top*), medium (*centre*), and short (*bottom*) sessions. The x-axis represents the membership distribution in percentage value for each type.

users vary less their skipping patterns, thus consistently be a *listener* or *skipper* type.

Finally, how does account type and shuffle mode affect users' skipping behaviour (**RQ-3.5**)? With regards to the account type, we can observe how there is an inverse trend by session length in the distribution of the main types between premium and free account type. In long sessions, premium users have a higher *listener* activity, with free users being more of a *skipper* type. As sessions become shorter, we observe how the free users are now associated with the highest degree of *listeners* in medium and eventually also the lowest degree of *skippers* in short sessions. In the short sessions, free users are associated with the highest *listeners* and least *skippers* activity. These findings partially meet my expectations. It is intuitive to think that premium users should have high listening rates because of their enhanced ability to browse and instantly play any song on-demand. With unlimited skips, we should also expect relatively balanced *skipper* ratios across session lengths. On the other hand, since free users have a limited number of skips per hour, their use may be more sporadic and controlled. In 2018, when the data was collected, the Spotify Free Plan had different limitations depending on the platform used by the user (e.g., desktop, tablet, or mobile device). Furthermore, users may also select playlists more accurately to avoid unwanted recommendations and unnecessary "spending" of the skips. All in all, we should expect high *listener* and low *skipper* rates for free users. However, it is clear from my analysis that we observe contrasting results. The proportion of *skippers* in long sessions is higher for free users, and the proportions among the four types for premium users substantially differ. My analysis suggests that free users, despite the limitations of their subscription plan, highly interact with the skipping functionality of the platform, and that a premium subscription plan has no strong correlation to the users' session-based skipping activity.

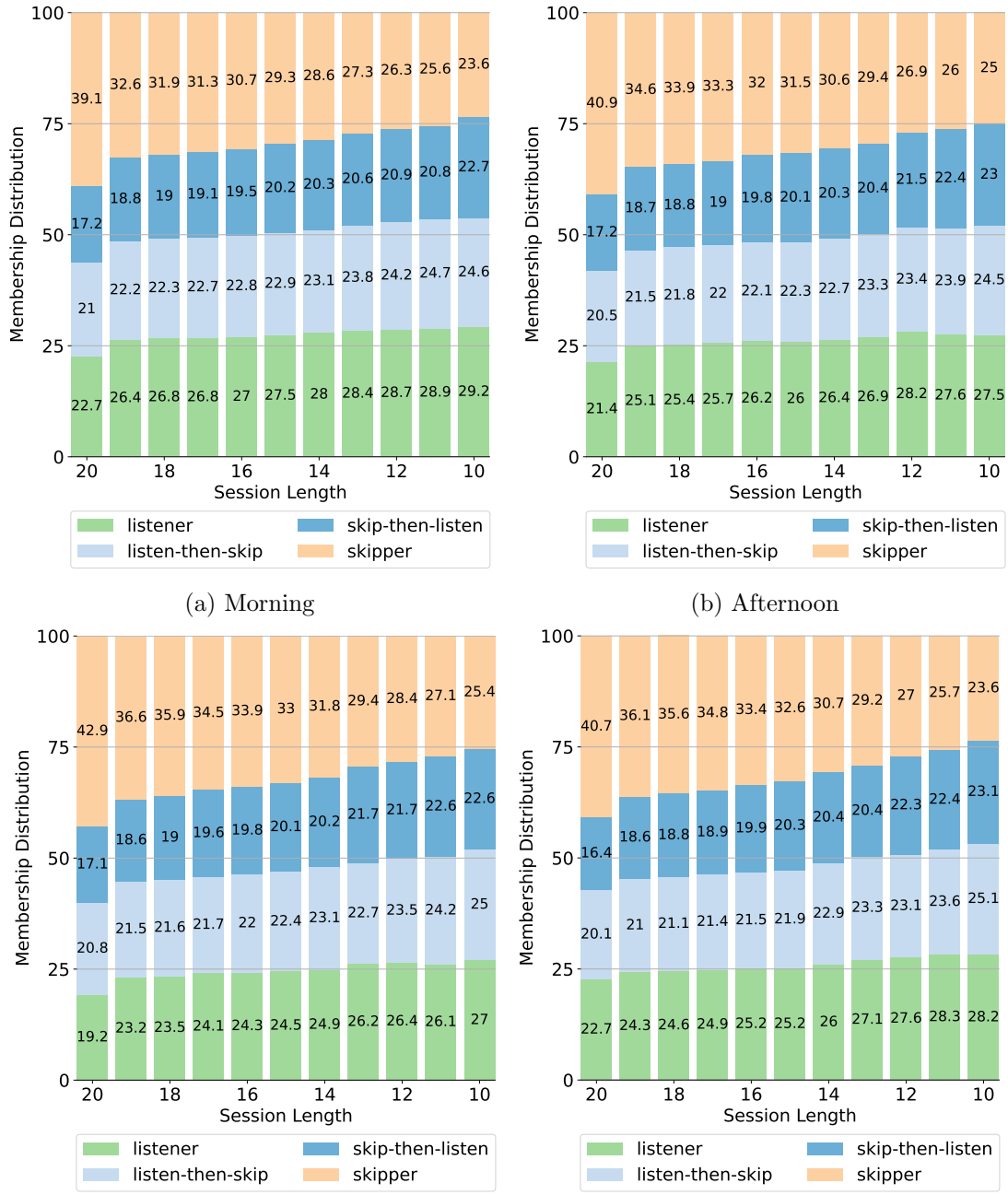
Regarding the shuffle listening mode, it may be unsurprising to see that it consistently leads to higher skipping rates. We can immediately note how there is a notably higher *skipper* activity when the shuffle levels are high. This listening mode appears to be negatively affecting the overall listening activity by leading users to skip more frequently, which in turn could impact on the UE and satisfaction. This is also usually reflected in lower *listener* activity compared to when there is low shuffle activity.

However, although it being the case for long and medium sessions, we observe that in short sessions a high shuffle level is instead, and perhaps surprisingly, associated with an overall higher *listener* activity. In particular, higher proportions can be seen on *listener* and *skipper*, with a consequent decrease in the *listen-then-skip* and *skip-then-listen* types. High shuffle users are reported with less varied skipping patterns, thus consistently be a *listener* or *skipper* type. This may suggest that the shuffle mode may or may not be enjoyed by the users.

### Proportions For Different Session Lengths

Major differences can also be seen when looking at the distribution proportions for different session lengths. In the example of the "Weekday" analysis, we note how that the proportion of *skippers* drops from a 41.1% in long sessions, to 31.4% in medium, and 25.1% in short. Conversely, the proportion associated with *listener* type increases from 21.2% to 26%, eventually reaching 27.6% in short sessions. Overall, as sessions become shorter, we see a decrease in *skipper* type behaviour and an increase of the other three types. If the *skipper* type prevails in long and medium sessions, we note how, for short sessions, the gaps are reduced, with *listener* becoming the highest representative behaviour. In other words, under my analytical settings and with the provided data, we observe that, in comparison with long listening sessions, in short sessions, users increase their listening activity and thus skip less frequently.

This finding can also be observed and further validated by Figure 3.5, which reports the types distribution for the "Time of The Day" analysis and on all available session lengths (i.e., [20..10]). The proportion of *skippers* always decreases as sessions become shorter. This suggests that users tend to be more of a *skipper* type as a streaming sessions gets prolonged over time. Conversely, when the listening sessions become shorter, the *listener* type prevails, with users increasing their listening activity and thus skip less frequently. On average, we observe an overall mean increase in the range of sessions from long to short of 2.8%, 1.9%, 3.1%, and $-5\%$ for *listener*, *listen-then-skip*, *skip-then-listen*, and *skipper* types, respectively. Across all the "Time of The Day" listening contexts, we note how the mean average difference between *listener* and




(a) Morning

(b) Afternoon

(c) Evening

(d) Night

Figure 3.5: Distribution of types under a "Time of the Day" (see Section 3.3.2) analysis and for all session lengths. The x-axis represents the various session lengths ([20..10]), with the y-axis representing the membership distribution in percentage value for each type.

*skipper* increases from a $-47.3\%$ for long sessions to $14.9\%$ for short sessions.

Overall, I note how the users' listening activity differs between long and short sessions. This could be associated with various factors, including the recommendation procedure, which may have failed to correctly estimate the users' true intentions and musical preferences. This is reflected in a higher *skipper* activity as sessions become longer. With these being initial hypotheses as to why we observe such large differences in the users' session-based skipping behaviour, more in-depth analysis is required in order to better understand and explain the factors that lead to such differences in users' behaviour. This is left as future work and further consideration.

## 3.5 Chapter Summary

Music skipping behaviour provides implicit feedback that can measure users' satisfaction and, therefore, the engagement of users with a streaming platform. An accurate representation of this behaviour would allow us to model the users' degree of interest and engagement with the platform. Thus, understanding session skipping patterns may play an important role in today's systems. Prior work toward understanding and modelling users' skipping behaviour in music listening sessions has mainly focused on analysing the skipping activity at an individual song level [15, 16].

Inspired by the research relevance around the music skipping behaviour, in this chapter I tackled the task of identifying different behaviours during entire listening sessions with regards to the users' session-based skipping activity. This analysis is performed on a large real-world dataset of music streaming listening sessions (MSSD). By adopting an effective data transformation and clustering-based approach (see Section 3.2), I identified four main skipping types, namely *listener*, *listen-then-skip*, *skip-then-listen*, and *skipper* (see Section 3.4.2 and Figure 3.2). Further analysis at both clustering performance level, using Cluster Validity Indices (CVIs) (see Section 3.4.1 and Figure 3.1), and on the partitioning schemes when varying the number of clusters (Section 3.4.2) indicates four to be the optimal number. These four main skipping types provide a generalised form of all the refined and fine-grained behaviours that emerge as

the number of clusters increases. Hence, they can be defined as the main and dominant skipping types. This is because they represent, despite having an inherently higher variance, the different facets and properties of users' session-based skipping behaviour that a music streaming platform can observe.

In my analysis, I observed the following key findings:

- The four types are consistent across sessions of varying length. Thus, they represent a generalisable and common behaviour that has no strong relation with the absolute length of a session (see Section 3.4.2).
- A closer examination of these types reveals that these four types exhibit two pairs of complementary behaviours: *listener* and *skipper*, as well as *listen-then-skip* and *skip-then-listen*. This is because, the behaviours of, for example, a *listener* and a *skipper* can be seen as contrasting or opposing, making them complementary to each other (see Section 3.4.2).
- The distribution of skipping types varies under different listening context information. With the exception of "Day Type" (i.e., weekday and weekend), which I note as being caused by the Spotify's global user distribution, cultural differences around the definition of the "weekend", and lack of available demographic information in the data, I observe marked differences in all the other scenarios (see Section 3.4.3).
- Finally, notable distributional differences can also be seen when varying the session lengths. I observe that, as sessions become shorter, users tend to increase their listening activity and thus skip less frequently (see Section 3.4.3).

In particular, I observe how the users' listening activity differs when varying the length of the sessions. I hypothesise this phenomenon could be explained and associated by how users interact with the streaming platform. When engaging for a relatively short amount of time (i.e., short sessions), users may have a more active involvement with the streaming platform. Songs may be individually picked and more accurately selected. Therefore, the UE and related satisfaction with the platform is higher. For

example, the streamed media may not be used as a background activity. Instead, the streaming process is highly controlled and interacted with by the user. Users and the underlying RS optimise the selection criteria for the users' current needs in order to achieve instant satisfaction. This is commonly referred to as the short-term optimisation of user satisfaction. On the other hand, as sessions become longer, I observe an inverse trend, with the *skippers* activity consistently increasing and with the *listener* type consequently decreasing. This could be justified by the total length of the listening session. A session of length 20 corresponds to at least an hour of continuous listening (if I assume full listening and no skipping). It is common in such cases that users listen to music in the background. Thus, they may be less inclined to accurately select songs of their liking. Rather, they may select large, and possibly mixed, playlists since their main attentional focus may not be on the music but rather on other tasks, such as driving or working. In this case, users may engage with a large playlist that contains a vast collection of songs. Such playlists, however, may have too much diversity in, for example, the music genres. In the case of RSs, a sub-optimal procedure may be associated with a misjudgement of the users' needs. This is commonly referred to as the long-term optimisation of user satisfaction, which is an open and current research direction.

The relevance of the extracted skipping types for potential downstream tasks is evident. A better understanding, and thus modelling, of this feedback signal could help in devising new recommendation and search processes. By better representing the user and their preferences, we can consequently observe an improved personalisation and tailoring of the results to a given user. In Chapter 4, I delve deeper into the dynamics of music skipping behaviour by studying the utility of users' historical data for the task of sequentially predicting users' skipping behaviour.

# Chapter 4

# On Predicting and Understanding Music Skipping using Deep Reinforcement Learning

## 4.1 Introduction

MRSs aspire to tackle the problem of providing the users the support they need to access the large available collections of music items and find songs that match their interests and needs. Recent research has seen a significant effort around black-box recommender approaches such as DRL [67, 216]. This is motivated by the possible radical changes of behaviour from one song to another, or even within the same song, but at different points in time. Users' behaviour is influenced by external (trends) and internal (individual changes of personal interests) factors. The users' shifting interests and behaviour make it hard to learn a generalisable model to tailor the user's specific needs at any given time; it is a case where DRL is required due to continuous learning and adaption [29–31]. These advances have led, together with the increasing concerns around users' data collection and privacy, to a strong interest in building responsible RSs [217]. Constraints should be put in place when considering what data is collected

and then presented to a model to measure user behaviours. This is due to the potential hazard of introducing errors and biases. Therefore, minimising and selecting high-quality data features is of important consideration.

With explicit rating data relatively scarce and rare in today's systems, modelling implicit feedback is becoming of acquired importance. For example, in a *lean-back* formulation, the case of automatic playlists or radio streaming, the user interaction is minimised. Users are presented with a single song at a time. The MRS needs to rely almost entirely on implicit feedback signals such as the skipping or scrubbing (i.e., seeking forward and backward by moving the cursor [93]) to predict satisfaction and engagement [29, 30]. By first understanding these interactions, insights can be drawn to enable the construction of more transparent and responsible systems. The skipping is a signal that can measure users' satisfaction, dissatisfaction or lack of interest, and engagement with the platform [14]. For example, in a *lean-back* formulation, the MRSs are often designed to be more conservative, prioritising *exploitation* over *exploration* to minimise negative feedback (in this context, skips) [84]. Thus, one of their goals may be determined as recommending songs that yield the highest listening activity (i.e. no skip). However, understanding the users' skipping behaviour is still an under-explored domain [6, 13, 14]. It is a challenging problem due to its noisy nature: a skip may suggest a negative interaction, but a user may skip a song that they like because they recently heard it elsewhere.

Existing prior research in analysing the skipping behaviour revealed an universal behaviour in skipping across songs, with geography, audio fluctuations or musical events [15–17], and contextual listening information affecting how people skip music (Chapter 3). Recently, the effectiveness of deep learning models has also been explored for the task of predicting the users' sequential skipping behaviour in song listening sessions [18–24]. While they made a significant contribution towards this direction, their process is usually seen as an independent and static procedure. They may not account for the dynamic nature of the users' behaviour, and do not intuitively optimise for the long-term potential of user satisfaction and engagement [25–30]. Overall, this motivates the investigation of the DRL's applicability in predicting music skips and a comprehensive

investigation on the relation of the skipping signal with users' behaviour, listening context, and content.

In this chapter, I aim to understand how a DRL-based model predicts music skips by comprehensively analysing the utility of users' historical data. In particular, I analyse the impact and effect of the users' behaviour (e.g., the user action that leads to the current playback to start), listening content (i.e., the listened song), and contextual (e.g., the hour of the day) features in the classification task of predicting the users' music skipping behaviour. I propose a novel approach that leverages and adapts DRL for this classification task. This is to most closely reflect how a DRL-based MRS could learn to detect music skips.

### 4.1.1 Research Questions

This chapter aims to investigate the following two important RQs, which are investigated through an extensive study on the real-world MSSD:

- **RQ-4.1:** Can DRL be applied to the users' music skipping behaviour prediction task, and if so, would it be more effective in the music skip prediction task than deep learning state-of-the-art models?
- **RQ-4.2:** What historical information is considered discriminative and serves as a high-quality indicator for the model in predicting music skipping behaviour?

### 4.1.2 Contributions

The main contributions of this chapter are:

- I demonstrate the applicability and effectiveness of DRL in predicting users' skipping behaviour from listening sessions. A framework is devised to extend the DRL's applicability to perform this classification and offline learning. This is the first time that DRL has been explored in this task. The effectiveness of my approach is empirically shown on a real-world music streaming dataset (MSSD). My proposed approach outperforms state-of-the-art models in terms of Mean Average and First Prediction Accuracy metrics.

- I perform a comprehensive post-hoc (SHAP) and ablation analysis of my approach to study the utility of users' historical data in detecting music skips. I reveal a temporal data leakage problem in the historical data. Further, my results indicate that overall users' behaviour features are the most prominent and discriminative in how the proposed DRL model predicts music skips. The listening content and context features are reported to have a lesser effect.

This chapter is organised as follows. Section 4.2 introduces RL, the Deep Q-Network (DQN) algorithm, its advances, and the differences between online and offline learning. In Section 4.3 my approach is outlined. Sections 4.4 and 4.5 present, respectively, the experimental settings and results. Finally, Section 4.6 concludes the chapter.

## 4.2 Preliminaries

### 4.2.1 Reinforcement Learning (RL)

RL [216] is a computational goal-oriented framework where an agent learns how to perform sequential decision tasks via interactions with the system environment. It is usually formulated as a Markov Decision Process (MDP) to define the interaction process. An MDP, a model for sequential stochastic decision problems, is defined by a tuple $\langle S, A, p, R, \gamma \rangle$, where S is the set of states, $A$ the set of actions, $p$ the transition probability function $p : S \times A \times S \to [0, 1]$, $R$ the reward function $R : S \times A \times S \to R$, and $\gamma \in [0, 1]$ is a discount factor which is used to balance the importance of the immediate and future rewards. The agent, at each discrete time step $t \in \{0, 1, 2, ...\}$ and at a particular state $s_t \in S$, moves to the next state $s_{t+1}$ by performing action $a_t$ via a transition function $p(s_t, a_t)$. The agent receives a numerical reward $r_t$ from the environment as internally defined by the reward function $R(s_t, a_t, s_{t+1})$. The action that yields the highest reward is selected by the policy $\pi(s)$, a mapping from states to probabilities of selecting each possible action. The ultimate goal is to find the optimal policy $\pi_*$ that maximises the expected discounted cumulative return $G_t$, defined as:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{4.1}$$

The value function serves as a measure for an agent to assess the goodness of a state. This concept is central to almost all algorithms, especially value-function-based methods such as Q-learning [218]. The *state-value function* for a policy $\pi$ of a state $s$ under policy $\pi$, denoted $V_\pi(s)$, is the expected return when starting in state $s$ and following $\pi$:

$$V_\pi(s) = \mathbb{E}_\pi[G_t|s_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}\middle|s_t = s\right] \tag{4.2}$$

The *action-value function* for policy $\pi$ of taking action $a$ in state $s$ under a policy $\pi$, denoted $Q_\pi(s,a)$, is the expected return starting from $s$, taking the action $a$, and following policy $\pi$:

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t|s_t = s, a_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}\middle|s_t = s, a_t = a\right] \tag{4.3}$$

The goal is to find an optimal policy $\pi_*$ that achieves the maximum expected return from all states for $V_*$ and $Q_*$:

$$V_*(s) = \max_\pi V_\pi(s) \tag{4.4}$$

$$Q_*(s,a) = \max_\pi Q_\pi(s,a) = \mathbb{E}\left[R_{t'} + \gamma \max_{a_{t'}} Q_*(s_{t'}, a_{t'})\middle|s_t = s, a_t = a\right] \tag{4.5}$$

Despite tabular RL methods being successful in a multitude of domains, they suffer from the curse of dimensionality. Estimating the action-value function $Q_*(s,a)$ for each state-action pair with large state and action spaces, as is the case of real-world systems, is infeasible due to memory and computational constraints. Neural networks (NNs) have been proposed as a nonlinear method to generalise across states and/or actions to estimate the action-value function, i.e. $Q_*(s,a) \approx Q(s,a;\theta)$, with $\theta$ representing the neural network parameters. The combination of RL and NNs defines DRL.

### Deep Q-Network (DQN)

DQN [119] is an off-policy DRL algorithm. In DQN, the neural network's weights approximate the Q-values, making it applicable to large real-world problems. The $Q(s, a)$ function is thus modelled as a deep neural network that predicts the value of all possible actions given the input state. Moreover, to increase sample efficiency, transitions are stored in a replay memory, namely experience replay. The loss function minimises the Mean Squared Error between the Q-network and its target network at each iteration $i$:

$$L_i(\theta_i) = \mathbb{E}_{s,a,r,s' \sim D_i}[(r + \gamma \max_{a'} Q(s', a'; \theta_i^-) - Q(s, a; \theta_i))^2] \tag{4.6}$$

where $D$ is the replay memory, and parameters from the previous iteration $\theta_i^-$ are fixed when optimising the loss function $L_i(\theta_i)$. That is to mitigate episodes of catastrophic forgetting. Finally, the loss function is optimised and updated by stochastic gradient descent.

### Advances to Vanilla DQN

In the past years, numerous architectural advances have been proposed to the original DQN algorithm. They have been demonstrated to improve training, stability, convergence, and achieve state-of-the-art performance on various benchmarks (although not in a classification context). Double DQN (DDQN) [219], Dueling DQN [220], and n-step DQN [221] are some examples. Vanilla DQN suffers from an overestimation and thus maximisation of the bias of the Q-values, which deteriorates the performance and may lead to a sub-optimal policy. DDQN proposes a means to alleviate this problem via the use of a target network with periodical step-wise updates. On top of that, in Dueling DQN, the Q-value function is explicitly split in the network's architecture into the value of the state and the advantage of performing an action at that state. This can lead to better performance with enhanced training stability and convergence. Lastly, n-step DQN improves the propagation speed of values, in turn improving convergence and therefore speeding up learning. In a typical DQN, the action-value function is learned

with one-step Temporal Difference (TD) learning. By varying the rollout length $n$, we can transition from pure TD to Monte Carlo (MC) updates, with the return estimated by bootstrapping after $n$ steps. TD methods have a lower variance of the target value but higher bias in comparison with MC. In practice, lower values are usually preferable. This is because the off-policyness assumption breaks for larger values, making it turn towards on-policy learning.

Finally, partial observability in DRL, and in particular for DQN, is a challenging problem. Observations (frames) stacking [119] is a well-established method where a history of observations is stacked into the observation of the agent; sequences of a fixed number of frames are stacked together and given as input to the agent at each step. Recurrent Neural Networks (RNNs) are a viable alternative that is increasingly adopted in the literature but at the cost of greater computational requirements. Although it holds great potential, especially in sequential learning, major questions still remain open in their adaptability to RL-based methods. For example, how to properly train an RNN from replay memory, on whole trajectories, and by adhering to DQN's random sampling policy is not obvious and creates various practical, computational, and algorithmic issues [222].

### 4.2.2 Online and Offline Learning

On-policy and off-policy learning are the two main approaches to train a DRL agent. It also serves as the main classification criteria for its algorithms. Despite name similarities, on-policy and online as well as off-policy and offline are not the same. On-policy methods are those that attempt to evaluate or improve the policy that is used to make decisions. The agent works with "fresh" data that is obtained and sampled from interactions with the environment following the current policy. On the other hand, off-policy methods evaluate or improve a policy different from the one used to generate the data. It enables learning from logged data since the agent's policy improves upon the behavioural policy that was used to generate the data. The difference between online and offline lie in how training is performed, i.e. from live interactions or from a static dataset. The DQN's standard training procedure is entirely online. It is also an off-

policy algorithm, with its "off-policyness" originating from the greedy max operation in the TD update.

### Online Learning

Online learning is the standard formulation when adopting RL. It is an iterative process where the agent collects new experiences by interacting with the environment (or simulator of user experiences), typically with its latest learned policy. That experience is then used to improve the agent's policy. It is able to explore the environment by performing any action $a$ from its current state $s$ and observing the goodness of the new state $s'$. The trade-off between exploration and exploitation is essential in order to avoid the agent becoming stuck in a local optimum by encouraging active exploration of unknown, potentially high-reward areas of the environment. However, in many settings, this system of live interactions is impractical because data collection is either expensive or dangerous. Furthermore, if the domain is complex and with a vast array of logged data, exploiting such previously collected information may be helpful and informative for the agent as a form of (pre)training.

### Offline Learning

Offline learning, also known as Batch RL [223], is a prominent research area where the agent's task is instead defined as learning from a static dataset. In other words, policies are learnt from logged data. No interactions with the underlying environment are required, effectively removing the ability to explore. It has great potential in pretraining a DRL agent to some level of optimality before being deployed online, given the increased stability and greater learning speed compared to standard RL. Despite this ambition, challenges remain, most notably, the distribution mismatch between the current policy and the offline data policy [224]. Since learning is based on fixed transitions, an action selected by the agent that is different from the one reported in the dataset may lead to misestimated trajectories in the agent's internal learning mechanisms. It is fundamentally impossible to explore the effect that those actions would have in the real setting since we do not know the reward that this state-action

pair should be provided with. The untruthful state trajectory being generated is a wrongly estimated sequence of experiences that may hamper performance by leading to instability and/or ineffective learning (i.e., extrapolation error) [225, 226].

## 4.3 Approach

In this section, I present my framework to facilitate the application of DRL to the problem of sequentially predicting users' skipping behaviour from listening sessions. To do so, I model this problem as a MDP and a mechanism is introduced in the RL problem formulation to correctly exploit logged interactions and thus perform offline learning. The details of this framework are as follows:

- **State**: it is the record-level representation of a listening session at a discrete time step (i.e., position in the session). The state, i.e. a record in a listening session, includes various user's contextual information about the stream, their interaction history with the platform, and information about the track that the user listened to. An episode is the entire listening session, with sessions containing from 10 up to at most 20 records.
- **Actions**: it is a discrete action space which is a binary indicator of whether the current track is going to be skipped or not by the corresponding user. Effectively, the problem formulation can also be thought of as a binary classification problem $A = \{0, 1\}$, where 0 represents a no skip operation and 1 represents a skip.
- **Reward**: a positive reward of 1 is given for a correctly predicted skip classification, 0 reward (i.e., no penalty) otherwise.

Motivated by the discrete action space and off-policy requirements of the music skip prediction task, I leverage DQN. These requirements preclude the use of algorithms such as Deep Deterministic Policy Gradient [227] (continuous action space) and Proximal Policy Optimization [228] (on-policy learning). Whilst the problem is formulated as an MDP, it is partially observable (POMDP) by definition. This is because only partial information about the listening context and of the user is available [229]. Hence, in my

problem formulation, I consider MDP and POMDP to be equivalent. This means that I do not perform any further processing of the state representation (e.g., masking of some features).

This classification formulation can be seen as a guessing game, where a positive reward is given for a correct guess, and no penalty is given for an incorrect one. Long-term optimisation via discount factor $\gamma$ can be thought of as a way to correctly guess as many records in an episode as possible. Since there is a sequential correlation among records within an episode (i.e., a music listening session), a high $\gamma$ value should be used. This corresponds to optimisation on the total number of correct guesses in an episode (long-term) rather than optimisation on the immediate ones (short-term). By taking into account previous points in time and the past interactions with the environment, the DRL agent makes fully informed decisions.

### 4.3.1 Offline Mechanism

The DQN's standard training procedure is entirely online (see Section 4.2.2). Online learning is an iterative process where the agent collects new experiences by interacting with the environment, typically with its latest learned policy. That experience is then used to improve the agent's policy. However, exploiting logged data may be helpful and informative for the agent as a form of (pre)training. In offline learning, the agent's task is instead defined as learning from a static dataset. Policies are learnt from logged data, and no interactions with the underlying environment are required. Whilst my prior formulation would work in an online learning setting, it presents a major problem when performing offline learning. A misclassification would cause a transition to a new state, which is, however, not part of the original trajectory and thus not represented in the dataset as well. The agent will generate and associate a (discounted) cumulative reward to a wrongly generated trajectory that is substantially different from the original. Thus, a pure offline algorithm has to exclusively rely on the transitions that are stored in the dataset provided in the input. From my initial formulation, I need to account for those out-of-distribution actions.

Within the definition of the reward function itself, the out-of-distribution, untruth-

ful action is marked as invalid and, if sampled by the agent throughout learning, it causes the current episode to be terminated. In other words, an incorrect guess (0 reward) leads to a terminal state. This simple constraint forces a minimisation of estimation errors and, therefore, it avoids the creation of potential estimation mismatches. As such, the untruthful action that causes the current episode to terminate avoids the future propagation of incorrect bootstrapped return estimations in the TD target. This is to minimise the distributional shift issues due to differences between the agent's policy and the behaviour policy. More specifically, it explicitly ensures that regardless of the next sampled action, the current policy $\pi(a'|s')$ is as close as possible to the behaviour distribution $\pi_\beta(a'|s')$. The Q-function is queried as little as possible on out-of-distribution and unseen actions, since this will eventually increase errors in the estimations.

This error, i.e. "extrapolation error" [225], is introduced when an unrealistic and erroneous estimation is given to state-action pairs. This is caused when action $a'$ from estimate $Q(s, a)$ is selected, and the consequent state-action pair $(s', a')$ is inconsistent with the dataset due to the pair being unavailable. It provides a source of noise that can induce a persistent overestimation bias and that cannot be corrected in an off-policy setting, due to the inability to collect new data [225, 226]. Directly utilising DQN in an offline setting may result in poorer performance and a resemblance to overfitting [224]. My proposed mechanism minimises these errors. It is important to note that the "correct" action is not forcefully fed to the agent, as in Behaviour Cloning based approaches. I let the agent deterministically decide as if it were a live interaction with the environment, thus keeping the general workflow of the original algorithm intact. This provides a single interface to easily transition from offline to online learning and vice versa.

Finally, it is important to note that the aim of this work is to enhance our understanding of how a DRL-based model predicts music skips and to identify the high-quality features for its detection. To this end, I analyse the applicability of DRL in predicting this behaviour. I leave further tailoring of the approach to the music skip prediction task and an evaluation with recently proposed offline model-free al-

gorithms [230–233] for future work. Nevertheless, my proposed approach requires no architectural or algorithmic modifications. It offers the potential for a swift transitioning from online to offline learning and vice versa. It can be also be considered as a swift pre-training of an agent that can later be deployed online for continual learning.

## 4.4 Experimental Settings

### 4.4.1 Dataset

I conduct my experiments on the real-world MSSD provided by Spotify [6]. The publicly available training set consists of approximately 150 million logged streaming sessions, collected over 66 days from July 15th and September 18th 2018. Each day comprises ten logs, where each log includes streaming listening sessions uniformly sampled at random throughout the entire day. Sessions contain from 10 up to at most 20 records and are defined as sequences of songs/tracks that a user has listened to (one record per song). Each record includes various user's contextual information about the stream (e.g., the playlist type) and interaction history with the platform (e.g., scrubbing, which is the number of seek forward/back within the track). Although the track titles are not available, descriptive audio features and metadata are provided for them (e.g., acousticness, valence, and year of release). It is important to note that there is no user identification, nor access to demographic or geographical information. Hence, by not knowing whether two sessions have been played by the same user or by two different users, this study revolves around the modelling and understanding of the users' skipping behaviour.

#### Temporal Correlation

There is no temporal correlation among listening sessions, i.e. the sessions are not presented in historical order, which is reflected in the chance of consecutive sessions having a considerably different hour of the day (e.g., morning and evening). Also, there is no order to the ten logs within a given day (i.e., the 1st log of the first day does not necessarily occur before the 2nd of the same day). This does not preclude the potential

Table 4.1: Summary of datasets used for experiments after pre-processing. log(s) # indicate which log(s) are selected out of the available ten. skip (%) refers to the ratio between True and False values.

| Dataset | Date | log(s) # | # of records | # of sessions | skip (%) |
|---|---|---|---|---|---|
| Training Set | 15/07/2018 | [0, 3] | 11,927,861 | 711,838 | 51.20% |
| Test Set (T1) | 15/07/2018 | 4 | 2,991,438 | 178,419 | 51.21% |
| Test Set (T2) | 19/07/2018 | 8 | 3,395,883 | 204,145 | 50.53% |
| Test Set (T3) | 27/07/2018 | 0 | 3,447,209 | 207,060 | 50.76% |
| Test Set (T4) | 10/08/2018 | 6 | 3,407,685 | 205,267 | 50.42% |
| Test Set (T5) | 09/09/2018 | 1 | 2,588,711 | 155,617 | 51.48% |

applicability of DRL for the skip prediction task since the hour of the day in which a song was played is provided. Thus, it allows for the modelling of skipping behaviour dependent on the hour of the day.

### Creation of Training and Test Sets

In this work, I only leverage the training set since, in the test set, most of the metadata and the skipping attributes used as ground truth in my evaluation are not provided. By selecting logs from the original training set, statistics for my training and test datasets are presented in Table 4.1. As it can be seen from the statistics, the ratio of skip values for all sets is balanced between True and False values. This balanced distribution is an intrinsic property of the dataset and of any of the available logs. Due to the large amount of data, and therefore computational and execution time requirements, the first four logs of the first available day are used for training. Testing is performed on various logs in order to test the models' generalisability for different days. Except for T1, which is the 5th and next immediate consecutive log after the training set collection, all the other logs are of a random index, day and/or month. This random selection approach is justified by the fact that there is no temporal correlation among logs of the same day. This is to show the generalisation capabilities of my proposed approach and to allow for the comprehensive analysis of the importance of the users' historical data.

### Data Preprocessing

All available features, with a full description available in [6], are included in the state representation, except for the skip features, session and song identifiers. Categorical features, such as the playlist type and the user's actions that lead to the current track being played or ended, are one-hot encoded. All the audio features are standardised to have a distribution with a mean value of 0 and a standard deviation of 1. Overall, this results in a state representation consisting of 70 features. For ease of discussion, they are grouped as follows:

- **User Behaviour (UB)**
  - **Reason End (RE)** is the cause of the current playback to end. This is a one-hot encoded feature that thus groups various encoded features such as *Trackdone*, *Backbtn*, *Fwdbtn*, and *Endplay*.
  - **Reason Start (RS)**. Similar to *Reason End*, it is the type of actions that cause the current playback to start.
  - **Pauses (PA)** is the length of the pause in between playbacks. It consists of *No*, *Short*, and *Long Pause*.
  - **Scrubbing (SC)** is the number of seeking forward or backward during playback. They correspond respectively to *Num Seekfwd* and *Num Seekback*.
  - **Playlist Switch (PS)** indicates whether the user changed playlist for the current playback.
- **Context (CX)**
  - **Session Length (SL)** is the length of the listening session.
  - **Session Position (SP)** is the position of the track within the session.
  - **Hour of Day (HD)** is the hour of the day in which the playback occurred ([0..23]).
  - **Playlist Type (PT)** is the type of the playlist that the playback occurred within. Examples are *User Collection*, *Personalized Playlist*, and *Radio*.

  - **Premium (PR)** indicates whether the user was on premium or not.
  - **Shuffle (SH)** indicates whether the track was played with shuffle mode activated.

- **Content (CN)**. This third and final category groups all the **Track (TR)** metadata and features, as they constitute the only content-based information in the MSSD. It includes 28 features such as *Beat Strength*, *Key*, *Duration*, and the eight *Acoustic Vectors* ([0..7]).

### 4.4.2 Evaluation Metrics

To perform an evaluation of my proposed approach, I adopt the evaluation metrics from the *Spotify Sequential Skip Prediction Challenge*. This is also to provide a fair comparison with the selected baselines, since they were proposed on this challenge and for the following task: *given a listening session, predict whether the individual tracks encountered in the second half of the session will be skipped by a particular user.* Therefore, every second half of a session in the selected test set is used for prediction. If a session has an odd number of records, the mid-value is rounded up. This is motivated by the fact that an accurate representation of the user's immediately preceding interactions can inform future recommendations generated by the music streaming service. Hence, it is important to infer whether the current track is going to be skipped as well as subsequent tracks in the session. First Prediction Accuracy and Mean Average Accuracy are adopted as metrics.

- **First Prediction Accuracy (FPA)** is the accuracy at predicting the first interaction for the second half of each session.
- **Mean Average Accuracy (MAA)** is defined as:

$$MAA = \frac{\sum_{i=1}^{T} A(i)L(i)}{T} \tag{4.7}$$

where $T$ is the number of tracks to be predicted within the given session, $A(i)$ is the

accuracy up to position $i$ of the sequence, and $L(i)$ indicates whether the $i^{th}$ prediction is correct or not. Intuitively, in these evaluation metrics, higher importance is given to early predictions. In my setting, however, I do not exploit this specification in the problem formulation. Instead, the agent is instructed to optimise the total number of correct predictions in the session. This is to keep the system's specifications simple and easily adaptable to different metrics and/or tasks. In the dataset schema, prediction is based on the *skip_2* feature. It indicates a threshold on whether the user played the track only briefly (no precise threshold is provided) before skipping to the next song in their session.

### 4.4.3 Models

#### Baselines

To identify state-of-the-art baselines on the music skip prediction task, I performed an extensive search on prior works that utilise the MSSD dataset. I identified the following 4 of the top-5 ranked submissions to the *Spotify Sequential Skip Prediction Challenge* and presented at the WSDM Cup 2019 Workshop:

- **Multi-task RNN**: RNN-based approach that predicts multiple implicit feedbacks (multi-task) [18].
- **Multi-RNN**: Multi-RNN with two distinct stacked RNNs where the second makes the skip predictions based on the first, which acts as an encoder [19].
- **Temporal Meta-learning**: A sequence learning, meta-learning, approach consisting of dilated convolutional layers and highway-activations [20].
- **Weighted RNN**: RNN architecture with doubly stacked Long Short Term Memory (LSTM) trained with a weighted loss function [21].

They respectively reported the 1st, 2nd, 3rd, and 5th best overall performance on the Spotify Challenge, with Multi-task RNN being the strongest and Weighted RNN being the weakest baselines. The exclusion of the 4th overall best model on the challenge in my evaluation is because no manuscript and code repository were

found. For the selected baselines, I use the code accompanying the papers (GitHub links available in cited manuscripts). I then reproduced their results locally by running their provided public code locally, to the best of my abilities and with an optimised set of parameters. However, despite my best efforts, I reported consistently worse results than the ones in the Spotify Challenge public leaderboard and/or accompanying papers. These discrepancies in replicating the public results could be attributed due to differences in the parameter tuning, the selection of test sets, and the randomness introduced in the initialisation of weights in the neural networks. Despite utilising the provided parameters for local replication and performing optimisation, the model configurations used for the challenge could have differed. Additionally, the test set used in challenge is not fully released. No ground truth is available, thereby not allowing for a local evaluation. However, given my procedure for the creation of the train and test sets (Section 4.4.1), i.e. the training is performed on the first available day and the evaluation is for different days/months, I make the strong assumption that the overall data distribution of my selected test sets and the one used in the public challenge are similar. These factors, along with the inherent variability introduced by neural network initialisation, underscore the difficulties in achieving a direct comparison with the challenge results. For a fair comparison, I thus report the results from the public leaderboard as well as the ones from my local evaluation.

#### DQN Architecture

For this work, I explored nine state-of-the-art DQN architectures. By adhering to my proposed framework, they have been thoroughly investigated in the users' music skipping behaviour prediction task. They are the Vanilla [119], Double [219], Dueling [220], and their respective n-step learning variants [221]. Partially observable architectures have also been explored, with observations stacking [119] and Gated Recurrent Units (GRU) and LSTM based architectures [234].

A comparison among all these architectural variants is reported in Appendix A.1. I note that Vanilla DQN achieves the best performance. This is given its comparable performance and the advantage of a significantly simpler architecture with lower

complexity. The superiority of the Vanilla DQN architecture in this context can be attributed to the unique characteristics of the proposed training paradigm (see Section 4.3). The intuition behind this can be found in the nature of the task at hand. In the early training stages, an agent would normally explore low-reward regions of the environment. Actions have a strong effect on what the next states are, and divergence may happen if there are no stability guarantees. This is most especially true with the maximisation bias, where wrong high estimates are given to bad regions, leading the agent to a local optimum. Whilst being the case for live interactions with the environment, this is not an explicit property in a "data-driven", supervised-like training scenario where exploration is not present and only an already truthful sequence of state-action pairs has to be experienced. A classification that would deviate from this sequence is immediately truncated by the environment's specifications. This is reflected in lower improvements of the more advanced DQN architectural advances than would otherwise be expected. Therefore, the reported results are only for the Vanilla DQN architecture (hereafter referred to as "DQN").

### 4.4.4 Experimental Procedure

I trained my DQN using the following set of parameters: experience replay memory is 10000, batch size and frequency of updates are set as 256, the learning rate is 0.001, and the discount factor $\gamma$ is 0.9. The policy network consists of three fully-connected layers (of size 128) and a final action-value linear output layer of size 2. This final layer computes the Q-values for each action. Hyperparameters were selected by random and Tree-structured Parzen Estimator search, with the best set selected for evaluation on the test collections. The implementation of the DQN agent is provided by the Tensorforce [235] library. For complete reproducibility of my work, the code for this work is available at `https://github.com/NeuraSearch/Spotify-XRL-Skipping-Prediction`

To explore the potential instabilities and divergences during training, the proposed DQN approach is run five times per test set. The reported results represent the mean across all test sets. Lastly, during the training phase, learning is constrained with out-of-distribution actions, and therefore, some state-value pairs in the dataset are

not experienced by the agent due to early termination. During the testing phase, all episodic records are sequentially retrieved, and the agent acts deterministically on the complete episodes for its evaluation.

### Post-Hoc Analysis

In order to carry out an analysis on the importance and validity of the users' historical data in predicting music skipping behaviour, I first leverage the Shapley Additive Explanations framework (SHAP) [236]. It is a game-theory based approach that explains the predictions of machine learning models. In particular, I adopt the Kernel Explainer, which is a model agnostic method to estimate the SHAP values. This is because there exists no DRL specific explainers. However, since the Kernel Explainer makes no assumptions about the model to explain the predictions of, it is a highly expensive computational approach. This means that it is slower than the other model type specific algorithms. By considering these extensive computational requirements, for each test set, I estimate the feature importance values for the first 50 episodes (i.e., listening sessions) and with 200 perturbation samples per record. Given the high similarity across all test sets, I only report the results for T1.

### Ablation Analysis

To validate the SHAP results, I perform an ablation analysis on the input state representation. I study the effect that the category (e.g., *UB*) and type (e.g. *RS*) features have on the DQN's performance. To this end, I train and evaluate (following the same above-mentioned experimental procedure) the proposed approach on a state representation that does not include the selected features' type. This iterative approach, whereby only a single type is removed for each evaluation, is repeated until all types that comprise the input state representation are evaluated.

### Temporal Data Leakage

A closer investigation of the MSSD dataset, and validated by the post-hoc and ablation analysis, reveals a temporal data leakage of some features. These features have been

left unnoticed and they have inadvertently affected the Spotify Challenge and thus the baselines. These features correspond to the length of session ($SL$) and the user actions that lead the current playback to end ($RE$). This is because they provide to the model information from the future that should not be available in a live predictive system. Although I recognise and acknowledge this to be a problem, the reported results on the comparison with the selected baselines are without the removal of such features. This is to provide a fair comparison with the selected baselines, since they include these features in their input representation. These features are removed from the state when I investigate the reasons for people skipping music from the model's perspective, and it is referred to as the "corrected" state.

## 4.5 Experimental Results

First, the validity of my approach to predict users' music skipping behaviour is demonstrated against the state-of-the-art deep learning based models. My analysis of the music skipping prediction task and of the MSSD dataset reveals a temporal data leakage problem (Section 4.4.4). With a "correction" of the state representation by removal of such features, I report the comprehensive investigation on how the skipping behaviour can be detected by analysing the importance of UB, CX, and CN.

### 4.5.1 Applicability of DRL to Music Skip Prediction

On my local evaluation, Multi-RNN and Temporal Meta-learning, despite outperforming Weighted RNN in the challenge submissions, perform consistently worse on my selected test sets. Multi-Task RNN, the best performing baseline on the public challenge, achieves slightly inferior performance compared to Weighted RNN. In Table 4.2, it can be observed how the ranking of the reproduced baselines differs from that of the public leaderboard. Notably, Weighted RNN, which was ranked the lowest on the public leaderboard, appears to be the best performing baseline on my local evaluation. Overall, I note that all the baselines perform consistently worse on my local evaluation

[1]Leaderboard results available on cited manuscripts and/or at https://www.aicrowd.com/challenges/spotify-sequential-skip-prediction-challenge

Table 4.2: MAA and FPA results for my proposed DQN approach and baselines. The reported results are the averages across all test sets for DQN (with 95% Confidence Interval (CI)). For the baselines, I report the publicly available results from the Spotify Challenge[1]and of their local evaluation. For the former, this is to provide a fair comparison, since they are better than those obtained from my local evaluation. No CIs are reported for the baselines' public results due to their unavailability. The best performing model is highlighted in bold.

| | | *MAA* | | *FPA* | |
|---|---|---|---|---|---|
| | | Mean | 95% CI | Mean | 95% CI |
| | DQN | **0.820** | [0.818 - 0.822] | **0.881** | [0.880 - 0.882] |
| Public Leaderboard | Multi-task RNN | 0.651 | — | 0.812 | — |
| | Multi-RNN | 0.641 | — | 0.807 | — |
| | Temporal Meta-learning | 0.637 | — | 0.804 | — |
| | Weighted RNN | 0.613 | — | 0.794 | — |
| Local Evaluation | Multi-task RNN | 0.602 | [0.596 - 0.608] | 0.788 | [0.782 - 0.794] |
| | Multi-RNN | 0.479 | [0.475 - 0.483] | 0.725 | [0.721 - 0.729] |
| | Temporal Meta-learning | 0.588 | [0.585 - 0.591] | 0.782 | [0.779 - 0.785] |
| | Weighted RNN | 0.605 | [0.599 - 0.611] | 0.790 | [0.784 - 0.796] |

than in the public challenge. I observe decreases in performance of 4.9, 16.2, 4.9, 0.8 (%) and 2.4, 8.2, 2.2, 0.4 (%) in MAA and FPA and for Multi-task RNN, Multi-RNN, Temporal Meta-learning, and Weighted RNN respectively. Therefore, in Table 4.2, I report results in terms of MAA and FPA metrics for my proposed DQN approach with the baselines' public results from the Spotify Challenge. This is because they are better than those that I obtained from my local evaluation and to provide an as fair as possible comparison. My proposed approach exhibits improvements over all baselines on both MAA and FPA metrics. My proposed DQN registers an increase of performance for both MAA and FPA of 17% and 7% respectively with regards to Multi-task RNN, the best performing baseline from the public challenge.

Overall, my results demonstrate the validity and applicability of DRL to predict users' music skipping behaviour (addressing **RQ-4.1**). A Vanilla DQN architecture can outperform the more complex deep learning based state-of-the-art models. Furthermore, the results and a thorough analysis (see Appendix A.2) also indicate that convergence is achieved using a considerably lower number of episodes, at around $2 \times 10^5$

($\sim$ 1/4 of the episodes in the training set). This suggests sample efficiency and swift convergence of my proposed approach. Thus, it also addresses the well-known problem of DRL, which is its computationally intensive and slow learning. My approach converges swiftly and, in contrast to the selected baselines, it does not require Graphics Processing Unit (GPU) access. The low variability in performance across multiple runs and during the learning process also indicates stable and effective learning.

### 4.5.2 Identification of Temporal Data Leakage

In the previous section, I compared my proposed DQN against the selected baselines in order to demonstrate the validity of my proposed DQN. By performing an as fair as possible comparison, empirical results indicate the superiority of my approach. However, this benchmarking introduced errors into the model. This is because, as described in Section 4.4.4, I recognise that there are data leaking features in MSSD. The *SL* informs the model of how many songs a given user will listen to. This should not be made available because it is impossible to know how many songs a user will listen to in their current listening session. Further, the *RE* features provide information about how the current stream ends. This information should also not be exposed to the model. However, to provide a fair comparison with the baselines, since they are included in their input representation, these features were not removed despite my acknowledgement.

The temporal data leakage problem is validated by Figure 4.1, which reports the analysis of the average impact on model output (SHAP) of all features in the input state representation. It can be noted how the most discriminative feature to detect music skips is *RE Trackdone*, followed by *RS Trackdone*, *RS Fwdbtn*, and *Short PA*. *SL* is also found to have a relative impact (19th). It is clear that the proposed DQN considers these features to be of high quality and prominent importance for predicting the users' music skipping behaviour. However, they introduce a data leaking problem. By their removal from the input state representation, I observe a decrease in performance for my proposed DQN of 16% and 11% in MAA and FPA respectively. Further, I observe decreases in performance of 5.2, 26.2, 7.6, 0.6 (%) and 3.5, 28.4, 6.0, 1.3 (%) in MAA and FPA for Multi-task RNN, Multi-RNN, Temporal Meta-learning,

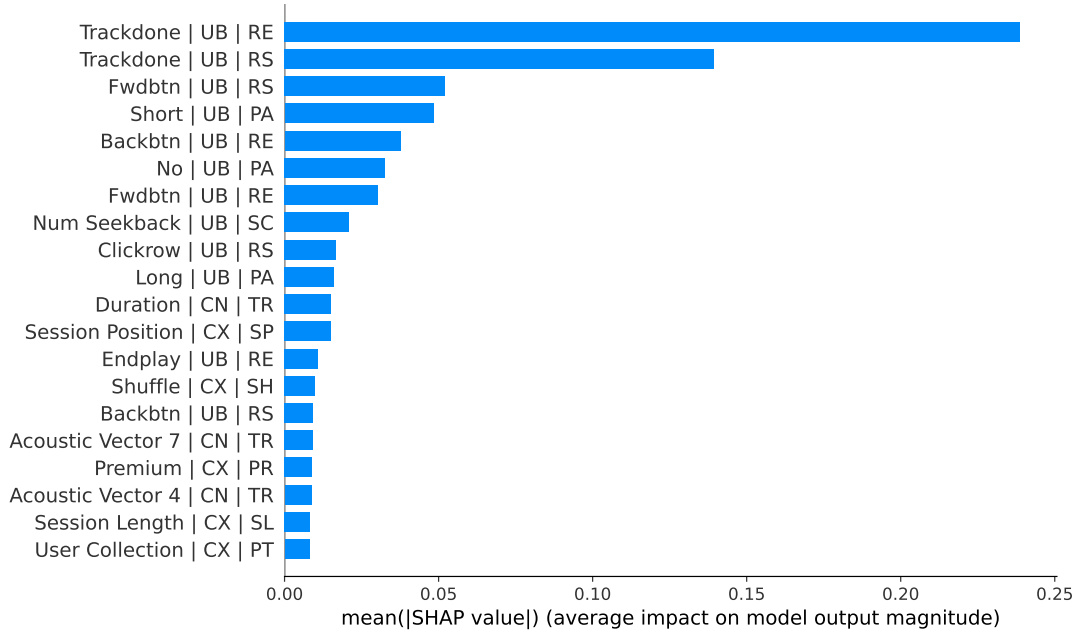



Figure 4.1: SHAP features importance analysis of the proposed DQN. The categorisation of the features and an explanation of the used acronyms are described in Section 4.4.1. Features are ranked in order of importance and they are reported as "[Name] — [Category] — [Type]".

and Weighted RNN respectively (differences calculated from the results obtained in my local evaluation after removal of the features with those reported in the public challenge). This higher decrease in performance of the proposed DQN compared to the baselines, upon correcting the temporal data leakage, can be attributed to the inherent reliance of DRL on user behaviour data to make predictions. Thus, the removal of such behavioural features (i.e., *RE*) had a more considerable impact on the proposed DQN than on the selected baselines. This could be attributed to the baselines not leveraging user behaviour to the same extent. This observation underscores the importance of DRL models in analysing nuanced user interactions. Overall, these results validate my initial intuition and demonstrate the data leakage problem. This finding provides a strong implication for a future outlook on creating attentive data collection procedures for transparent measurements of user behaviours. Offline benchmarks should be an as truthfully as possible reflection of real-world (online) tasks.

### 4.5.3 The Role of User Behaviour, Context, and Content in Detecting Music Skips

In this final section, I aim to address my main RQ: how does the proposed DRL model predicts music skips? (**RQ-4.2**). To this end, I acknowledge and thus remove the leaking features from the state representation to enable for a correct modelling of the users' music skipping behaviour.

#### User Behaviour (UB)

Figure 4.2 reports the SHAP features importance analysis of the proposed DQN on the "corrected" state representation. It can be observed that how the user interacted with the underlying platform to start the current playback (i.e., the *RS* type) is considered being the most discriminative feature to detect music skips. *Trackdone* and *Fwdbtn* are the highest negatively and positively correlated features in predicting a skip. They correspond to the user starting the current playback having listened in full or having pressed the forward button (i.e., skip) on the previous playback. These findings validate the observations of Chapter 3. By considering the listener and skipper user types, I hypothesise that the user behaviour that can inform the membership of a user to one of these two types is *RS Trackdone* or *Fwdbtn*. From my results, it is clear that how a person interacted with the previous song appears to greatly affect the DRL's ability to detect how they will interact next. Another UB that appears to have a prominent effect is the pause in between playbacks. A *Short PA* and a *No PA* are shown to highly and weakly suggest a music skip respectively. In the case of a *Long PA*, my results strongly indicate that the user will not skip their current song. This finding validates my initial hypothesis. It may correspond to a person searching the catalogue for a song they would like to listen, and hence a long pause. Therefore, it is intuitive that it may not be skipped. However, the effect of a short pause in detecting music skips is of surprising effect. This may be justified by a user's exploratory state, where they browse the catalogue and briefly listen to multiple songs until they find a match for their needs.

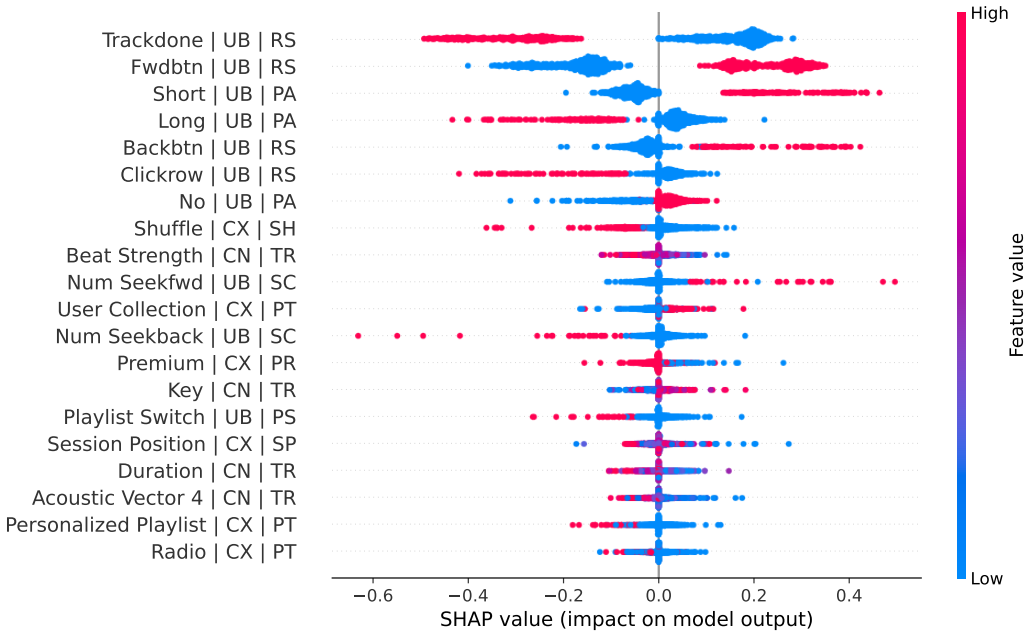



Figure 4.2: SHAP features importance analysis with positive (skip) and negative (no skip) impact values of the proposed DQN on a "corrected" state representation (i.e., after addressing temporal data leakage). The Feature Value axis refers to high or low observational values. For Boolean features (e.g., *RS Trackdone*), high/red is a True value, and low/blue is False. The categorisation of the features and an explanation of the used acronyms is described in Section 4.4.1. Features are ranked in order of importance and they are reported as "[Name] — [Category] — [Type]".

### Context (CX)

I observe that users that listening in *Shuffle* mode and/or with a *Premium* account are associated with less skipping activity. Listening with a *User Collection* PT is associated with a higher skipping rate. It is also shown that listening under a *Personalised Playlist* or *Radio* is subject to more listening and thus less skipping activity. This finding could suggest that they have a higher users' engagement. However, this is not possible to quantify, and further evaluation is required in order to understand this phenomenon. This could be explained by the noisy nature of the skipping activity and the possibility, as in the example of radio listening, of passive (background) consumption of the music. Although the *PT* findings appear to partially validate prior work [14], in my ablation analysis I see that their removal from the state representation registers no significant effect on the DRL's ability to predict music skips.

### Content (CN)

The only content-based features in the MSSD are related to the track being listened by the user (*TR*). The correlation between skipping activity and the *TR* features is less obvious since they appear to be less discriminative and promiment in detecting music skips. *Beat Strength* and *Key*, although mostly centred around a zero impact, suggest that a high beat strength is associated with more listening, and a high-pitched song (*Key*) with higher chances of skipping. Further, longer songs (*Duration*) are usually associated with higher listening activity, although they may also correspond to skips. However, in my ablation analysis, I observe the no effect on the DQN's performance by the removal of all *TR* features. I find this to be of surprising effect, since it appears to contradict prior research suggesting that audio characteristics influence how people skip music [16, 17].

### Ablation Analysis

In order to validate my findings and to demonstrate the impact, whether statistically significant or not, that these features have on the DQN's performance, in Table 4.3 I report the results for the ablation analysis. I performed paired t-tests on the prediction

Table 4.3: MAA and FPA results for my ablation analysis of the proposed DQN on the corrected state representation. The reported results are the average across all test sets and the 95% CIs. (*) and (**) indicate that the selected type of features had a statistically significant effect in performance in the proposed DQN (on a "corrected state") on MAA or FPA. This is based on confidence levels ($p < .05$) and ($p < .001$) respectively.

| | | *MAA* | | *FPA* | |
|---|---|---|---|---|---|
| | | Mean | 95% CI | Mean | 95% CI |
| | Corrected State | 0.664 | [0.662 - 0.666] | 0.773 | [0.772 - 0.774] |
| UB | Reason Start (RS) | 0.389 (**) | [0.378 - 0.400] | 0.479 (**) | [0.464 - 0.494] |
| UB | Pauses (PA) | 0.659 (*) | [0.657 - 0.661] | 0.769 (*) | [0.768 - 0.770] |
| UB | Scrubbing (SC) | 0.659 | [0.655 - 0.663] | 0.770 (*) | [0.768 - 0.772] |
| UB | Playlist Switch (PS) | 0.662 | [0.659 - 0.665] | 0.773 | [0.772 - 0.774] |
| CX | Hour of Day (HD) | 0.663 | [0.661 - 0.665] | 0.773 | [0.772 - 0.774] |
| CX | Playlist Type (PT) | 0.663 | [0.661 - 0.665] | 0.772 | [0.771 - 0.773] |
| CX | Premium (PR) | 0.664 | [0.662 - 0.666] | 0.773 | [0.772 - 0.774] |
| CX | Shuffle (SH) | 0.663 | [0.660 - 0.666] | 0.774 | [0.773 - 0.775] |
| CN | Track (TR) | 0.664 | [0.661 - 0.667] | 0.773 | [0.772 - 0.774] |

accuracy of the proposed DQN (on the "corrected" input state representation) with each of the selected type of features (e.g., *RS*). I use (*) and (**) to denote the fact that the removal of the selected type of features had a statistically significant effect in performance in the proposed DQN on MAA and FPA. This is based on confidence levels ($p < .05$) and ($p < .001$) respectively. I note how the *RS* features type, as previously shown in Figure 4.2, is the highest quality estimator to detect music skips. Its removal registers a decrease in performance of 28% and 29% in MAA and FPA respectively. The *PA*s also register a significant impact. All the remaining features, including the CX and CN categories, do not appear to show a statistically significant effect on the DQN's performance. These results, therefore, suggest that a limited amount of users' data can be indeed leveraged to predict the users' music skipping behaviour, with only the *RS* and *PA* user behaviours showing a statistically significant effect.

## 4.6 Chapter Summary

In Chapter 3, I demonstrated that different and distinct skipping types can be successfully identified and categorised. This comprehensive analysis was performed on the users' session-based skipping behaviour. The analysis was conducted by considering various listening contextual factors, including day type, time of the day, playlist type, account type, and shuffle mode. Overall, it improved our overall understanding of how people skip music and how listening contextual factors affect such behaviour. However, the main limitation of this work lies in its focus on a limited set of features and, most importantly, it does not delve into the underlying reasons behind why people skip music.

In this chapter, I aim to understand how people skip music from a DRL-based model perspective. To perform such an analysis, I first proposed to leverage DRL to the task of sequentially predicting users' skipping behaviour in song listening sessions. By first understanding how a DRL model learns individual user behaviours, we can then help the process of explaining recommendations of a DRL-based MRS. To this end, I extended the DRL's applicability to this classification task (see Section 4.3). Results on a real-world music streaming dataset (Spotify) indicate the validity of my approach by outperforming state-of-the-art deep learning based models in terms of MAA and FPA metrics (**RQ-4.1**; see Section 4.5.1 and Table 4.2). By empirically showing the effectiveness of my proposed approach, my main post-hoc (see Section 4.5.2 and 4.5.3) and ablation analysis (see Section 4.5.3) revolves around a comprehensive study of the utility and effect of users' historical data in how the proposed DRL detects music skips (addressing **RQ-4.2**).

In my analysis, I observed the following key findings:

- The most discriminative indicator for an accurate detection of skips is how users interact with the platform (i.e., *RS* and *PA*; see Figure 4.2).
- Surprisingly, the listening CX and CN features explored in this work do not appear to have an effect on the DRL model for the prediction of music skips (see Section 4.5.3 and Table 4.3).

- My analysis also reveals a temporal data leakage problem derived from some features in the dataset and used in the public challenge, since they provide information from the future that should not be made available to a live predictive system (see Section 4.5.2 and Figure 4.1).

Overall, this work shows that an accurate representation of the users' skipping behaviour can be achieved by leveraging a limited amount of user data. This offers strong implications for the design of novel user-centred MRSs with a minimisation and selection of high-quality data features to avoid introducing errors and biases. The results and a thorough analysis of my proposed approach indicate sample efficiency, swift convergence, and long-term stability of my proposed approach (see Appendix A). With convergence reached using a considerably lower number of episodes, training time can be greatly reduced by early termination. With no GPU access required (in contrast to the state-of-the-art deep learning based models), my approach also clearly addresses the well-known limitation of DRL being a computationally extensive approach. These findings and the consistent performance with no signs of instability make this work of great interest for future research.

Given the importance of modelling and understanding the users' skipping behaviour, I consider Chapter 3 and 4 to be important advancements in enhancing user modelling techniques. This is because they provide a more precise understanding of the patterns and motivations behind music skipping. This knowledge can be leveraged to enhance the user modelling techniques, thus leading to considerable improvements in the building of novel user-centred MRSs.

# Part III

# The Podcast

Recognising podcasts' growing importance as a significant medium for online information seeking, as outlined in Chapter 2, Section 2.5, this part of the thesis delves into the relatively under-explored domain of podcast research. The rise of podcasts on unified platforms alongside music content introduces new research challenges, particularly in understanding user behaviour, optimise podcast streaming platforms, and exploring the concept of relevance in the context of podcasts [3, 4].

Therefore, a key research direction in optimising podcast platforms is understanding user behaviour and exploring the concept of relevance within a podcast streaming experience. This necessity forms the basis of this user study, inspired by the challenges and datasets presented by the TREC Podcast Track and the Spotify Podcast Dataset [34] (see Chapter 2, Section 2.5.2).

Part III aims to investigate the following two important RQs, which relate to the impact of integrating textual components in the UI of podcast streaming platforms:

- **RQ-7.1**: Can captions and full-text transcripts enhance user experience and engagement on podcast platforms?
- **RQ-7.2**: How do these textual features influence users' ability to determine the relevance of podcast content?

These RQs are investigated through a user study on the *Podify* platform, detailed in the upcoming chapters. Chapter 5 introduces the platform, Chapter 6 outlines the study's methodology, and Chapter 7 presents the findings.

# Chapter 5

# Podify: A Podcast Streaming Platform with Automatic Logging of User Behaviour for Academic Research

## 5.1 Introduction

As previously discussed in Chapter 2, online audio streaming services have a long-lasting connection with the music industry, which has been their main pivotal point for decades. In recent years, more and more streaming services are now expanding their catalogues to support both music and podcasts on the same platform (e.g., Amazon Music and Spotify). This poses a UI challenge on what and how much information has to be presented to the user [4]. Podcasts are spoken documents (they can be represented by their transcripts of their spoken content) that have gained significant interest and widespread popularity in recent years. They have recently started a rapidly growing process and have swiftly become, although under-researched [3], an essential part of listening habits.

In 2020, Spotify identified the podcast as an important research domain and released the Spotify Podcast Dataset [35]. It is a large corpus of over 100,000 episodes, each

comprising an audio file, automatically transcribed text via Google's Cloud Speech-to-Text APIs, and associated metadata. Although the dataset's great applicability to various tasks in fields such as speech and audio processing, natural language processing, information retrieval, and computational linguistics, it is unsuitable for those where logged user behaviour is required [44]. This is the case of analyses of user information needs, their characteristics and behaviour, relevance, search, recommendation, and personalisation systems. Further, there are currently no platforms available to the academic community to conduct experiments and user studies, thereby significantly hindering the advances in the field. Despite the growing research interest in this domain, conducting user studies remains challenging due to the lack of datasets that include user behaviour. In particular, there is a need for a podcast streaming platform that reduces the overhead of conducting user studies.

To address these needs, and given the considerable growing importance of conducting research in podcasts, in this chapter, I propose and release a new web-based platform named *Podify*. It is the first podcast streaming service specifically designed for academic research. The platform highly resembles existing streaming systems to provide users with a high level of familiarity on both desktop and mobile. A catalogue of podcast episodes can be easily created via RSS feeds. The platform also offers Elasticsearch-based indexing and search that is highly customisable, allowing research and experimentation in podcast search. With a scalable design to accommodate large-scale user studies, it implements manual playlist creation and curation, podcast listening, and explicit (i.e., liking and disliking behaviour) and implicit feedback (i.e., user interactions) collection mechanisms. With all user interactions automatically logged by the platform and easily exportable in a readable format for subsequent analysis, *Podify* aims to reduce the overhead researchers face when conducting user studies. A demonstration of the platform is available at `https://youtu.be/k9Z5w_KKHr8`, with the code and documentation available at `https://github.com/NeuraSearch/Podify`.

### 5.1.1 Motivation

To foster research in the podcast domain, an online streaming service with a catalogue search for academic purposes is required. This is due to the rapidly increasing notoriety of podcasts and the ever-increasing worldwide consumption of this medium. However, the lack of such a platform poses a significant challenge, and it hinders progressive research. The *Podify* platform aims to address this problem. It is a podcast web-based streaming platform that resembles existing streaming services. The search is powered by Elasticsearch, and it can be easily adapted according to the researchers' needs to conduct research in this domain.

### 5.1.2 Contributions

This contribution of this chapter is the creation, development, and release of *Podify*. It is the first web-based podcast streaming platform explicitly designed for academic research. This platform allows researchers to conduct user studies in the podcast domain to alleviate the lack of user interactions in the Spotify Podcast Dataset [44]. This work can foster academic research toward podcast consumption patterns and listening activities and thus elicit research in novel personalisation techniques.

## 5.2 System Architecture

In this section I describe the *Podify*'s UI, search and catalogue creation functionalities, the user behaviour that is collected, and the implementation details.

### 5.2.1 User Interface (UI)

*Podify* is optimised for both desktop and mobile use. It features a user authentication process that requires users to sign up with their username (which could be Subject IDs or Worker IDs in the case of crowdsourcing studies on Amazon MTurk) for cross-referencing their behavioural data. Upon successful registration, users are redirected to the homepage (Figure 5.1), which includes a sidebar (A) with links to liked episodes, disliked episodes, and playlists (with an option to create, delete, or rename). An empty

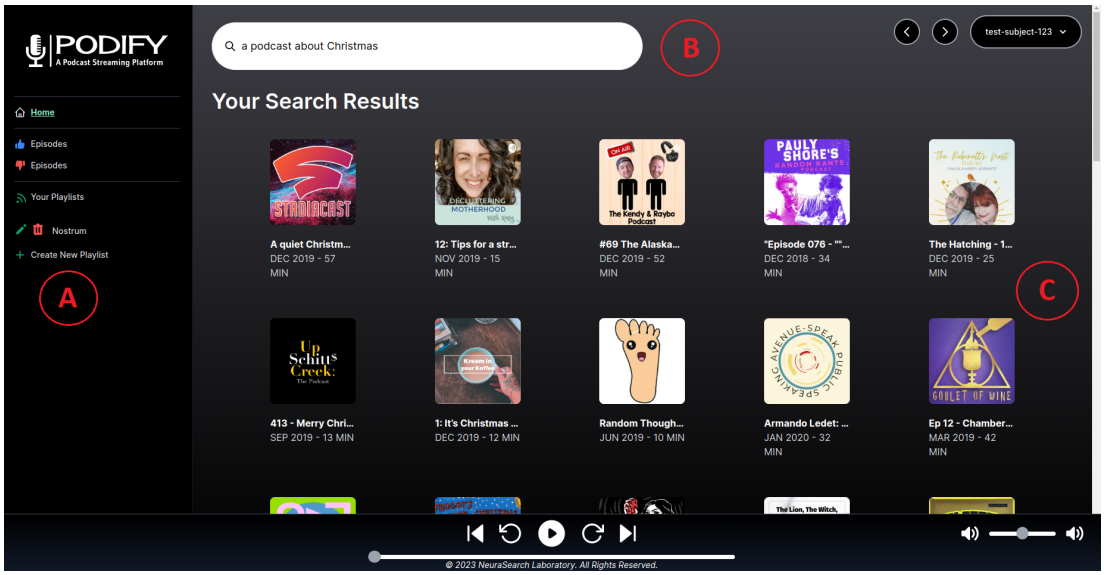



Figure 5.1: *Podify*'s UI with an example of catalogue search. Top@50 results for the query "a podcast about Christmas".

playlist for the user is generated by default. The homepage also features a search bar (B) for catalogue browsing. When a user performs a search, the results are displayed to them as shown in Figure 5.1 (C). They can be inspected by clicking on individual episodes to view their metadata (Figure 5.2). Users can also provide explicit feedback (like, dislike, or textual feedback with a 1-5 rating) and add them to their playlists.

To listen to podcast episodes, a user must first populate a playlist with episodes of their choice and then select it for consumption. Figure 5.3 shows an example of a playlist. The episodes are played sequentially, allowing the user to perform jump-ins, re-arrange the sequence order, provide explicit feedback, or remove episodes from the playlist. During playback, the user has access to various controls at the bottom of the screen (Figure 5.3 (D)). The controls include: skip to the previous episode, seek back (for 30 seconds), play/pause, seek forward (for 30 seconds), and skip to the next episode.

*Podify* also includes an admin dashboard[1], providing:

1. A visual representation of the underlying *Podify*'s database. The dashboard allows for the creation, editing, and deletion of database records (i.e., experiments,

[1]Demonstration available at https://youtu.be/k9Z5w_KKHr8

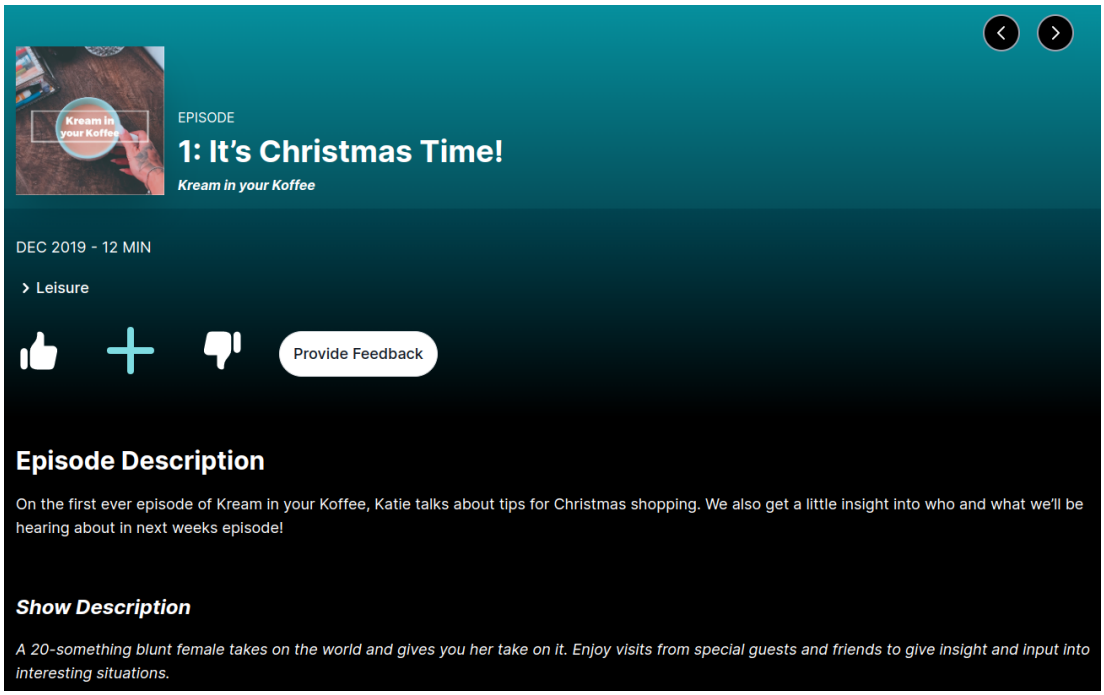



Figure 5.2: Episode's page with metadata (e.g., publication date), like, add to a playlist, dislike, and textual explicit feedback.

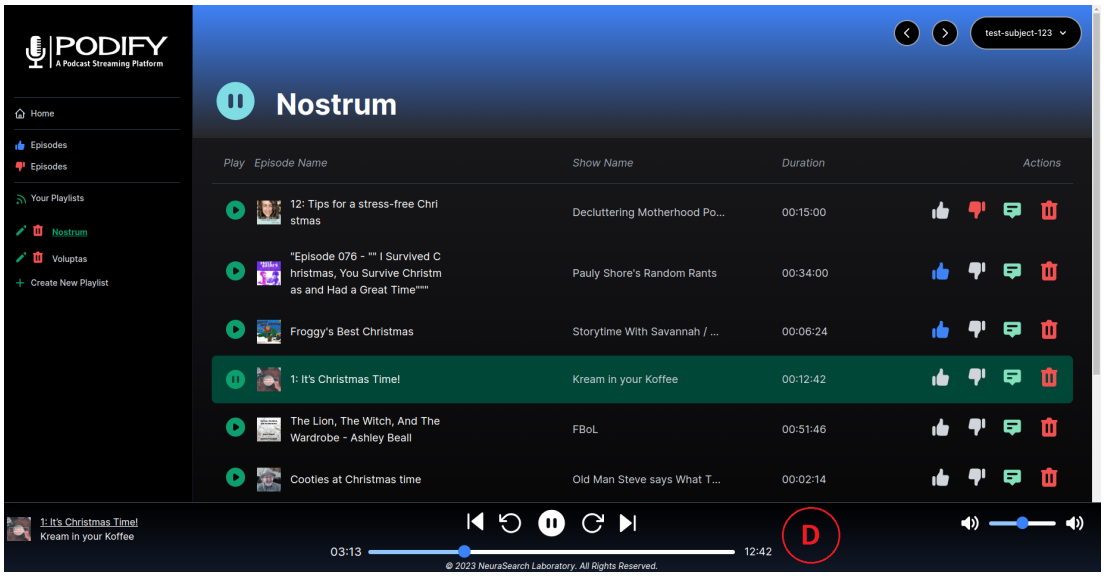



Figure 5.3: *Podify*'s UI with an example of a manually curated playlist (i.e., "Nostrum") and episode consumption.

systems, episodes, playlists, users and their logged interactions).

2. All the collected users' behavioural data can be inspected and easily exported in a readable Comma-Separated Values (CSV) format.

3. Creation of different variations (systems) of the platform for conducting user studies. At the moment, systems can be set to have different catalogues. The systems can be changed for a user through specific URLs provided during, for example, surveys. I expect future versions to extend this functionality to include more control variables (e.g., search intents or recommendation procedures).

### 5.2.2 Search Functionality

Searching for episodes in the *Podify* catalogue requires a running Elasticsearch instance. This can be achieved through (i) a local Docker container or (ii) by connecting to a Bonsai Elasticsearch cluster on, for example, Heroku.

With an a priori running Elasticsearch instance, the catalogue can be created. The instance indexes the catalogue and makes it searchable, as demonstrated in Figure 5.1 (B). The search functionality uses the Okapi BM25 with default values as the ranking model and a Snowball-generated stemmer for word stemming for the indexing. The indexed fields include a transcript (double weight), episode name, episode description, show name, and show description. The inclusion of the ASR transcript is motivated by prior works [36, 137] suggesting that its inclusion significantly increases the search quality compared to a metadata-only based search. This also motivates the need for a double weight. Finally, the top@50 results are presented. It is important to note that such search functionality (including, for example, the ranking model) is highly customisable.

Elasticsearch was selected as the backbone system for indexing and retrieval because of its scalability, robustness, and speed, which are essential for managing large datasets. However, despite its strengths, it may not fully meet all specific research needs because of its inherent constraints (e.g., ranking models). This is a limitation I aim to mitigate through ongoing customisation and optimisation. For example, by considering a "plug-

and-play" approach, it would be possible to integrate a variety of engines and algorithms into *Podify* to provide additional options for user study customisation. Given the flexibility of the platform, this adaptability ensures that *Podify* can evolve in response to the changing needs of academic research, enabling more tailored and effective studies in the podcast domain.

### 5.2.3 Catalogue Creation Procedure

The procedure to create the *Podify*'s catalogue of podcast episodes is straightforward and based on the RSS feeds. A JavaScript Object Notation (JSON) array containing a set of RSS feeds (one for each episode that is to be added to the catalogue) has to be created. It is then asynchronously processed by the backend, and the catalogue is thus created and automatically indexed as described in Section 5.2.2. It is important to note that although all the metadata is obtained from the RSS feeds, the audio and transcript files of each episode have to be made available to the creation procedure via Amazon Web Services (AWS) S3 buckets. The creation procedure leverages the unique episode filename field of each episode in the RSS feed to fetch the audio and transcript files remotely. *Podify* has a scalable architectural design, allowing for any catalogue size to be efficiently served by the backend.

Since *Podify* expects RSS feeds, it does not restrict its usage to only, for example, the Spotify Podcast Dataset [35]. However, the RSS feeds originating from the Spotify Podcast Dataset were used for this demonstration. In particular, I selected the first 1,000 episodes that have a valid RSS feed.

### 5.2.4 User Behaviour

All user interactions with *Podify* are automatically logged, and they are easily exportable in a readable format (CSV). Each interaction is also associated with a username, experiment, system, and timestamp. When there is a new visit to the platform, a visit record is also created. It contains information (General Data Protection Regulation (GDPR) compliant) about the traffic source (referrer, referring domain, landing page), country-level geocoding, and technology (browser, operating system,

device type). Internet Protocol (IP) addresses are masked, and cookies are switched to anonymity sets. The user interactions that *Podify* collects are:

- **Navigation**. page changes; catalogue search query; rank of the clicked item from the search results.
- **Episode**. explicit feedback (like, dislike, unvote, textual comment with rating); update selection (due to end of streaming or manual change); current time in playback (a new entry is logged every three seconds to monitor listening activity or when **listening** actions are performed).
- **Playlist**: create; delete; add or remove of episodes; update episodes' order; selection of playlist for playback.
- **System**: update of user's current system.
- **Listening**. play; pause; seek forward (30s or manual); seek back (30s or manual); volume (up, down, manual).

It is important to note that more interactions can be easily integrated into the logging system in future versions.

### 5.2.5 Implementation

*Podify* is built with Ruby on Rails (7.0.2), a modern and notorious server-side web application framework. With a PostgreSQL relational database management system, the front end is implemented with the Hotwire Stimulus JavaScript and the Tailwind CSS frameworks. I provide background asynchronous job processing to (i) accommodate the generation of podcast catalogues irrespective of length and (ii) for automatic backups (to an AWS S3 Bucket) of the collected user behaviour with a cron schedule of every 15 minutes. Although user behaviour can be manually downloaded via the admin dashboard (Section 5.2.1), this cron schedule is also implemented to avoid any potential data loss. Given *Podify*'s modern and state-of-the-art architectural implementation, I expect its integration into modern pipelines for scaling and swift content delivery to conduct large-scale user studies.

## 5.3 Chapter Summary

Podcasts, being spoken documents and a medium for online information seeking activities, have seen a significant surge in interest and popularity in recent years. Despite being an under-researched area [3], they have swiftly become an essential part of listening habits. However, the lack of datasets that include user behaviour has made conducting research studies in this domain a challenging task. This highlights the need for a dedicated podcast streaming platform for research purposes.

To address these challenges, and given the considerable growing importance of conducting research in this domain, in this chapter, I proposed *Podify*. It is a web-based platform for podcast streaming and consumption. With high resemblances to existing podcast streaming services, it is specifically designed to support academic research in this domain, especially in the under-explored areas of search and user behavioural analysis.

*Podify* is the first available research-purpose podcast streaming service that also features a highly customisable search functionality (Section 5.2.2) and an easy-to-create catalogue of podcast episodes (Section 5.2.3) that facilitates large-scale user studies (Section 5.2.5). Episodes can be organised in manually curated playlists and then played by users for consumption. All users' interactions with *Podify* are automatically logged, and they can be easily exported in a readable format for successive experimental analysis (Section 5.2.4). Mechanisms to collect explicit feedback (i.e., liking and disliking of an episode) are also provided.

As we progress to Chapter 6, I will leverage this platform and its functionalities to design a user study. The analysis and findings of such study will then inform the analyses presented in Chapter 7.

# Chapter 6

# Research Methodology

## 6.1 Introduction

In this chapter, I outline the methodology for the analyses conducted in Chapter 7. This study evaluates the impact of incorporating text-based components, such as captions and full-text transcripts, into the *Podify*'s UI, on the users' ability to accurately determining the relevance of podcast content in relation to their INs. This chapter is organised as follows. Section 6.2 presents the study's design, and Section 6.3 the apparatus employed. In Section 6.4, the chosen topics and corpus are outlined, with Section 6.5 discussing the questionnaires used in the study. Section 6.6 and Section 6.7 present, respectively, the qualitative and quantitative measures and the procedure. Finally, Section 6.8 concludes the chapter.

## 6.2 Experimental Design

To address the RQs presented in Chapter 7, Section 7.1.1, and to create a realistic podcast information access scenario, I conducted a within-subjects experiment with three factors: search intent, system, and task complexity. For the search intent, I used the *Topical* (*TO*) and *Known-Item* (*KI*) search intents introduced in the 2020 and 2021 TREC Podcast Track [34]. The *TO* search intent refers to the finding of general information about the topic. On the other hand, *KI* refers to finding something that is known to exist but under an unknown name [35,137]. For the system factor, my *baseline*

(*BA*) is based on the originally proposed UI of *Podify* [2], while the *enriched* (*EN*) system incorporated captions and full-text transcriptions. Finally, the complexity of the task was controlled and determined by the grading and relevance within the top@10 recommendations, resulting in *easy (Ea)* and *difficult* (*Di*) tasks. This design allowed us to simulate a realistic IN and recommendation scenario as closely as possible. This combination of factors results in a $2x2x2$ factorial design. Each participant completed eight unique tasks across the eight different conditions, as outlined below:

- (***TO.BA.Ea***) Topical IN, on the baseline system, and easy.
- (***TO.BA.Di***) Topical IN, on the baseline system, and difficult.
- (***TO.EN.Ea***) Topical IN, on the enriched system, and easy.
- (***TO.EN.Di***) Topical IN, on the enriched system, and difficult.
- (***KI.BA.Ea***) Known-Item IN, on the baseline system, and easy.
- (***KI.BA.Di***) Known-Item IN, on the baseline system, and difficult.
- (***KI.EN.Ea***) Known-Item IN, on the enriched system, and easy.
- (***KI.EN.Di***) Known-Item IN, on the enriched system, and difficult.

To mitigate any potential bias (e.g. the effect of task or fatigue), the tasks were evenly split (by search intent) between the morning and afternoon sessions. The allocation was randomised. This approach was taken to divide the first level of independent variables between the two sessions. Thus, this ensured that participants completed the four tasks related to one search intent in the morning, and the other four in the afternoon. The order sequence of the four tasks within a search intent (e.g., *TO.BA.Ea*, *TO.BA.Di*, *TO.EN.Ea*, and *TO.EN.Di*) was also randomised. Conditions were assigned using a Latin square rotation to mitigate potential ordering effects. The dependent variables comprised the qualitative (questionnaires) and quantitative (user interactions) data.

Last, it is important to highlight that participants were briefed on the search intent for each session. However, they were not provided with the information regarding the

system being used (although they could observe the differences in the UI when starting the task) or the complexity of the task.

## 6.3 Apparatus

For my experiment, I employed a custom web-based survey platform developed at the NeuraSearch Laboratory (Computer & Information Sciences Department (CIS), University of Strathclyde), along with a specially customised version of the *Podify* podcast streaming platform (Chapter 5). The entire experiment was conducted remotely, with participants using their personal computers to engage in the two sessions of the study. Upon recruitment, participants were redirected to the survey platform, where both sessions of my study occurred.

The interface of *Podify* was changed to include two key system variations, namely *BA* (based on the original *Podify* interface) and *EN* (enhanced to incorporate the textual components of captions and full-text transcripts). Figure 6.1 presents the captions (A) and full-text access (B) components. Figure 6.2 depicts how participants could access a complete segment's transcript. This also includes an exact word-match search feature (component (C)).

Thorough testing was conducted across various browsers, operating systems, and screen resolutions to ensure a consistent experience among participants. This approach ensured that the experiment could be performed without technical difficulties, irrespective of the diverse hardware and software environments that individual participants may have employed.

## 6.4 Topics & Corpus

My experiment utilised the train and test collections from the segment retrieval task of the 2020 TREC Podcasts Track. These collections included three search intents, namely *TO*, *KI*, and refinding (*RE*), with 41, 8, and 9 topics for each intent, respectively. Each topic was accompanied by a set of relevance assessments (*qrels*), specifying a segment's relevance score in relation to a two-minute of a podcast episode.

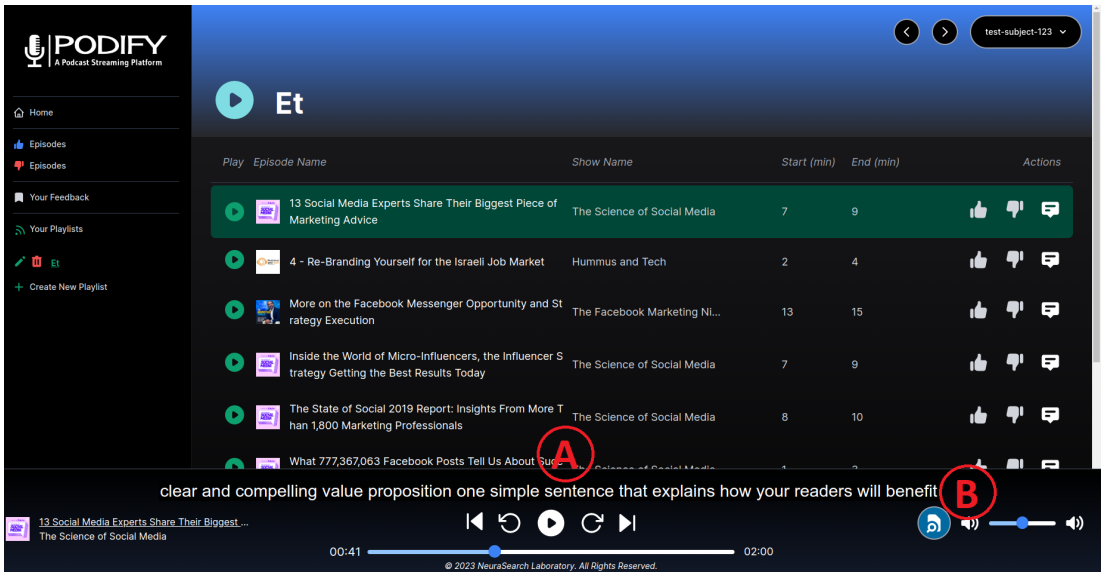



Figure 6.1: The *Podify*'s version used in this study (*EN* system). The *BA* system does not include the captions (A) and the access to the full-text transcript (B) textual components. The auto-generated playlist shows the segments for the topic "social media marketing".

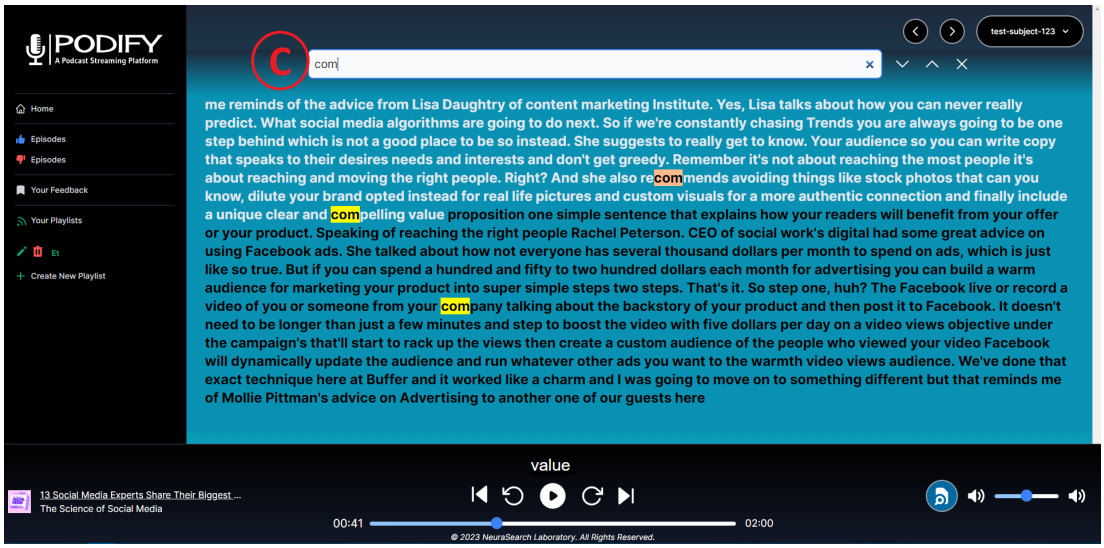



Figure 6.2: The *Podify*'s interface with full-text transcript inspection. It is accessed by clicking component (B) of Figure 6.1. The captions (component (A)) change from a sentence-level to a word-level granularity. (C) is the exact word-match search functionality.

For the scope of this study, I considered the 2020 data (as the *qrels* for the 2021 version of the task were not available) and two search intents, *TO* and *KI*, as detailed in Section 6.2. I excluded the *RE* search intent due to its overlap with *KI* and the subsequent merge into a single type in the 2021 collection [143]. This approach supported the generalisation of my findings and compatibility with future work using the 2021 collection.

### 6.4.1 The Relevance Assessments (Grades)

Relevance assessments (*qrels*) of the segments (two-minute snippets) were conducted by the National Institute of Standards and Technology (NIST). These assessments served as the gold standard for comparison with the participants' judgements on *Podify* during their experimental tasks. The *qrel* assessments were made on the PEGFB graded scale (Perfect, Excellent, Good, Fair, Bad) and as follows [1]:

- **Perfect (4)**: the earliest entry point into the episode that fulfils the user's IN.
- **Excellent (3)**: highly relevant information.
- **Good (2)**: highly-to-somewhat relevant information.
- **Fair (1)**: somewhat relevant information.
- **Bad (0)**: no relevance.

### 6.4.2 Segments Ranking

For each task, participants were presented with a playlist of ten segments retrieved from the 2020 TREC Podcasts Track. I controlled the complexity of the task by manipulating the ordering of these segments based on their relevance. By controlling the ordering of these top@10 segments, I controlled the complexity of the task and created a realistic scenario where recommendations were more (*Ea*) or less (*Di*) relevant to the participant's IN. It is important to note that I did not disclose the task's complexity

[1] Descriptions available at: https://trecpodcasts.github.io/participant-instructions-2020.html

to the participants. The top@10 segments had the following assessments (where 444 indicate three segments with agrade of four):

- *Ea* Task: 444 - 333 - 2222
- *Di* Task: 22 - 11 - 000000

Hence, with my experimental design, a *Di* task contained little to no relevant information. On the other hand, an *Ea* task contained highly relevant information that can be effectively leveraged to complete the task.

I also introduced an additional layer of realism in my design and to closely mimic a typical recommendation scenario (where least relevant items may appear in the early ranks due to, for example, diversity control measures). From the aforementioned assessments list, I employed arbitrary normalized discounted cumulative gain (nDCG) values of 0.9 and 0.5 for the *Ea* and *Di* tasks, respectively. This resulted in the two ordering schemes used during the experimental tasks:

- *Ea* Task: 4 2 2 4 2 2 4 3 3 3
- *Di* Task: 0 0 1 0 2 2 1 0 0 0

### 6.4.3 Topics

In my experiment, I selected eight topics, as presented in Table 6.1. These were equally divided between the two search intents: four topics related to the *TO* intent, while the remaining four to the *KI*. The selection process began with a candidate pool comprising 41 *TO* and 8 *KI* topics [2]. From this pool, I first excluded segments linked to episodes without valid RSS feeds.

Following this, I analysed the remaining topics and the corresponding segments based on the distribution of their grades. For the *Ea* topics, I randomly selected two topics that fulfilled the requirement of having at least three grades of 4, three grades of 3, and four grades of 2 in their *qrels*, as described in Section 6.4.2. Conversely, for

[2]The complete list of topics is available at https://trecpodcasts.github.io/resources/podcasts_2020_topics_train.xml and https://trecpodcasts.github.io/resources/podcasts_2020_topics_test.xml

Table 6.1: The selected topics with their task description. The topics are organised by complexity (*Ea* and *Di*) and search intent (*TO* and *KI*).

| | Compl. | Topic | Task Description |
|---|---|---|---|
| TO | *Ea* | social media marketing | *I'm looking for tips and suggestions for creating a social media marketing strategy. What kinds of things do I need to think about? How can I reach the most people online? Are there companies that can help?* |
| | *Ea* | hvac industry environmentalism | *What is the HVAC industry doing to promote sustainability and "green" technology? Has it been effective? What is the effect on people working in the industry?* |
| | *Di* | how to cook turkey | *I'm looking for podcasts that talk about how to cook turkey well, in particular podcasts that have some information on how I can make my turkey juicier.* |
| | *Di* | causes and prevention of wildfires | *2019 saw a large number of wildfires, in Australia, California, and the Amazon. What were people saying about them? What caused them? How could they be prevented? I am interested in news reports but also speculation, rumor, and unverified information.* |
| KI | *Ea* | malcolm x biography | *I heard about a podcast biography of Malcolm X and I would like to listen to it.* |
| | *Ea* | nefertiti | *I want to hear a podcast I know about that is about Nefertiti.* |
| | *Di* | recommended books for entrepreneurs | *I was told about a podcast that recommended good books for entrepreneurs. I'm trying to find that episode.* |
| | *Di* | bias in college admissions | *I read an article that mentioned a podcast about bias in college admissions. I would like to listen to it but I don't know the name of the show.* |

the *Di* topics, I randomly selected two topics from the candidate set (excluding the already selected *Ea* topics) that met the requirement of having at least two grades of 2, two grades of 1, and six grades of 0 in their *qrels*. This approach ensured that each chosen topic contained enough *qrels* to comply with the top@10 ranking specifications outlined in Section 6.4.2.

To mitigate potential topic bias in system selection, two topics were allocated to each combination of search intent and complexity. The assignment of topics was done randomly. As an example, both "social media marketing" and "hvac industry environmentalism" topics were available for assignment to (*TO.BA.Ea*) and (*TO.EN.Ea*). While some participants completed *TO.BA.Ea* with the former topic, others completed the same task as the latter topic. This randomised distribution minimised any risk of topic assignment bias, reinforcing the validity of my responses to the RQs.

### 6.4.4 Playlist Generation

To create the playlists for your experiment, I devised the following procedure. For each identified topic, I randomly selected ten segments from the TREC *qrels*, in line with the grades described in Section 6.4.2. For instance, for the "social media marketing" topic in the *Ea* category, I sampled three segments of grade 4, three of grade 3, and four of grade 2.

These selected segments were then used as input for the catalogue creation procedure of *Podify* (see Chapter 5, Section 5.2.3). By leveraging the functionality of *Podify*, I generated the catalogue, incorporating all the identified segments. Subsequently, I used *Podify*'s "systems" to create different platform variations based on the search intent (*TO* or *KI*), system (*BA* or *EN*), and complexity (*Ea* or *Di*). In total, I created 16 unique "systems", representing the different combinations of search intents, system, complexity, and topic (i.e., 2 search intents $x$ 2 systems $x$ 2 complexities $x$ 2 topics per task).

During session execution via the survey platform, changing a participant's platform "system" with the provided URLs auto-generated a default playlist for the user. This playlist followed the top@10 recommendation procedure, with the segments arranged

as specified by the task's complexity and detailed in Section 6.4.2.

To ensure the necessary RSS metadata's availability, I used the Spotify Podcast Dataset [35] to complement my retrieved segments with the required information. Finally, it is important to note that only the two-minute segments (and not the entire podcast episode) were made accessible for listening to the participants.

## 6.5 Questionnaires

The participants encountered three different questionnaires at various stages of the experiment, designed to collect insights into their backgrounds, experiences, and perceptions.

At the beginning of the experiment, participants were introduced to an *entry questionnaire* (see Appendix B.4). This questionnaire comprised three distinct sections. The first section delved into standard background and demographic details. The next section explored the participants' prior engagement with streaming services, their weekly podcast consumption estimates, concurrent activities while listening, main reasons for listening, and the factors that they consider important when considering a new podcast. The third and concluding section evaluated their familiarity and past interactions with existing podcast streaming platforms. This information was collected to estimate their level of familiarity with such services and the experiment's tasks. To minimise any bias, the order of the questions in the second and third sections was randomised. Finally, to reduce attrition, the demographic questions were placed at the beginning of the experimental procedure [237].

After each task, participants were asked to complete a *post-task questionnaire* (see Appendix B.6). In this questionnaire, I elicited their opinions on diverse facets of the procedure, including all the UE dimensions. The first section focused on the participants' perception of the task they had just undertaken.

Subsequently, their thoughts on the playlist's segments and their influence on task completion were collected. The third and final section covered the UE aspects related to the podcast system experience. All questions in these questionnaires were of a forced-choice type, with their order randomised to mitigate ordering effects.

Finally, an *exit questionnaire* was introduced at the end of the study (see Appendix B.7). This questionnaire aimed to gather information regarding the user study, tasks, and systems. For example, participants were asked whether they found some tasks more challenging than others and which system they preferred. They were also given the opportunity to provide general comments and feedback about the user study.

## 6.6 Qualitative and Quantitative Measures

To evaluate UE, I considered the six dimensions introduced by O'Brien and Toms [195]: perceived usability, aesthetics, novelty, felt involvement, focused attention, and endurability. The *post-task questionnaire*, comprising a series of forced-choice questions, served as the tool to measure these dimensions.

The questionnaire was structured into 12 questions: (i) "*The task that I performed was: [difficult / easy to perform / enjoyable / interesting / involving / tiring]*", and (ii) "*The podcast system that I used was: [annoying / attractive / confusing / enjoyable / frustrating / visually appealing]*". The participants were instructed to respond on a 5-point Likert scale (1: "Strongly Disagree", 2: "Disagree", 3: "Neither Agree Nor Disagree", 4: "Agree", 5: "Strongly Agree").

To mitigate biases and fatigue effects, the order of the questions was randomised. Besides these qualitative measures, I employed quantitative measures to evaluate the impact of the textual components (i.e., the *EN* system). The analysis encompassed participant listening activities, their access to, and interactions with, the full-text transcript (evidenced by the click of component (B) in Figure 6.1 and Figure 6.2), and an evaluation of the accuracy of their relevance judgements compared to the TREC gold set. These quantitative data were collected by tracking the participants' interactions with the platform, offering a robust and comprehensive assessment of both qualitative and quantitative aspects of the study.

## 6.7 Experimental Procedure

### 6.7.1 Ethics

Ethical permission (no. 2077) to conduct this study was obtained from the CIS Ethics Committee at the University of Strathclyde, and the experiment was conducted in accordance with the ethical guidelines.

All participants were assigned a unique ID number to ensure participation anonymity. Answers in the questionnaires were automatically collected by the survey platform. Participants were not asked for any personal details, such as name, address, and location. The current study did not involve any invasive procedure for data collection. Each participant signed up to *Podify* using their randomly assigned ID. This was consequently used for all the data accumulated during the experiment, and to identify the participants in the data analysis stage. The data was kept confidential and only the results of the analysis were included in this thesis and considered for publication. If an individual's response is published, they cannot be identified since no identifiable information was collected (e.g., name). Furthermore, any material that can be potentially identify a participant (i.e., their assigned ID) was permanently replaced and destroyed after the data was collected and validated. Thereby, this guarantees the anonymisation and confidentiality of the collected data. Only data that is adequate and sufficient to fulfil my research purpose was collected. Limited or no personal data was collected as a result.

During the experiment, a log of interactions was recorded (e.g., browsing activity, search queries, listening activity by play/pause, seek forward/backward, and volume adjustments). This collection procedure is GDPR compliant[3]: IP addresses were masked, anonymity sets were used instead of cookies, and there was no link between page visits and participants. Collected data were processed with strict adherence to the Code of Practice and with the GDPR. Answers to the questionnaires and logged behavioural data in CSV format were securely stored in an encrypted network hard drive within the department of CIS at the University of Strathclyde, and under the management of CIS

[3] https://github.com/ankane/ahoy#gdpr-compliance-1

Systems Support. Finally, as per UK Research and Innovation (UKRI)'s requirements, the data will be kept for ten years, after which it will be securely disposed.

### 6.7.2 Procedure Outline

Participants were briefed about the two-session structure of the experiment and it would take approximately 240 minutes to complete. The first session was to be completed in the morning, and the second in the afternoon (as detailed in Section 6.2). Each participant was required to complete eight tasks, one for each level of independent variable (search intent, system, and complexity), and as outlined in Section 6.2. Participants received a payment of £15 upon completion. The total cost of the experiment was £360.

The session started with a general overview of the study (see Appendix B.1), including its purpose, duration, and an *information sheet* (see Appendix B.2). This sheet provided an in-depth presentation of the study, informing participants that looking at the *information sheet* did not impose any obligation, and without negative consequences, should they choose not to participate. After reviewing the *information sheet*, participants were directed to a *consent form* (see Appendix B.3). By agreeing to all the conditions stated in the form, they started the session with the *entry questionnaire* (see Appendix B.4).

After completing the *entry questionnaire*, the participants proceeded to the assigned tasks (see Appendix B.5). Each task consisted of a distinct topic and task description (as outlined in Table 6.1). Detailed instructions on how to complete the task were also provided. In particular, the participants were instructed to learn more about the provided segments by clicking on them and inspecting their episode's metadata, such as description. They were asked to listen to as many segments as necessary to complete the task, in any order, and by making use of the features provided by the podcast streaming platform (e.g., the listening controls). Further, they were also asked to use the thumbs up/down feature to show relevance and provide feedback for the segments they listened to. The task was preceded by a brief training video, designed for the system (i.e., *BA* or *EN*), highlighting the most important UI features using an

example task. Participants were subsequently redirected to *Podify* for task execution. On *Podify*, the participants were presented with an automatically generated playlist of ten podcast segments that were relevant (to different degrees) to the task. This was based on the task's system (*BA* or *EN*) and complexity (*Ea* or *Di*). This approach aimed to simulate a realistic scenario of podcast consumption and recommendation.

After completing each task, participants were redirected back to the survey platform, where they were asked to fill in a *post-task questionnaire* (see Appendix B.6). To mitigate the impact of fatigue, the order of the questions in the *post-task questionnaire* was randomised. Once all tasks in the session were completed, participants were then asked to fill in an *exit questionnaire* to conclude the session (see Appendix B.7).

### 6.7.3 Participants

To conduct this user study, 24 participants were recruited. In the *information sheet*, the participants were informed about the importance of accurately assessing the relevance of the segments they listened to (see Section 6.7.2). Their judgements were then compared with the TREC 2020 relevance gold set *qrels* for performance evaluation.

All the recruited participants completed the experimental tasks by following the instructions and thus successfully completing the experiment. 54.2% identified as males and 45.8% as females. The age distribution of participants indicated all participants were under the age of 41, with the largest age group between the ages of 24-29 (70.8%), followed by the group between 30-35 (16.7%). In terms of educational background, participants had a bachelor's degree (45.8%), master's degree (25%), PhD (20.8%), or a high school diploma or equivalent (8.3%) The majority of the participants were employed by a company or organisation (66.7%) or students (29.2%), with the remaining identifying as not employed. Last, 54.2% of the participants reported having a native speaker level of English. The remaining 45.8% described themselves as fluent.

### 6.7.4 Pilot Studies

Prior to conducting the main study, I ran two pilot studies using a total of eight participants. The purpose of the pilot studies was to gather detailed feedback from

the participants. This feedback was then used to adjust the study's presentation and design. Based on the feedback received, changes consisted of moving the full-text transcript access button (component (B) in Figure 6.1) closer to the listening controls. After the second pilot study, it was determined that the participants could complete the user study with no difficulties and that the platform accurately logged all the required interaction data. The data from these pilot studies were not included in the analyses presented in Chapter 7.

## 6.8 Chapter Summary

Chapter 6 provided information regarding the methodology used in this thesis, specifically for the analyses presented in Chapter 7. In particular, this chapter outlined:

- **Study Design** (Section 6.2): introduced the study's design that revolves around three main independent variables: search intent, system, and task complexity. This design allowed us to simulate realistic IN and recommendation scenarios as closely as possible.
- **Apparatus** (Section 6.3): detailed the technical setup and apparatus, including the specially customised version of the *Podify* podcast streaming platform, which were central to the experiment's design and execution.
- **Topics & Corpus** (Section 6.4): presented the study's topics and their respective podcast segments, including details on grading, ranking and playlist generation for the experimental tasks.
- **Questionnaires** (Section 6.5): provided an overview of the questionnaires' design, deployment, and objectives.
- **Qualitative and Quantitative Measures** (Section 6.6): delved into the diverse measures adopted to assess UE.
- **Procedure** (Section 6.7): outlined the procedural steps, encompassing ethical considerations, pilot studies, and participants' recruitment.

# Chapter 7

# Influence of Text for Assessing Content Relevance in Podcast Information Access

## 7.1 Introduction

Podcasts have recently emerged as a significant medium for online information-seeking, a trend noted in Chapter 2, Section 2.5 [3]. Despite their increasing popularity and widespread recognition, the podcast domain remains a relatively under-researched domain [3]. For example, their recent incorporation into unified platforms with the music media has created numerous challenges and research opportunities. In particular, these platforms require robust and effective search systems to aggregate diverse content into user-friendly UIs [4, 32]. The coexistence of diverse media within a single platform raises questions about the optimal design of audio-focused information access systems. Specifically, there is a need for dedicated research to understand user behaviour and optimise podcast streaming platforms. This can be achieved by considering the unique characteristics of podcasts and the concept of relevance in this context [4, 33]. Another key unexplored area is the concept of "relevance" in the context of podcasts [3, 36]. In this chapter, we seek to investigate the concept of podcast relevance by exploring its relationship with user context (i.e., search intent and task complexity) and observed

implicit feedback (i.e., listening activity).

One of the main challenges in podcast IR lies in identifying specific, relevant information within podcast episodes. This issue was the focus of the 2020 and 2021 TREC Podcast Track, which attracted numerous submissions to its tasks of retrieval of fixed two-minute segments and episode summarisation [34]. The track utilised the Spotify Podcast Dataset, which contains over 100,000 episodes featuring audio files, transcriptions, metadata, and RSS feeds [35]. These auto-generated transcriptions through ASR systems allow content-based search and user navigation. However, they pose challenges to standard IR methods because of their length and errors [36, 37]. Therefore, there is a need to segment podcast episodes into, for example, fixed two-minute chunks [34]. These chunks, referred to as segments, are two-minute snippets of a podcast episode.

Transcriptions hold the potential to transform podcast accessibility and engagement across diverse audiences and domains. They facilitate access to content for the hearing-impaired community, aligning with principles of Universal Design [9,10], which support the ability to cater to various learning and comprehension styles through theories such as Dual Coding [38–40] and the Cognitive Theory of Multimedia Learning [41]. This multi-modal approach also aids in the retention of complex information [11, 12]. They also serve as a useful tool in conducting research, since they enable effective pattern recognition, thematic analysis, and support grounded theory methodologies [42, 43]. Searchable text enhances IR capabilities and aligns with the Information Foraging Theory. It aids in navigation and helps users to find information efficiently [7,8]. These facets motivate my investigation into incorporating captions and full-text transcripts into the UI of the *Podify* podcast streaming platform (Chapter 5).

This chapter focuses on examining the impact of incorporating textual components, specifically captions and full-text transcripts, on user experience and engagement. In particular, I comprehensively analyse their impact on the users' process of assessing the relevance of podcast segments to their INs. By combining qualitative (the participants' reported relevance judgements of podcasts) and quantitative (listening activity) data, I investigate the importance of these textual components in enabling users to better assess the relevance of podcast content.

### 7.1.1 Research Questions

This chapter aims to investigate the following two important RQs:

- **RQ-7.1**: Does the inclusion of captions and full-text transcripts improve the user experience and engagement in podcast information access?
- **RQ-7.2**: How do these textual components affect the users' ability to accurately assess the relevance of podcast content?

I investigated my RQs by conducting a user study on the *Podify* streaming platform (see Chapter 5, conducted in alignment with the methodology outlined in Chapter 6).

### 7.1.2 Contributions

The contributions of this chapter are two-fold:

- I demonstrate the novel approach of incorporating text-based components, specifically captions and full-text transcripts, in the UI design of the *Podify* podcast streaming platform. My comprehensive evaluation considers multiple measures of user experience and engagement, providing insights into the potential for improved content comprehension, effective IR, and enhanced navigation. This exploration advances the understanding of how text-based components can enrich the users' experience in a podcast information access context.
- I present the first work that rigorously analyses the users' relevance assessment process in the podcast context. By examining the participants' accuracy, I highlight the positive influence of text-based components in enhancing their abilities to accurately judge the relevance of podcasts. This investigation reveals new insights into the users' behaviour and interaction with podcasts, crucial for the development of novel user-centric interfaces.

This chapter is organised as follows. Section 7.2 details the experimental settings. The experimental results are reported in Section 7.3, followed by a summary and discussion in Section 7.4.

## 7.2 Experimental Settings

This section outlines the experimental settings, focusing on how the relevance judgements were collected and assessed, the accuracy measure employed, and a description of the specific textual components integrated into the *Podify* platform and investigated in this study.

### 7.2.1 Relevance Judgements: Relevant and Non-Relevant

The process of collecting relevance feedback in this study consisted of diverse interaction methodologies within the *Podify* platform, as illustrated in Chapter 6, Figure 6.1 and outlined in Chapter 6, Section 6.7.2. Participants' judgements were collected via thumbs up/down and textual feedback that used a star rating scale that ranged from 1 to 5. This interaction data was subsequently compared with the TREC gold standard grades that are described in Chapter 6, Section 6.4.1 for a performance evaluation of how the user performed in the task. Relevance, inherently subjective, is influenced by a user's perception of information pertinent to their current IN [238]. The binary relevance scale (relevant vs. non-relevant) has been the prevalent approach in the IR field due to its guarantees in stability during measurement [238–240]. Therefore, in this work, we investigate the concept of podcast relevance under a binary scale.

In my analysis, I considered segments that received thumbs up by the participants as relevant, and those with a thumbs down as non-relevant. Similarly, star ratings greater than or equal to three were deemed relevant, with lower scores classified as non-relevant. It is important to note, however, that by analysing the logged behaviour of participants, I had to account for the case of participants providing multiple feedback for the same segment, such as a sequence of "*relevant, non-relevant, relevant, non-relevant*". To address this, and ensure consistency, my analysis focused solely on the final feedback given and also by excluding any judgements made before the segment was played. This is to remove any potential noise in my analysis. Thus, in the sequence mentioned above, I consider the judgement provided by the participant to that specific segment as being a non-relevant judgement.

Participants could also remove prior assessments. Such "feedback removal" process was considered in my analysis. If the last feedback was a removal, I considered that the participant deemed the segment as unassessed in terms of relevance. This may correspond to accidental feedback, later rectified by the participant.

The final stage of the process involved aligning these judgements with the TREC grades. Segments with associated *qrel* grades of 4, 3, or 2 (representing perfect, excellent, and good) were classified as relevant, whereas those with grades of 1 or 0 (fair and bad) were categorised as non-relevant.

### 7.2.2 Evaluation Metric

A participant's accuracy in the context of relevance judgements is determined by the mutual agreement between their assessments and the TREC gold set. In other words, their provided relevance judgements for specific segments are compared to the gold set provided by TREC. A judgement is accurate when both the participant and the TREC gold set align in their assessment of a segment's relevance (e.g., both the participant and the gold set describe a segment as relevant). On the other hand, an inaccurate judgement by the participant is with a misalignment (e.g., the participant deemed a segment to be relevant, but the TREC gold set described the segment as non relevant). The results presented in this study reflect the mean and standard deviation of accuracies across all individual participants, providing a statistical summary of how well the participants' judgements matched the established TREC standards.

### 7.2.3 The Textual Modality Components

In this study, the textual components that have been investigated were the captions and the full-text transcripts. The textual data for this study was sourced from the Spotify Podcast dataset, which included both the captions and the full-text transcripts. The latter were utilised as provided. However, the former required specific processing to align with the study's needs. This processing involved segmenting the full-text transcripts into five-second intervals to generate captions that accurately represented the spoken content within these time frames. Each caption was thus created to provide

the dialogue over these brief periods, ensuring that they provided a clear and concise textual snapshot of the podcast content. I aim to understand the role of these components in how participants provide relevance judgements and interact within the *Podify* platform. Each of these components, detailed below, was systematically categorised and analysed.

**Captions.** They are the brief textual descriptions that accompany podcast segments (component (A) in Chapter 6, Figure 6.1). They provide a textual representation of the current spoken content, offering participants a snapshot of the current discourse. Within the *EN* system, any judgements made by participants were automatically categorised as being related to the role of captions.

**Full-Text Transcript.** They provide a more comprehensive textual representation of the podcast segments. While the captions offer glimpses of the current spoken material, the full-text transcripts encapsulate the entire dialogue and content of the segment. Within the *EN* system, logged behaviour was analysed to detect access and utilisation of this component on the assessment of segments. These judgements were then categorised as being related to the role of full-text transcripts.

## 7.3 Experimental Results

In this section, I present the results of my study. I begin by exploring the data gathered by in the questionnaires, and the participants' overall perception of the user study. Then, I evaluate the two observed systems (*BA* and *EN*), examining aspects such as participants' preferences, engagement, and task perception within each system. Importantly, I provide evidence that the users' perception of the *EN* system remains unaltered, affirming that it does not pose a UI challenge. This allows us to focus on the additional value that these textual components bring to task completion. This investigation delves into the relevance assessments process, providing insights on how the textual modality aids participants in more accurately assessing the relevance of podcast segments.

### 7.3.1 Participant Questionnaires

The findings presented in this section are informed from a targeted *entry questionnaire* (refer to Appendix B.4). This questionnaire was formulated to gather data on the participants' experiences with streaming platforms, their podcast consumption patterns, the activities they undertake during listening, the primary motivations for listening, and the factors they consider important when choosing new podcasts. Additionally, I collected their familiarity with, and the nature of, their previous engagements with existing podcast streaming platforms. Details on participant demographics, such as age and education, have been previously presented in Chapter 6, Section 6.7.3.

#### Consumption Estimates and Utilisation of Podcast Streaming Services

The participants' habits surrounding podcast consumption varied, particularly regarding the time of day and their estimated weekly listening duration. When queried about their preferred podcast listening time, 50% indicated the evening hours (18-23), with the morning (6-11) and afternoon (12-17) were equally preferred (25% each) by the participants. In terms of weekly consumption duration, 29.2% of participants primarily listened for two hours, and 20.8% for one hour. The rest had a diverse range, with several participants committing over four hours weekly; notably, three participants indicated listening for over ten hours.

Considering previous utilisation of podcast streaming platforms, Spotify was the most popular, having been used by 75.0% of the participants. It was closely followed by YouTube and Apple Podcasts at 62.5% and 41.7%, respectively. Audible had been used by 16.7% of participants, with Google Podcasts being the least used (4.2%).

#### Activities and Reasons while Listening to Podcasts

Figure 7.1 shows the distributions of activities participants undertake while listening to podcasts. 79.2% of the participants indicated they listen to podcasts during housework or chores, which includes activities such as cooking. Utilising podcasts as a travel medium is popular, with 66.7% listening while riding public transportation. Leisure time is also an important activity, as reported by 62.5% of the participants. Podcasts

are also utilised during physical activities: 41.7% of the participants use this medium while walking or riding a bike, with 37.5% during outdoor activities such as walking, running, or walking dogs. Additionally, 37.5% find podcasts ideal for relaxing before going to sleep. 33.3% listen to podcasts while studying or working. Finally, 25% incorporate podcasts into meal times or driving sessions, and a 12.5% choose to only focus on listening.

In terms of the motivations behind podcast listening, Figure 7.2 highlights that hobbies and interests are the most common motivations (87.5%). The desire to learn and explore new topics is indicated by the participants as another important reason (79.2%). Furthermore, 45.8% exploit podcasts as a tool to gain practical knowledge and skills. The entertainment and relaxation motivates are reported by 62.5% of the participants. 29.2% listen to podcasts because of the hosts or guests involved. Finally, emotional companionship via podcasts is a less prevalent motive (8.3%).

#### Important Factors when Considering a new Podcast

Participants were asked about the important factors when considering a new podcast. My results show that the episode description was the most predominant factor, with 83.3% of the participants deeming it crucial when considering a new podcast. This was closely followed by the episode title, which was considered important by 79.2% of the participants. The interviewed guest was also deemed important, as 41.7% reported it was important that they had previously heard of them. Similarly, 33.3% of participants would consider a new podcast based on their familiarity with the presenter. Other factors, such as the podcast's ratings and reviews, and the frequency of new episode releases were deemed less important (25.0% and 16.7% of the participants, respectively).

#### Familiarity and Experience with Existing Streaming Platforms

Figure 7.3 shows the participants' responses to how they rated their experiences across various aspects in existing podcast streaming platforms. Responses were gathered using a 5-point Likert scale, as described in Chapter 6, Section 6.6.

On the topic of finding general information (i.e., *TO* search), participants reported

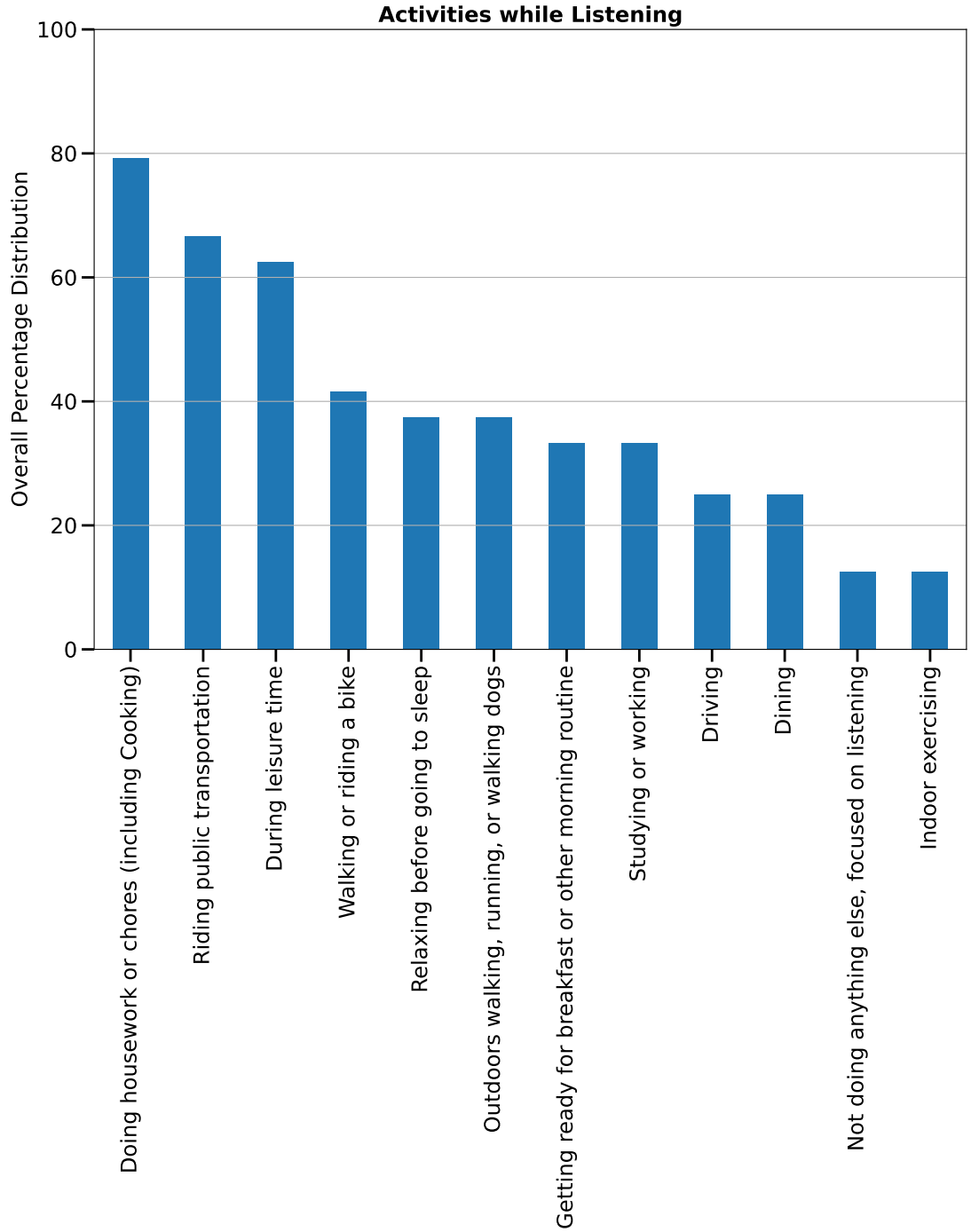



Figure 7.1: Distribution of the participants' activities usually performed while listening to podcasts and based on the responses collected through the *entry questionnaire*.

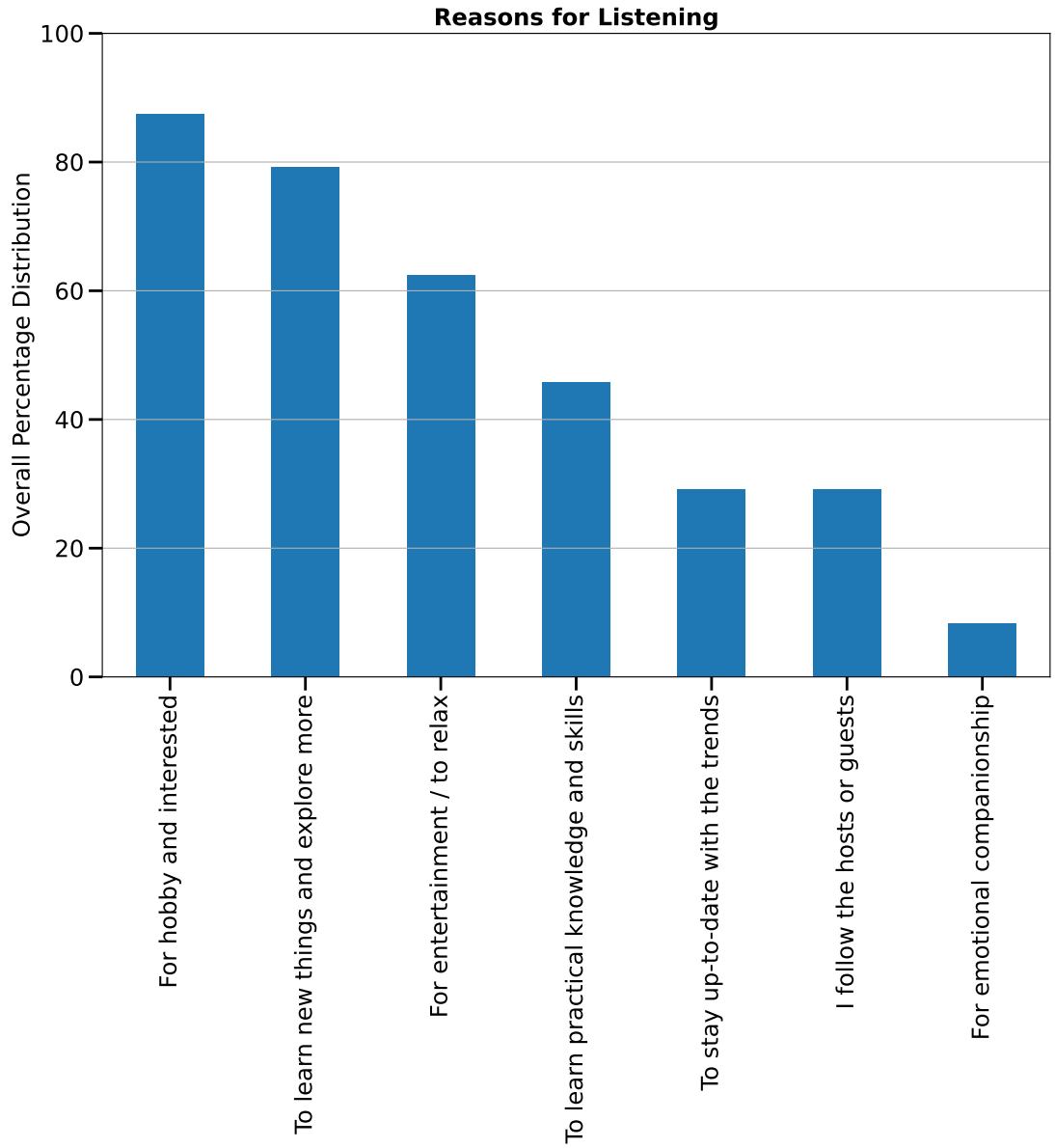



Figure 7.2: Distribution of the reasons for listening to podcasts and based on the responses collected through the *entry questionnaire*.

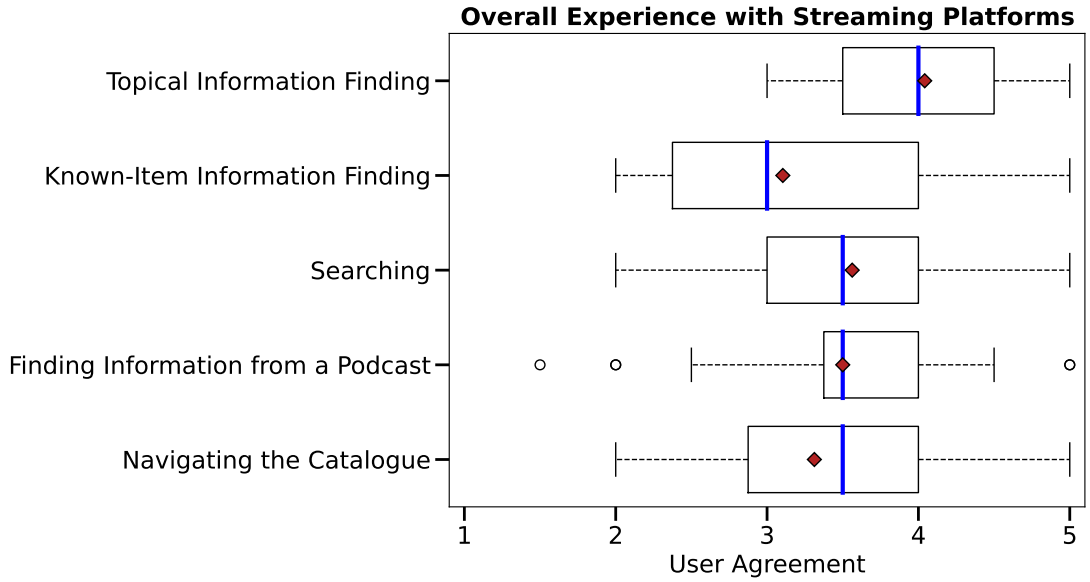



Figure 7.3: Box plot of the participants' overall experience and based on the responses collected through the *entry questionnaire*. The higher the value on the x-axis, the higher the level of user agreement. The diamond represents the mean value.

a high satisfaction level (M = 4; SD = 0.7). However, finding something that is known to exist, but under an unknown name (i.e., *KI* search), participants indicated a lower satisfaction rate (M = 3.1; SD = 0.9) with existing streaming services. Searching for podcasts is reported having a slightly better experience (M = 3.6; SD = 0.8). Accessing specific information within a podcast episode also received a comparable satisfaction level (M = 3.5; SD = 0.9). Last, when navigating through the platforms' catalogues, my results indicate the participants find the existing search functionalities not to be accurate. This is indicated by a lower satisfaction level (M = 3.3; SD = 0.9).

### 7.3.2 Study Perception

Before analysing the main results, I examine the participants' perception of the study and their overall experience during the experiment. To assess this, in the *exit questionnaire*, the participants were asked to rate their agreement to the following questions: "*I feel that, during the study, I was: [at ease with the procedure / bored / comfortable / given clear instructions / given enough time / interested / motivated / tired / under pressure]*". Responses were gathered using a 5-point Likert scale, as described in

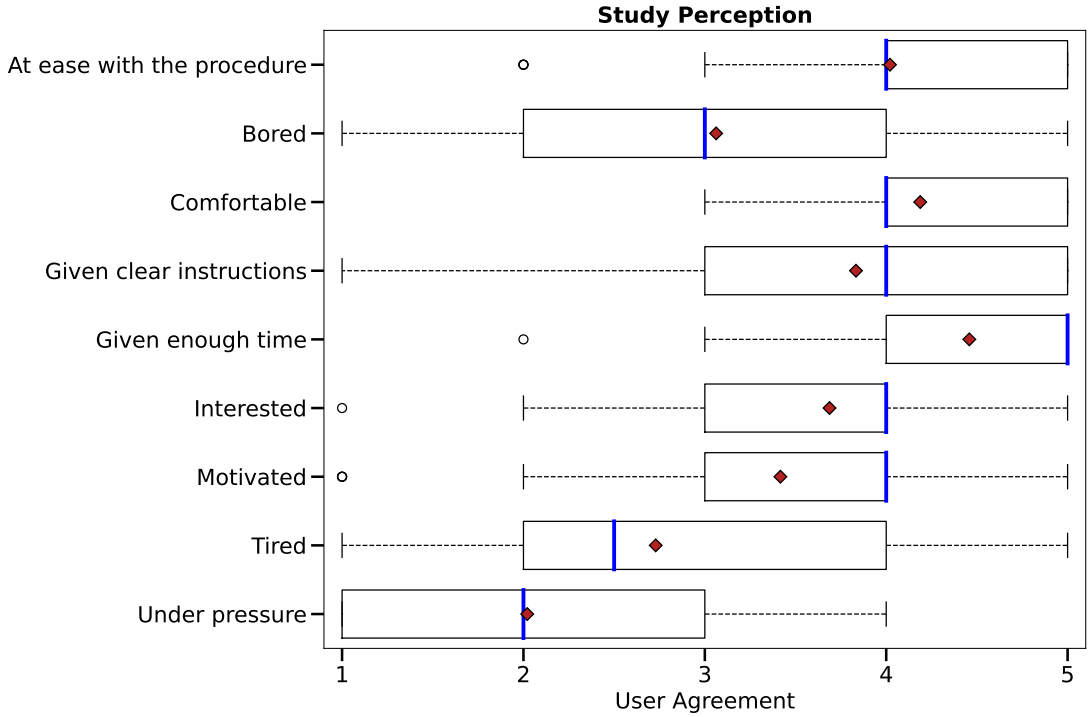



Figure 7.4: Box plot of the participants' perception of the study and based on the responses collected through the *exit questionnaire*. The higher the value on the x-axis, the higher the level of user agreement. The diamond represents the mean value.

Chapter 6, Section 6.6.

Figure 7.4 reveals a positive overall perception of the study by the participants. They expressed having had enough time to complete the tasks (mean score, M = 4.46; standard deviation, SD = 0.77), feeling comfortable (M = 4.19; SD = 0.64), and at ease with the procedure (M = 4.02; SD = 0.86). These results show the clarity of the study and instructions (M = 3.83; SD = 1.12). Additionally, participants reported being interested (M = 3.69; SD = 0.9) and motivated (M = 3.42; SD = 1.13) in completing the study. This positive experience was not hindered by excessive feelings of pressure (M = 2.02; SD = 0.96) or fatigue (M = 2.73; SD = 1.16), despite these results reporting a minor level of tiredness. While there is a slight overall agreement towards feeling bored (M = 3.06; SD = 1.04), it appears to not affect the overall engagement or investment in the study.

In summary, these findings suggest that the experimental design was well-received,

and the participants were largely engaged and invested in their participation.

### 7.3.3 System Evaluation

In this section, I aim to address my first RQ: how does incorporating captions and full-text transcripts enhance the user experience in a podcast information access context? (**RQ-7.1**). To this end, I evaluate the *BA* and *EN* systems, examining the participants' preferences, engagement, and task perception within each system.

#### System Preferences

The *exit questionnaire* asked participants to indicate their preferred system by answering the question "*Which system did you prefer?*" (answer options: 1: Baseline, 2: Enriched). Overall, 75% of the participants preferred the *EN* system over the *BA* system. The participants also elaborated on their choices through comments on their preference.

Participants highlighted the advantages of incorporating the textual components alongside the audio modality. This integration facilitated greater concentration on the content, aiding comprehension when speech was unclear because of accents or other factors. As one participant noted, "*I preferred enriched, as I could focus on the transcripts and I also felt the system transitioned better between segments*". The presence of text was also reported to be a useful aid to more swiftly assess the relevance of segments. One participant reported that "*the enriched system made it faster to complete the task, [since] I could skim the transcript to see if the segment was relevant to the question at hand*". The *EN* system was also valued for its ability to enhance engagement with the podcast content. Participants detailed how the presence of transcripts aided them in following the speech, finding specific words, and focusing on relevant details. Integrating text alongside the audio modality enabled them for a more efficient processing of information. This was noted by a participant as follows: "*I process information quicker in visual form rather than audio form, so the transcripts were helpful*". Finally, some participants also reported concerns about the accuracy of the transcriptions. Whilst this is recognised and attributed to the errors introduced by the ASR system (18%

error was reported for the Spotify Podcast Dataset [35]), participants acknowledged that they were still able to infer what was being said. This was noted by the following comment from a participant: "*the transcript may be useful for finding out specific words, but it must also be remembered that it was far from 100% accurate*".

A further analysis, shown in Figure 7.5, delves into the participants' perception of the text-based components on the *Podify*'s UI (and thus their addition to the *BA* system), as informed by their agreement to the following questions in the *exit questionnaire*. "*With regards to completing the task, having the transcript was: [informative / unhelpful / easier / useful / irrelevant / engaging / undesirable]*". Responses were gathered using a 5-point Likert scale, as described in Chapter 6, Section 6.6.

The results reveal that the text components were perceived to be useful (M = 4.23; SD = 0.88), informative (M = 4.08; SD = 0.87), and they facilitated the completion of the tasks (M = 4.06; SD = 0.86). Additionally, they were reported to make the UI more engaging (M = 3.69; SD = 0.99) with regards to task completion. Moreover, there is a strong indication that participants did not perceive the text modality as irrelevant (M = 1.98; SD = 1.08), undesirable (M = 1.85; SD = 0.87), or unhelpful (M = 1.79; SD = 0.87). These results further reinforce the positive role and perception of integrating the text modality on the *BA's* UI (i.e., the *EN* system) for improved task completion.

Overall, my analyses reveal a strong users' preference towards the *EN* system. The integration of the textual components with the audio content are reported to have improved the users' comprehension and engagement, and they were desirable with regards to task completion. This is despite the minor concerns about transcription accuracy, which is an intrinsic property of the Spotify Podcast Dataset and ASR systems. The insights from the *exit questionnaire* and the detailed analysis further corroborate the potential of the text modality in augmenting the user experience with the podcast content.

### Engagement Perception within Systems

To assess the UE levels with both the *BA* and *EN* systems, I analysed the responses to the following question from the *post-task questionnaire*: "*the podcast system that I used*

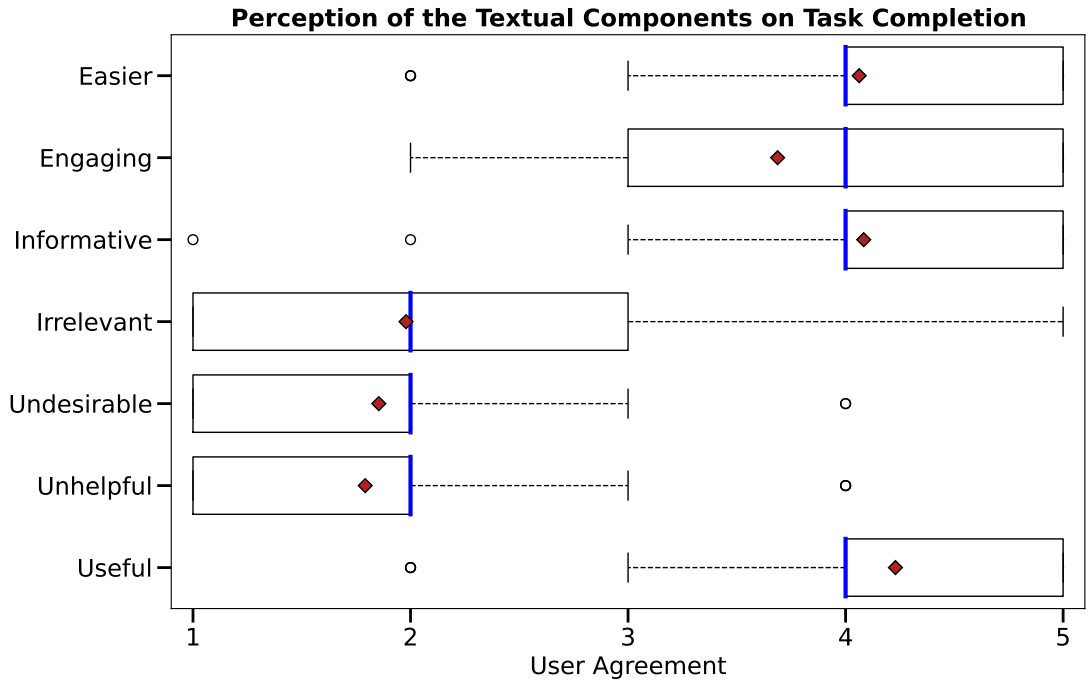



Figure 7.5: Box plot of the participants' perception of having access to text-based components on the *Podify*'s UI and based on the responses collected through the *exit questionnaire*. The higher the value on the x-axis, the higher the level of user agreement. The diamond represents the mean value.

*was: [annoying / attractive / confusing / enjoyable / frustrating / visually appealing]*". Figure 7.6 presents the box plot for the UE analysis of the two systems, employing the Likert rating scheme outlined in Chapter 6, Section 6.6.

The results indicate that participants perceived the *EN* system to be less annoying and more attractive, enjoyable, and visually appealing than the *BA* system. Interestingly, while *EN* was perceived as more confusing, it was found to be less frustrating than the *BA* system. This positive perception of the *EN* system's attractiveness, enjoyment, and visual appeal seem to indicate a positive response towards its UI design. The increased confusion reported by the participants may be attributable to the novelty of the text-based components explored in this work. This is because existing podcast streaming platforms do not include such features. Therefore, this may have initially caused some disorientation. Despite this, the overall perception of the *EN* system is positive and favourable over the *BA* system. The higher confusion is not reported to hinder the user experience, with the textual components contributing positively to the overall engagement with the podcast content.

However, it is important to note that my analysis did not reveal any statistically significant differences between the perceptions of the two systems. Although the inclusion of the textual modality did not significantly enhance UE under the examined measures, it also importantly did not have a negative impact on the *Podify*'s UI. Thus, the findings suggest that the impact of the textual modality on task completion, such as assessing podcast relevance, is solely related to the unique value that the textual components provide. This is without negatively affecting the overall user experience.

### Task Perception within Systems

Last, I conducted an analysis of the participants' perception of the experimental tasks. This analysis is conducted on the responses from the *post-task questionnaire*. Figure 7.7 shows the results for the question: "*The task that I performed was: [difficult / easy to perform / enjoyable / interesting / involving / tiring]*". This analysis reveals clear differences in how the participants perceived the tasks between the *BA* and *EN* systems.

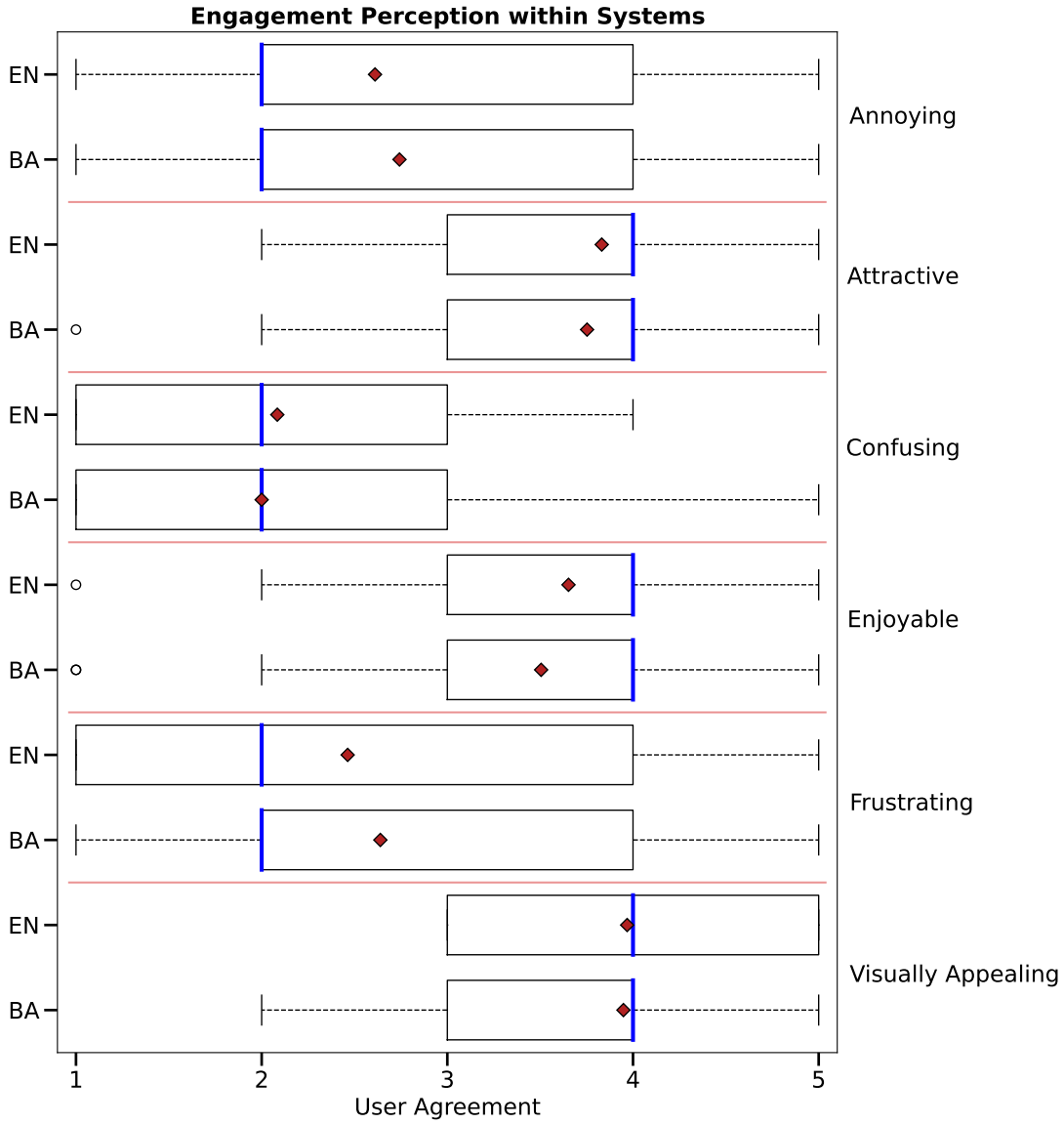



Figure 7.6: Box plot of the UE measures for system perception (*EN* and *BA*) based on the responses collected through the *post-task questionnaire*. The higher the value on the x-axis, the higher the level of user agreement. The diamond represents the mean value.

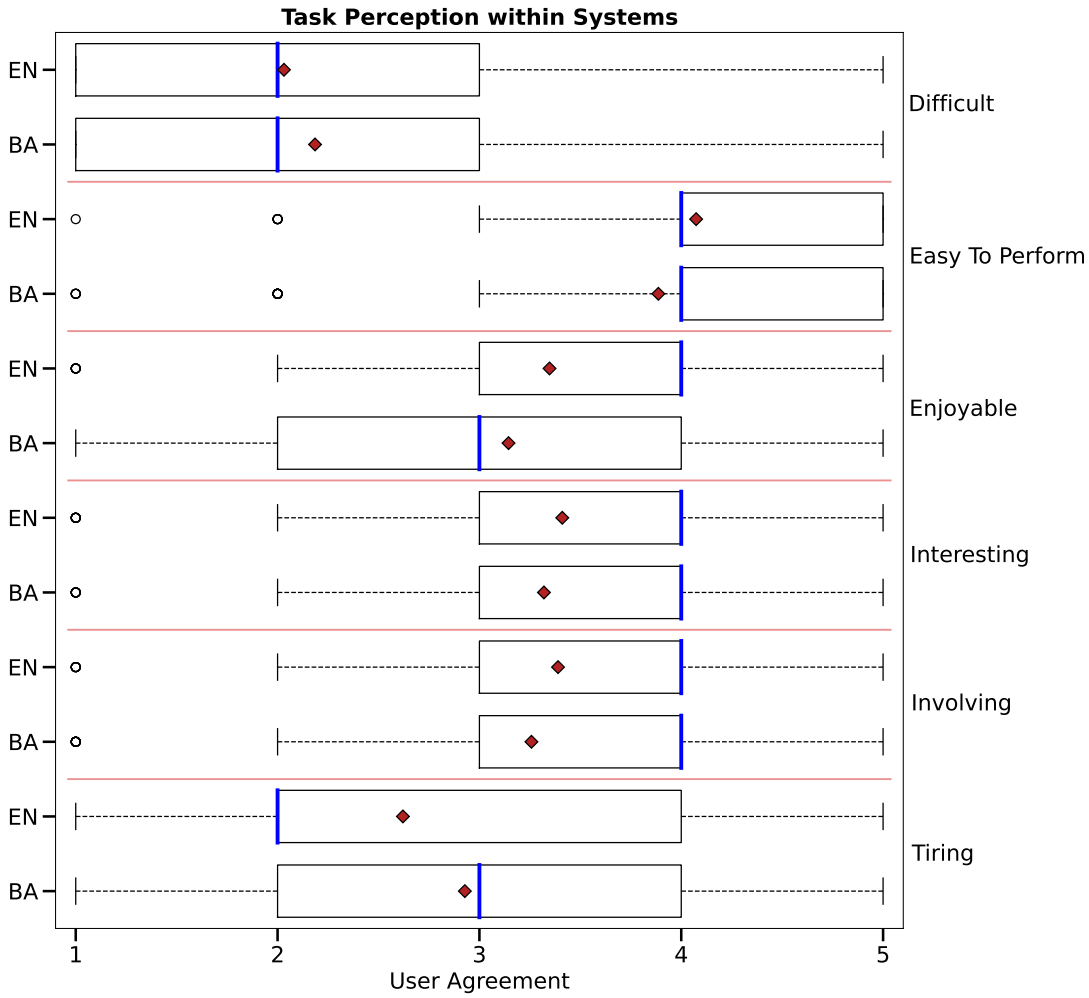



Figure 7.7: Box plot of the UE measures for task perception by system (*EN* and *BA*). This is based on the responses collected through the *post-task questionnaire*. The higher the value on the x-axis, the higher the level of user agreement. The diamond represents the mean value.

The tasks performed on the *EN* system were perceived as less difficult and tiring, yet more involving, interesting, enjoyable, and easier to perform than those on the *BA* system. These findings suggest that the *EN* system may facilitate a more engaging experience. Such positive feedback reinforces the hypothesis that incorporating textual components in the UI of podcast streaming platforms can enhance UE, without adding complexity or fatigue.

Nevertheless, it is important to note that these perceptual differences also did not translate into statistically significant differences between the two systems. This consistency with the prior findings of Section 7.3.3 further validating my claim that including of the textual modality in podcast streaming platform does not pose a UI challenge. Importantly, the enhancements attributed to the *EN* system do not negatively affect the user's experience.

### 7.3.4 Relevance Assessment Analysis

In the previous sections, I analysed the participants' study perceptions, engagement levels, and the positive impact of integrating the textual components in the *Podify*'s UI. This section delves into exploring the value these components add to task completion, specifically in how the participants assess podcast content relevance (addressing **RQ-7.2**).

Figure 7.8 shows the participants' reported performance in assessing the podcast segments, classified by search intent, system, and task complexity (see Chapter 6, Section 6.2). The participants' accuracy, classified by the three experimental independent variables, was determined by comparing their relevance judgements to the TREC gold set (see Section 7.2.1). I further categorise the assessments by the system employed: "*BA*", "*(Captions) EN*", and "*(Full-Text) EN*" (see Section 7.2.3). This enables us to investigate the particular role of captions (i.e., the *EN* system) as well as how the full-text transcripts variation of the *EN* system further influences the participants' accuracy of their assessment process.

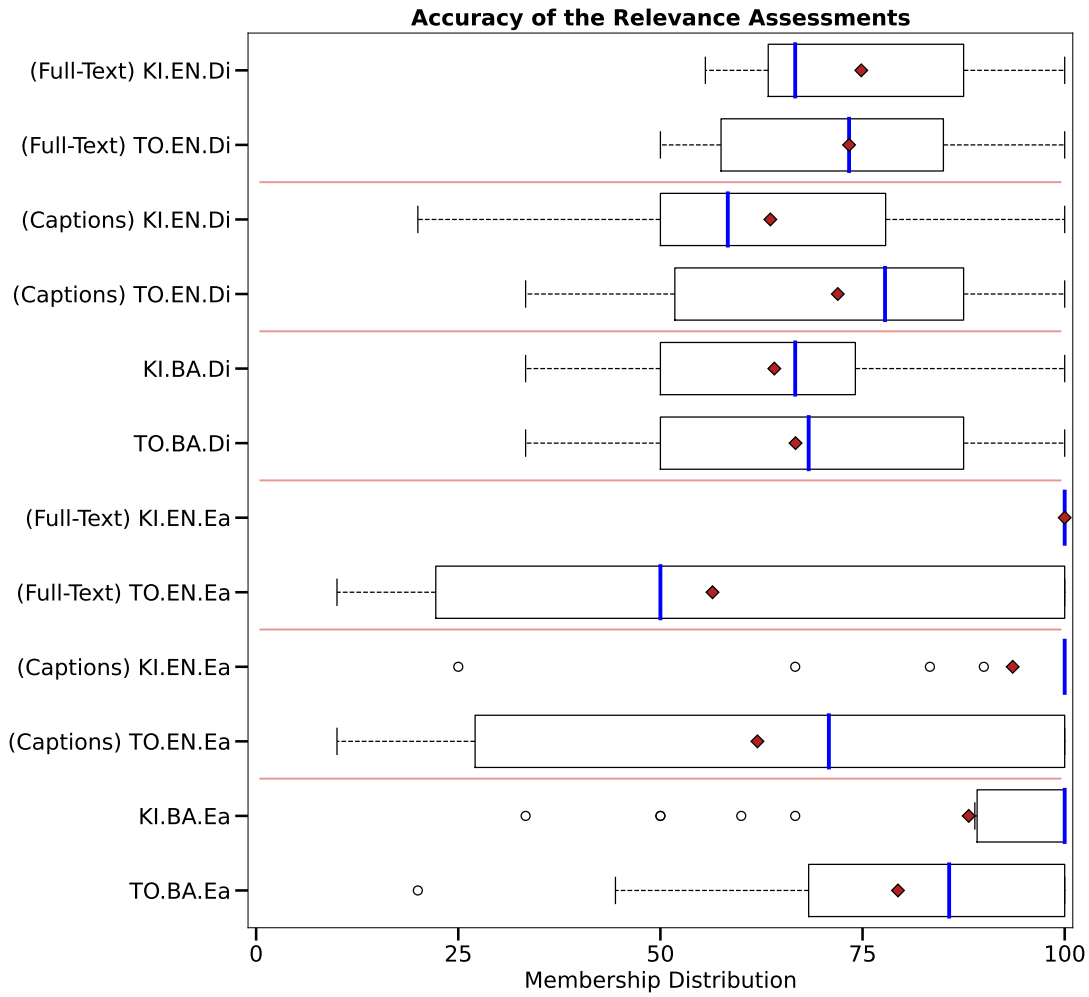



Figure 7.8: Box plot of the participants' accuracy of their relevance assessments, categorised by the three experimental independent variables: search intent (*TO* and *KI*), system (*BA* and *EN*, with the latter further categorised as captions or full-text transcripts), and task complexity (*Ea* and *Di*). The x-axis represents the membership distribution as a percentage value. The diamond represents the mean value.

### Task Complexity

Using the textual components revealed contrasting impacts on the relevance assessment process across the different task complexities (i.e., *Ea* and *Di*).

In *Ea* tasks, and with a *KI* search intent, including the textual components leads to increases in the accuracy levels. I observe increases from 88.1% in the *BA* system (*KI.BA.Ea*), to 93.6% and 100% in *(Captions) KI.EN.Ea* and *(Full-Text) KI.EN.Ea*, respectively. Conversely, the *TO* search intent registers a decrease in the overall accuracy and with higher variance (from 79.4% $\pm$ 22 in *TO.BA.Ea* to 56.4% $\pm$ 42.3 in *(Full-Text) TO.EN.Ea*). These two inverse trends may be because of an IO problem that is introduced by the *TO* search intent. The users misjudge the relevance by swiftly skimming the content and by searching for specific keywords. However, such keywords may not translate to a relevance of the segment's content to their IN. Thus, the textual components enhanced the completion of *Ea* tasks for the *KI* intent, but caused decreased accuracy and greater variance for the *TO* search intent.

In contrast, including the textual components in *Di* tasks has a positive impact. Within these *Di* tasks, the *TO* search intent initially appeared to be performed better than *KI* by the participants with an accuracy of 66.7% and 64.1% for *TO.BA.Di* and *KI.BA.Di*, respectively. However, the addition of full-text transcripts led the *KI* tasks to outperform the *TO* ones (74.8% and 73.3% for *(Full-Text) KI.EN.Di* and *(Full-Text) TO.EN.Di*, respectively). Further, despite captions are shown to improve the participants' accuracy for the *TO* search intent, it is noted how this component slightly reduced their performance in the *KI* tasks. Overall, this highlights the crucial role that text plays in search contexts, whereby full-text access is reported to achieve the highest accuracy in comparison to both captions and the *BA* system.

The contributions of the text components are clear, since they aid participants in better assessing the relevance of podcast segments. This is particularly important since a *Di* task corresponds to a case where the recommendations process is not optimal and the resulting recommendations are mostly not relevant to the participant's IN. Therefore, the text modality appears to be able to mitigate the issues of misjudging the content's relevance in the complex domain of podcast recommendations. This

improvement in relevance assessment can be attributed to the Dual Coding Theory [38–40], which posits that information is more effectively processed when presented through both verbal and visual channels. This dual modality approach reduces cognitive load [241, 242], allowing users to make more accurate relevance judgements.

#### Search Intents: Perception vs Performance

Finally, my results reveal an interesting finding when comparing the participants' performance with their self-reported experiences (see Section 7.3.1). This self-reporting data was collected through the *entry questionnaire* and for the questions "*How would you rate your experience in finding general information about a topic?*" and "*How would you rate your experience in finding something that you know exists but under an unknown name (i.e., known-item)?*" *[Very Easy / Easy / Neither Easy Nor Difficult / Difficult / Very Difficult]*. These correspond to the *TO* and *KI* search experiences, respectively.

The participants' overall performance, as shown by Figure 7.8 and the analysis conducted in Section 7.3.4, show that, expect for the single case of captions in *Di* tasks, the *KI* tasks were overall performed better by the participants when the textual components have been integrated into the *Podify*'s UI. However, their responses from the *entry questionnaire* reveal that they find performing *KI* tasks on existing podcast streaming services more challenging than the *TO* ones ($3.1 \pm 0.9$ and $4 \pm 0.7$ for *KI* and *TO*, respectively. The higher the value, the easier the experience [Very Easy: 5 - Very Difficult: 1]). However, my analysis of their performance in relevance assessments suggests they overall better judge the relevance of content when with a *KI*IN. Therefore, I report a misalignment between the participants' perceived difficulty and their actual performance. I leave an exploration of this phenomenon for future work.

Overall, this analysis provides a more in-depth and nuanced understanding of the users' relevance assessment process in the context of podcast information access. The textual components in *Podify* aid the users in correctly judging the relevance of the podcast content, with their impact varying and depending on task complexity and search intent. The positive influence on the *Di* tasks, in particular, advocates for

integrating the textual components. This is despite a decrease in performance for the *Ea* tasks conducted with a *TO* search intent.

## 7.4 Chapter Summary

A pervasive challenge in podcast IR centers on finding specific information within episodes. This topic has been thoroughly covered in the 2020 and 2021 TREC Podcast Track, which was released with the Spotify Podcast Dataset [35]. This track highlighted the challenges of retrieval of fixed two-minute segments and episode summarisation [34], by outlining the importance of leveraging the transcriptions of the spoken content. Despite their utility in content-based search and user navigation, automatically generated transcriptions via ASR systems present challenges, particularly because of their length and error rates. This poses significant challenges to standard IR methods [36, 37]. However, transcriptions hold broad implications for enhancing inclusivity and content comprehension [7–12, 38–41]. Further, inferring the relationship between context, relevance, and observed implicit feedback in podcasts is challenging and requires further investigation [3]. These are research gaps that this work aims to address.

This chapter delves into the effects of incorporating the text modality, specifically captions and full-text transcripts, into the UI of *Podify* (see Chapter 5), and on the users' process of assessing the relevance of podcast content. To this end, I designed a user study (see Chapter 6), where participants were asked to perform various information-seeking tasks that included different search intents (*TO* and *KI*) as well as complexities (*Ea* and *Di*). Further, these tasks were performed using two system variations: the *BA* interface, embodying the original UI of *Podify* and existing podcast streaming services and an *EN* version that seamlessly integrates both captions and full-text transcripts. The experiment was designed to simulate real-world INs and recommendation scenarios as closely as possible.

An analysis of the questionnaires' responses (see Chapter 6, Section 6.5) shows that the experimental design was well-received by the participants (see Section 7.3.2 and Figure 7.4). They were largely engaged and invested in their participation (see Section 7.3.3), with my analysis showing the positive value of incorporating the textual modality

within the design of the *Podify* podcast streaming platform. While the *EN* system improved the participants' perception of the platform (see Figure 7.6) and tasks under the explored UE measures (see Figure 7.7), no statistically significant differences in their overall engagement were observed. This is despite a consensus of the participants preferring the *EN* system over the *BA* (addressing **RQ-7.1**; see Figure 7.5). A more in-depth analysis of the participants' responses revealed several benefits associated with the text-based components. Participants found these textual components useful for following podcast content, locating specific information, understanding content in case of audio misunderstandings, facilitating easier skimming of content, and conducting keyword searches. These findings suggest that incorporating text-based components in a podcast streaming platform has the potential to enhance user perception, preference, and task experience, without significantly impacting their overall engagement. This also suggests that the impact of the textual modality on task completion is solely related to the unique value that the textual components provide.

An investigation of how users assess the relevance of podcast content (**RQ-7.2**; Section 7.3.4 and Figure 7.8) shows that the textual components add unique value to their assessment process. My results show varying impacts, based on the independent variable factors explored in this work, that are task complexity and search intent. For the *Ea* tasks, their accuracy in providing relevance assessments increased with the *KI* intent, but showed decreased accuracy with the *TO* intent. This might be because of an IO problem that is introduced by adding this modality to this specific experimental scenario. For the *Di* tasks, including the textual components had an unequivocally positive impact, particularly when full-text transcripts were utilised. Finally, my comprehensive analysis also revealed an interesting misalignment between the participants' perceived difficulty of the tasks (see Section 7.3.1) and their actual performance. In particular, my results show that performing *KI* tasks, despite the participants' perception of finding them harder in existing streaming services, is easier than the *TO* ones.

In my analysis, I observed the following key findings:

- Participants show a preference for the *EN* system, finding the textual components

beneficial for enhancing concentration, comprehension, engagement, and efficient IR (e.g., through keyword-based search and swift skimming through the content).

- Incorporating the textual components into the *Podify*'s UI positively impacts the participants' perception of the platform, without significantly altering their overall engagement.

- The assessment of podcast content relevance is uniquely influenced by the textual components, with varied effects based on task complexity and search intent. For the *Ea* tasks, the participants' accuracy of relevance assessments increased with the *KI* intent, but showed decreased accuracy with the *TO* intent. This might be because of an IO problem that is introduced by adding this modality in this specific experimental scenario. For the *Di* tasks, including the textual components had an unequivocally positive impact, particularly when full-text transcripts were utilised.

- A misalignment is observed between the participants' perceived difficulty of tasks and their actual performance. In particular, my results show that performing *KI* tasks, despite the participants' perception of finding them harder in existing streaming services, is easier than the *TO* ones. This is with regards to how accurately they assess the relevance of content.

Overall, these findings highlight the importance and potential for integrating the textual modality into the UI of podcast streaming services. Further, it provides insights into the concept of podcast relevance, particularly on the users' process of assessing the relevance of podcast content. By analysing the users' behaviour based on context and task, the insights of this work may inform future UI design and recommendation procedures.

# Part IV

# Conclusions

# Chapter 8

# Conclusions & Future Work

After laying the foundations to understand the various concepts and methodologies utilised throughout this thesis in Chapter 2, followed by the identification of significant research gaps and subsequent investigation to related RQs (Part II - Chapter 3 and 4 and Part III - Chapter 5, 6 and 7), this thesis now reaches its conclusive chapter. The remainder of this chapter is organized as follows. Section 8.1 presents a summary of the work presented in this thesis. Subsequently, Section 8.2 outlines the main key contributions and findings. Next, the limitations of the presented research, as well as directions for future work, are presented in Section 8.3. Finally, Section 8.4 concludes this thesis.

## 8.1 Thesis Summary

The ability to understand, model, and predict users' interactions on audio streaming services is of paramount importance for enhancing the underlying recommendation procedures. Despite its clear importance, this area remains largely under-researched in the existing literature [3, 6]. In response to the thesis statement posed in Chapter 1, Section 1.2 - "*can I uncover and comprehend users' behavioural patterns that may potentially improve their engagement within the music and podcast streaming content?*" - the empirical works presented in this thesis aimed to address this research gap.

Throughout this thesis, I rigorously tackled the challenge of unravelling users' be-

havioural patterns in both the music and podcast domains. By proposing novel approaches and methodologies, I expanded upon the existing literature and provided novel perspectives on users' behaviour. The results presented in this thesis showed distinct behavioural patterns, enabling us to accurately predict and model their interactions. In turn, these insights have the potential to inform the devising of novel user modelling, recommendation, and personalisation techniques. Overall, this thesis addresses the gap between theoretical knowledge and practical applications by setting a foundation for future user-centric research in the music and podcast domains.

## 8.2 Contributions & Findings

The main contributions of this thesis are as follows:

- In Chapter 2, I summarised of the current research in the music, specifically the skipping behaviour, and podcast domains, identifying research gaps and emphasising the need to understand users' behaviour in contemporary audio streaming platforms.
- In Chapter 3, I conducted an extensive investigation into users' music skipping behavior, focusing on identifying and categorising behaviours during entire music listening sessions with regards to the users' session-based skipping activity. This involved an effective data transformation and clustering-based approach. This analysis was performed on a large real-world dataset of music streaming listening sessions (MSSD).
- In Chapter 4, I demonstrated the applicability and effectiveness of DRL in predicting users' music skipping behaviour from listening sessions. A framework was devised to extend the DRL's applicability to perform this classification and offline learning. This is the first time that DRL has been explored in this task. The effectiveness of my approach was empirically shown on the MSSD. Furthermore, I performed a comprehensive post-hoc (SHAP) and ablation analysis of my approach to study the utility of users' historical data in detecting music skips.

- In Chapter 5, I introduced *Podify*, the first web-based podcast streaming platform specifically designed for academic research. This platform allows researchers to conduct large-scale user studies in the podcast domain to alleviate the lack of user interactions in the Spotify Podcast Dataset. Through logging and exporting of all user interactions for subsequent analysis, *Podify* reduces the overhead researchers face when conducting user studies in the podcast domain.

- In Chapter 6, I detailed a user study that simulates realistic IN and recommendation scenarios by focusing on three independent variables: search intent, system, and task complexity. This study involved a specially customised version of the *Podify* platform and a curated selection of topics and podcast segments. Comprehensive questionnaires were utilised to gauge the users' experience and engagement, with a systematic method to collect and categorise the participants' relevance judgements.

- In Chapter 7, the effects of incorporating text-based components, such as captions and full-text transcripts, into the *Podify*'s UI were explored. The effectiveness and positive influence of these components on the users' process of determining the relevance of podcast content was shown empirically. An in-depth analysis was also conducted on the users' behaviour dependent on the task's complexity (i.e., *Ea* and *Di*) and their search intent (i.e., *TO* and *KI*).

In the rest of this section, I elaborate on each of the findings in details.

### 8.2.1 On Skipping Behaviour Types in Music Streaming Sessions

Chapter 3 investigated users' skipping behaviour during entire listening sessions. In contrast with prior works which focused on analysing the skip patterns as a function of the time at which this occurs within a song [15, 16], this work aimed to obtain a broader insight into how users skip music.

To achieve this, I formulated the following RQs (see Section 3.1.2):

- **RQ-3.1:** What are the main types of skipping behaviour that we can identify at a session level?

- **RQ-3.2:** For those types, how does weekday/weekend affect their overall distribution?
- **RQ-3.3:** Furthermore, do different times of the day, i.e. morning, afternoon, evening, and night, affect users' skipping interaction with the streaming service?
- **RQ-3.4:** In what ways do playlist types (e.g., personalised playlist, radio, etc.) affect users' skipping behaviour?
- **RQ-3.5:** Finally, how is the users' skipping behaviour influenced by account type (premium or free subscription plan) and when listening is performed in a shuffle mode?

In order to answer these RQs, I proposed an effective data transformation and clustering-based approach to identify and categorise different types of skipping behaviour during entire listening sessions (Section 3.2). This analysis, conducted on the real-world MSSD, was performed by considering various listening contextual factors. This included the day type, time of the day, playlist, account type (premium or free subscription plan), and shuffle listening mode (see Section 3.3).

The first outcome of my analyses, reported in Section 3.4.1 and 3.4.2, provided an empirical analysis on the clustering performance level (CVIs) and on the partitioning schemes when varying the number of clusters, and how I empirically found four to be the optimal number of clusters (Figure 3.1). By addressing **RQ-3.1**, I reported how the four identified types appear to be consistent across sessions of different length. This suggests that such dominant behaviours have no strong relation to the absolute length of a session and are thus generalisable. Additionally, a closer examination reveals that in fact I observe two main distinctive behaviours: *listener* and *listen-then-skip*. The remaining two types, i.e. *skipper* and *skip-then-listen*, can be seen as their respective complimentary behaviour (see Figure 3.2 and 3.3).

Following the previous finding, I then answered **RQ-3.[2-5]** in Section 3.4.3. I performed an analysis and discussed on the effect that various listening context information have on the distribution of the skipping types and how the users' listening activity differs when varying the length of the sessions (see Figure 3.4 and 3.5).

Overall, the main observed findings in this chapter are:

- (Section 3.4.2) The four identified types, namely *listener*, *listen-then-skip*, *skip-then-listen*, and *skipper*, are consistent across sessions of varying length. Thus, they represent a generalisable and common behaviour that has no strong relation with the absolute length of a session.
- (Section 3.4.2) A closer examination of these types reveals that these four types exhibit two pairs of complementary behaviours: *listener* and *skipper*, as well as *listen-then-skip* and *skip-then-listen*. This is because, the behaviours of, for example, a *listener* and a *skipper* can be seen as contrasting or opposing, making them complementary to each other.
- (Section 3.4.3) The distribution of skipping types varies under different listening context information. With the exception of "Day Type" (i.e., weekday and weekend), which I note as being caused by the Spotify's global user distribution, cultural differences around the definition of the "weekend", and lack of available demographic information in the data, I observe significant differences in all the other scenarios.
- (Section 3.4.3) Significant distributional differences can also be seen when varying the session lengths. I observe that, as sessions become shorter, users tend to increase their listening activity and thus skip less frequently.

### 8.2.2 On Predicting and Understanding Music Skipping using Deep Reinforcement Learning

While prior research has identified universal skipping behaviours across various contexts [15–17], other works have explored deep learning based approaches to predict the sequential skipping behaviour in music listening sessions [18, 19]. However, these models, with their static and independent processes, overlook the evolving nature of user behaviour and do not intuitively optimise for the long-term potential of UE [25–30]. To address this research gap, in Chapter 4, my focus was on understanding how people skip music from a model's perspective. I analysed the utility of the users' historical

data and explored the applicability of DRL in predicting this behaviour. Specifically, I analysed the impact and effect of the users' behaviour (e.g., the user action that leads to the current playback to start), listening content (i.e., the listened song), and contextual (e.g., the hour of the day) features in the classification task of predicting users' music skipping behaviour. My proposed approach leverages and adapts DRL for this classification task (see Section 4.3). This is to most closely reflect how a DRL-based MRS could learn to detect music skips.

To achieve this, I formulated the following RQs (see Section 4.1.1):

- **RQ-4.1:** Can DRL be applied to the users' music skipping behaviour prediction task, and if so, would it be more effective in the music skip prediction task than deep learning state-of-the-art models?
- **RQ-4.2:** What historical information is considered discriminative and serves as a high-quality indicator for the model in predicting music skipping behaviour?

To investigate my RQs, I conducted an extensive study on the MSSD. My experimental results indicated the validity of my approach by outperforming state-of-the-art deep learning based models in terms of MAA and FPA metrics (**RQ-4.1**; see Section 4.5.1 and Table 4.2). By empirically showing the effectiveness of my proposed approach, my main post-hoc (see Section 4.5.2 and 4.5.3) and ablation analysis (see Section 4.5.3) revolved around a comprehensive study of the utility and effect of users' historical data in how the proposed DRL detects music skips (addressing **RQ-4.2**). This analysis revealed a temporal data leakage problem in the historical data (Section 4.5.2 and Figure 4.1) and that the most discriminative features in predicting music skips are some users' behaviour features (Section 4.5.3). The content and contextual features were reported having a lesser effect (Table 4.3 and Figure 4.2). This points towards the potential for more focused and responsible data collection procedures in constructing user-centred MRSs.

Overall, the main observed findings in this chapter are:

- (Section 4.5.3) The most discriminative indicator for an accurate detection of skips is how users interact with the platform (i.e., *RS* and *PA*).

- (Section 4.5.3) Surprisingly, the listening CX and CN features explored in this work do not appear to have an effect on the DRL model for the prediction of music skips.

- (Section 4.5.2) My analysis also reveals a temporal data leakage problem derived from some features in the MSSD and used in the public *Spotify Sequential Skip Prediction Challenge*, since they provide information from the future that should not be made available to a live predictive system.

### 8.2.3 Influence of Text for Assessing Content Relevance in Podcast Information Access

A critical challenge in podcast IR is finding specific information within episodes. This issue was the focus of the 2020 and 2021 TREC Podcast Track, which was released with the Spotify Podcast Dataset [35]. This track highlighted the challenges of retrieval of fixed two-minute segments and episode summarisation [34], by outlining the importance of leveraging the transcriptions of the spoken content. These transcriptions, auto-generated through ASR systems, allow content-based search and user navigation. However, they pose challenges to standard IR methods because of their length and errors [36, 37]. Despite these challenges, transcriptions have broad implications for inclusivity and comprehension [7–12, 38–41]. Hence, in Chapter 7, I investigated the impact of incorporating the text modality, specifically captions and full-text transcripts, into the UI of the *Podify* podcast streaming platform (see Chapter 5). The aim of this work was to provide insights into the influence of these textual components on the users' perception of the platform and the podcast content relevance process.

To achieve this, I formulated the following RQs (see Section 7.1.1):

- **RQ-7.1**: How does incorporating captions and full-text transcripts enhance the user experience in a podcast information access context?

- **RQ-7.2**: How do these components impact the users' ability to assess the relevance of podcast content?

To explore these RQs, I designed a user study. The participants were asked to perform various information-seeking tasks, encompassing varying search intents (*TO* and *KI*) and complexities (*Ea* and *Di*). These tasks were performed on two system variations: the *BA* interface, embodying the original UI of *Podify* and existing podcast streaming services and an *EN* version that seamlessly integrates both captions and full-text transcripts. This experimental design aimed to simulate a real-world IN and recommendation scenario as closely as possible (see Chapter 6, Section 6.2).

Questionnaire results showed that the experimental design was well-received (see Section 7.3.2 and Figure 7.4). The participants were largely engaged and invested in their participation (see Section 7.3.3). My analysis also highlighted that there is a strong participants' preference and positive perception towards including the textual modality (**RQ-7.1**; see Figure 7.5, 7.6, and 7.7). This inclusion did not impose challenges on the *Podify*'s UI. Moreover, my in-depth analysis of the users' relevance assessment process further reinforced the overall positive effects of integrating the textual components (see Section 7.3.4 and Figure 7.8). Overall, the text modality was shown to aid users is better assessing the relevance of the podcast content, in a swifter and more effective way (**RQ-7.2**).

Overall, the main observed findings in this chapter are:

- (Section 7.3.3) Participants showed a preference for the *EN* system, finding the textual components beneficial for enhancing concentration, comprehension, engagement, and efficient IR (e.g., through keyword-based search and swift skimming through the content).

- (Section 7.3.3) Incorporating the textual components into the *Podify*'s UI positively impacted the participants' perception of the platform, without significantly altering their overall engagement.

- (Section 7.3.4) The assessment of podcast content relevance was uniquely influenced by the textual components, with varied effects based on task complexity and search intent. For the *Ea* tasks, the accuracy of relevance assessments increased with the *KI* intent, but showed decreased accuracy with the *TO* intent.

This might be because of an IO problem that is introduced by adding this modality to this specific experimental scenario. For the *Di* tasks, including the textual components had an unequivocally positive impact, particularly when full-text transcripts were utilised.

- (Section 7.3.4) A misalignment was observed between the participants' perceived difficulty of tasks and their actual performance. In particular, my results showed that performing *KI* tasks, despite the participants' perception of finding them harder in existing streaming services (see Section 7.3.1 and Figure 7.3), is easier than the *TO* ones. This is with regards to how accurately they assess the relevance of content.

## 8.3 Limitations & Future Work

The research presented in this thesis significantly deepens our understanding of users' behaviour in the music and podcast domains. However, there are limitations that need to be noted.

**Chapter 3**. While the work presented in this chapter deepens our understanding of how people skip music, the main limitation of this chapter lies in its scope and the dependence on the MSSD (see Section 3.3.1). It does not extensively explore the reasons (i.e., why) of users skipping. Instead, it focuses on identifying behavioural patterns, rather than identifying the motivations behind skipping music. This limitation serves as the foundation for Chapter 4, which explores the reasons for skipping from a DRL model perspective. Another significant limitation pertains to the contextual factors. While my study considered certain contextual aspects, such as day type and time of the day, there are external factors that have not been considered. This includes user mood, weather, user activities while listening (such as exercising or commuting to work), or even global events that can influence the skipping behaviour. This potentially limits the granularity and applicability of my findings, which are, however, inherent to the constraints posed by the MSSD.

I believe this work to be a valuable step towards understanding users' skipping be-

haviour. With my results showing a clear ability to detect four dominant skipping types, this work also lays the foundation for further analysis and future work. With regards to my findings, by identifying in real-time the user's skipping type, the recommendation procedure can adapt and devise a new strategy. This result can to be particularly useful in tasks aimed at content personalisation and optimisation of users' satisfaction and engagement. With the demonstrated stability of my approach, I hope my work will inspire future work in the tasks of modelling, predicting, and, most importantly, understanding the users' music skipping behaviour.

**Chapter 4**. Similar to the limitations noted for Chapter 3, the dependence on the MSSD in this chapter constrained the broader applicability of these experimental results. The users' behaviour is influenced by external (trends) and internal (individual changes of personal interests) factors. The users' shifting interests and behaviour make it hard to learn a generalisable model to tailor the user's specific needs at any given time. While DRL is suitable for tackling these challenges, the MSSD does not capture all these multi-faceted dimensions. The potential biases, such as user demographics or geographical locations, further restrict my findings (as discussed in Section 4.4.1). A more granular dataset, featuring richer behavioural data and non-anonymised sessions, would allow for deeper insights into the interplay between the skip signal and individual user's preferences, thus allowing for the investigation of situation-aware MRSs.

With the importance of modelling and understanding users' skipping behaviour, I believe this work to be an important step towards improving user modelling techniques. An accurate representation of the skipping behaviour can provide an invaluable stream of information to the underlying recommendation process. For example, I expect my findings, e.g. the *RS* type, to be highly relevant in the downstream task of capturing, in real-time, a user's skipping type (Chapter 3). By extending my approach to predict and understand other users' behaviours, I can create a holistic representation of the listeners' preferences, interests, and needs. I also advocate for thoughtful considerations when collecting and then presenting data to a model for measuring user behaviours. With increasingly rising concerns around users' data collection and privacy, the need for minimal data collection is paramount. My proposed approach can be extended in future

works to predict *when* the song is likely to be skipped. This level of information could allow to predict moments in a song where skips are most likely to occur, which could be of great value for the underlying platform. Considering *how* user's emotions or current psychological state affect their skipping behaviour is also an interesting venue for further research. With access to richer behavioural data and non-anonymised listening sessions, another line of research can investigate the relation between skipping signal and the individual user's preferences (e.g., situation-aware MRS). Finally, although not the aim of this work, performance improvements are to be expected by further tailoring my approach to the music skip prediction task. Given the user-based exploratory nature of this work, I leave further experimentation and evaluations with emerging DRL model-free offline algorithms and architectures (e.g., extending my analysis to transformer-based DRL models [243]) for future investigation.

**Chapter 5**. *Podify* is a significant advancement in podcast research, addressing a longstanding need in the academic community. One of its main features, the search functionality, leverages Elasticsearch. This ensures a modern and robust approach. However, despite the Elasticsearch's capabilities, this search functionality might not accommodate every nuanced requirement or experimental design that researchers envision. This limitation originates from the inherent constraints of Elasticsearch and the ranking models it provides. Beyond search, while *Podify*'s feedback mechanisms are user-centric, they might not fully capture the array of emotional responses and intricate nuances of the listeners' experience. Additionally, the *Podify*'s logging tools can be further refined. Delving deeper, tracking sequences of user actions with higher granularity could lead to insights into the users' behaviour. Integrating advanced behavioral tracking tools, such as eye-tracking, could further enhance the platform's capability, offering deep insights into user behaviors and thus enriching podcast research.

There are several avenues for improving *Podify* in future work. Building on the previously discussed advanced logging tools, incorporating eye-tracking could yield deeper insights into users' interactions and their behaviours, offering invaluable data for user-centric studies. Another promising direction involves the development of adaptive UIs for *Podify*. These UIs, specifically tailored to change based on the user's activity or

goals, have the potential to revolutionise the user experience. Moreover, by integrating *Podify* with established survey platforms and analytical tools, the platform could become a central hub for podcast research. Exploring beyond the current scope of Elasticsearch, the platform could benefit from a "plug-and-play" approach to various search algorithms, further enhancing its adaptability and thus catering to various research needs. By leveraging NLP techniques, the platform can delve deeper into user feedback, allowing for a nuanced understanding of the listener's sentiments. In addition, integrating state-of-the-art recommendation and personalisation systems, coupled with a show-level search and browsing experience, is another direction for future work.

**Chapter 6 & 7**. While the study detailed in these chapters provides invaluable insights into integrating textual components in podcast streaming services, there are limitations to be acknowledged. First and foremost, my findings are predominantly contextualised within the *Podify* platform. While *Podify* resemblances existing streaming platforms, its unique design elements and functionalities may influence user behaviour, potentially limiting the generalisability or introducing biases to my results. Additionally, the design of my user study primarily simulates specific real-world IN and recommendation scenarios, suggesting that the observed effects might vary under different experimental setups. Another consideration is the potential influence of my participant pool's size and diversity on the outcomes. Despite my recruitment process targeting a diverse range of participants, there might exist unaccounted biases or patterns that have not been captured. Last, while the textual modality introduced in *Podify* showed significant potential in short-term usage, its long-term implications and adaptability over extended duration remain to be explored. As users continuously engage and interact with these features, their experiences and perceptions may evolve, leading to potentially different outcomes as they familiarise with the interface and its functions.

My results highlight the advantages of incorporating the textual modality within the UI of podcast streaming services, as reported by my findings on *Podify*. Through a nuanced analysis of users' behaviour within various contexts and experimental tasks, this study offers invaluable insights that can inform future UI designs and novel content recommendation procedures. One primary concern emerging from my findings is

the potential IO problem that might arise from incorporating the textual modality. Understanding this phenomenon provides a future outlook, where streaming services provide textual aids to users in scenarios that are fruitful to enhance their overall experience. These adaptive systems, therefore, would dynamically adjust the amount and type of textual content, leveraging the users' current intentions and past interactions, thus creating an optimal user-centric experience. Of particular interest is the observed misalignment between the users' perceived task difficulty and their actual performance. Delving into this misalignment can offer crucial insights into user cognition and the decision-making processes they employ while interacting with podcast content. This, in turn, could be pivotal in devising more tailored, user-centric systems. Such systems, by leveraging the unique characteristics of podcasts, might bolster users' satisfaction and engagement. Additionally, considering *Podify*'s high resemblance to existing streaming services, future research could examine the generalibility of my findings across diverse platforms. Such cross-platform explorations could help establish more universal design and interaction guidelines.

## 8.4 Chapter Summary

In this concluding chapter, I revisited the primary objectives outlined at the beginning of this thesis, comparing them with my research findings. The broader implications and relevance of my results in the evolving landscape of audio streaming services were comprehensively discussed. Alongside my findings, I also acknowledged the inherent limitations of my research, emphasising areas of future work.

My empirical results underscore the paramount importance of comprehensively understanding, modelling, and predicting users' interactions and behaviors on audio streaming platforms. The insights of this work provide perspectives on how this can be achieved. Thus, in the future, I expect subsequent research to further refine my proposed approaches, ensuring that audio streaming services are effectively aligned with the users' multifaceted needs and preferences.

# Bibliography


[1] B. Fields *et al.*, "Contextualize your listening: the playlist as recommendation engine," Ph.D. dissertation, Goldsmiths College (University of London), 2011.

[2] F. Meggetto and Y. Moshfeghi, "Podify: A podcast streaming platform with automatic logging of user behaviour for academic research," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '23, Association for Computing Machinery, New York, NY, USA. New York, NY, USA: Association for Computing Machinery, 2023, p. 3215–3219. [Online]. Available: https://doi.org/10.1145/3539618.3591824

[3] R. Jones, H. Zamani, M. Schedl, C.-W. Chen, S. Reddy, A. Clifton, J. Karlgren, H. Hashemi, A. Pappu, Z. Nazari *et al.*, "Current challenges and future directions in podcast information access," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 2021, pp. 1554–1565.

[4] M. Tian, C. Hauff, and P. Chandar, "On the challenges of podcast search at spotify," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, M. A. Hasan and L. Xiong, Eds. ACM, 2022, pp. 5098–5099. [Online]. Available: https://doi.org/10.1145/3511808.3557518

[5] A. Li, A. Wang, Z. Nazari, P. Chandar, and B. Carterette, "Do podcasts and music compete with one another? understanding users' audio streaming habits," in *Proceedings of the web conference 2020*, 2020, pp. 1920–1931.

[6] B. Brost, R. Mehrotra, and T. Jehan, "The music streaming sessions dataset," in *Proceedings of the 30th World Wide Web Conference*, 2019, pp. 2594–2600.

[7] G. Chowdhury, *Introduction to Modern Information Retrieval*, 3rd ed. Facet Publishing, 2010.

[8] P. Pirolli and S. Card, "Information foraging," *Psychological review*, vol. 106, no. 4, p. 643, 1999.

[9] S. Burgstahler, *Equal access: Universal design of instruction*. DO-IT, University of Washington, 2008.

[10] M. F. Story, "Maximizing usability: the principles of universal design," *Assistive technology*, vol. 10, no. 1, pp. 4–12, 1998.

[11] R. E. Mayer, "The promise of multimedia learning: using the same instructional design methods across different media," *Learning and instruction*, vol. 13, no. 2, pp. 125–139, 2003.

[12] ——, "Multimedia learning," in *Psychology of learning and motivation*. Elsevier, 2002, vol. 41, pp. 85–139.

[13] P. Lamere. (2014, may) The skip. [Online]. Available: https://musicmachinery.com/2014/05/02/the-skip/

[14] F. Meggetto, C. Revie, J. Levine, and Y. Moshfeghi, "On skipping behaviour types in music streaming sessions," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, ser. CIKM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 3333–3337. [Online]. Available: https://doi.org/10.1145/3459637.3482123

[15] N. Montecchio, P. Roy, and F. Pachet, "The skipping behavior of users of music streaming services and its relation to musical structure," *Plos one*, vol. 15, no. 9, p. e0239418, 2020.

[16] J. Donier, "The universality of skipping behaviours on music streaming platforms," *arXiv preprint arXiv:2005.06987*, 2020.

[17] A. Ng and R. Mehrotra, "Investigating the impact of audio states & transitions for track sequencing in music streaming sessions," in *Proceedings of the 14th ACM Conference on Recommender Systems*, ser. RecSys '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 697–702. [Online]. Available: https://doi.org/10.1145/3383313.3418493

[18] L. Zhu and Y. Chen, "Session-based sequential skip prediction via recurrent neural networks," in *Proceedings of the 2019 WSDM Cup Workshop, February 15th 2019, Melbourne, Australia.* ACM Press, 2019.

[19] C. Hansen, C. Hansen, S. Alstrup, J. Simonsen, and C. Lioma, "Modelling sequential music track skips using a multi-rnn approach," in *WSDM '19: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining.* Association for Computing Machinery, 2019, null ; Conference date: 11-02-2019 Through 15-02-2019.

[20] S. Chang, S. Lee, and K. Lee, "Sequential skip prediction with few-shot in streamed music contents," in *Proceedings of the 2019 WSDM Cup Workshop, February 15th 2019, Melbourne, Australia.* ACM Press, 2019.

[21] O. Jeunen and B. Goethals, "Predicting sequential user behaviour with session-based recurrent neural networks: our approach to the 2019 wsdm cup sequential skip prediction challenge," in *Proceedings of the 2019 WSDM Cup Workshop, February 15th 2019, Melbourne, Australia.* ACM Press, 2019, pp. 1–4.

[22] S. Adapa, "Sequential modeling of sessions using recurrent neural networks for skip prediction," *arXiv preprint arXiv:1904.10273*, 2019.

[23] C. Tremlett, "Preliminary investigation of spotify sequential skip prediction challenge," in *Proceedings of the 2019 WSDM Cup Workshop, February 15th 2019, Melbourne, Australia.* ACM Press, 2019.

[24] F. Béres, D. M. Kelen, A. Benczúr *et al.*, "Sequential skip prediction using deep learning and ensembles," in *Proceedings of the 2019 WSDM Cup Workshop, February 15th 2019, Melbourne, Australia.* ACM Press, 2019.

[25] G. Zheng, F. Zhang, Z. Zheng, Y. Xiang, N. J. Yuan, X. Xie, and Z. Li, "Drn: A deep reinforcement learning framework for news recommendation," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, p. 167–176. [Online]. Available: https://doi.org/10.1145/3178876.3185994

[26] X. Zhao, L. Xia, L. Zhang, Z. Ding, D. Yin, and J. Tang, "Deep reinforcement learning for page-wise recommendations," in *Proceedings of the 12th ACM Conference on Recommender Systems*, 2018, pp. 95–103.

[27] G. Shani, D. Heckerman, and R. I. Brafman, "An mdp-based recommender system," *J. Mach. Learn. Res.*, vol. 6, p. 1265–1295, dec 2005.

[28] F. Liu, R. Tang, X. Li, W. Zhang, Y. Ye, H. Chen, H. Guo, and Y. Zhang, "Deep reinforcement learning based recommendation with explicit user-item interactions modeling," *arXiv preprint arXiv:1810.12027*, 2018.

[29] D. Jannach, M. Quadrana, and P. Cremonesi, "Session-based recommender systems," in *Recommender Systems Handbook*. Springer, 2022, pp. 301–334.

[30] S. Wang, L. Cao, Y. Wang, Q. Z. Sheng, M. A. Orgun, and D. Lian, "A survey on session-based recommender systems," *ACM Computing Surveys (CSUR)*, vol. 54, no. 7, pp. 1–38, 2021.

[31] A. Beutel, P. Covington, S. Jain, C. Xu, J. Li, V. Gatto, and E. H. Chi, "Latent cross: Making use of context in recurrent recommender systems," in *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, 2018, pp. 46–54.

[32] M. Lalmas, "Aggregated search," *Advanced Topics in Information Retrieval*, pp. 109–123, 2011.

[33] R. Rezapour, S. Reddy, A. Clifton, and R. Jones, "Spotify at TREC 2020: Genre-aware abstractive podcast summarization," vol. 1266, 2020. [Online]. Available: https://trec.nist.gov/pubs/trec29/papers/Spotify.P2.pdf

[34] R. Jones, B. Carterette, A. Clifton, M. Eskevich, G. J. F. Jones, J. Karlgren, A. Pappu, S. Reddy, and Y. Yu, "Trec 2020 podcasts track overview," *CoRR*, vol. abs/2103.15953, 2021.

[35] A. Clifton, S. Reddy, Y. Yu, A. Pappu, R. Rezapour, H. R. Bonab, M. Eskevich, G. J. F. Jones, J. Karlgren, B. Carterette, and R. Jones, "100, 000 podcasts: A spoken english document corpus," in *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, D. Scott, N. Bel, and C. Zong, Eds. International Committee on Computational Linguistics, 2020, pp. 5903–5917. [Online]. Available: https://doi.org/10.18653/v1/2020.coling-main.519

[36] B. Carterette, R. Jones, G. F. Jones, M. Eskevich, S. Reddy, A. Clifton, Y. Yu, J. Karlgren, and I. Soboroff, "Podcast metadata and content: Episode relevance and attractiveness in ad hoc search," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 2021, pp. 2247–2251.

[37] M. Eskevich, W. Magdy, and G. J. Jones, "New metrics for meaningful evaluation of informally structured speech retrieval," in *Advances in Information Retrieval: 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012. Proceedings 34*. Springer, 2012, pp. 170–181.

[38] R. Vanderplank, "Déjà vu? a decade of research on language laboratories, television and video in language learning," *Language teaching*, vol. 43, no. 1, pp. 1–37, 2010.

[39] A. Paivio, "Imagery and language," in *Imagery*. Elsevier, 1971, pp. 7–32.

[40] ——, "Intelligence, dual coding theory, and the brain," *Intelligence*, vol. 47, pp. 141–158, 2014.

[41] R. E. Mayer, "Cognitive theory of multimedia learning," *The Cambridge handbook of multimedia learning*, vol. 41, pp. 31–48, 2005.

[42] K. Charmaz, *Constructing grounded theory*. sage, 2014.

[43] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[44] F. Meggetto, Y. Moshfeghi, and R. Jones, "On building a podcast collection with user interactions," 2021.

[45] F. Meggetto, C. Revie, J. Levine, and Y. Moshfeghi, "Why people skip music? on predicting music skips using deep reinforcement learning," in *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, 2023, pp. 95–106.

[46] S. Frith and L. Marshall, *Music and copyright*. Edinburgh University Press, 2004.

[47] P. Wikström, *The music industry: Music in the cloud*. John Wiley & Sons, 2020.

[48] J. Menn, *All the rave: The rise and fall of Shawn Fanning's Napster*. Crown Business, 2003.

[49] S. Jones, "Music that moves: popular music, distribution and network technologies," *Cultural studies*, vol. 16, no. 2, pp. 213–232, 2002.

[50] D. Hesmondhalgh, "Is music streaming bad for musicians? problems of evidence and argument," *New Media & Society*, vol. 23, no. 12, pp. 3593–3615, 2021.

[51] T. McCourt and P. Burkart, "When creators, corporations and consumers collide: Napster and the development of on-line music distribution," *Media, Culture & Society*, vol. 25, no. 3, pp. 333–350, 2003.

[52] A. Leyshon, "The software slump?: digital music, the democratisation of technology, and the decline of the recording studio sector within the musical economy," *Environment and planning A*, vol. 41, no. 6, pp. 1309–1331, 2009.

[53] C. Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More.* Hachette Books, 2006.

[54] J. W. Morris and D. Powers, "Control, curation and musical experience in streaming music services," *Creative Industries Journal*, vol. 8, no. 2, pp. 106–122, 2015.

[55] M. L. Barata and P. S. Coelho, "Music streaming services: understanding the drivers of customer purchase and intention to recommend," *Heliyon*, vol. 7, no. 8, 2021.

[56] P. Knees, M. Schedl, and M. Goto, "Intelligent user interfaces for music discovery," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, 2020.

[57] K. Jacobson, V. Murali, E. Newett, B. Whitman, and R. Yon, "Music personalization at spotify," in *Proceedings of the 10th ACM Conference on Recommender Systems*, 2016, pp. 373–373.

[58] Ò. Celma Herrada *et al.*, *Music recommendation and discovery in the long tail.* Universitat Pompeu Fabra, 2009.

[59] J. L. Sullivan, "Podcast movement: Aspirational labour and the formalisation of podcasting as a cultural industry," *Podcasting: New aural cultures and digital media*, pp. 35–56, 2018.

[60] P. Resnick and H. R. Varian, "Recommender systems," *Communications of the ACM*, vol. 40, no. 3, pp. 56–58, 1997.

[61] P. Knees, M. Schedl, B. Ferwerda, and A. Laplante, "User awareness in music recommender systems," in *Personalized human-computer interaction.* DeGruyter Berlin, Boston, 2019, pp. 223–252.

[62] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi, "Current challenges and visions in music recommender systems research," *International Journal of Multimedia Information Retrieval*, vol. 7, pp. 95–116, 2018.

[63] M. Schedl, P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas, "Music recommender systems," pp. 453–492, 2015. [Online]. Available: https://doi.org/10.1007/978-1-4899-7637-6_13

[64] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender systems: an introduction*. Cambridge University Press, 2010.

[65] C. C. Aggarwal *et al.*, *Recommender systems*. Springer, 2016, vol. 1.

[66] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Computing Surveys (CSUR)*, vol. 52, no. 1, pp. 1–38, 2019.

[67] M. M. Afsar, T. Crump, and B. Far, "Reinforcement learning based recommender systems: A survey," *ACM Computing Surveys (CSUR)*, 2021.

[68] X. Chen, L. Yao, J. J. McAuley, G. Zhou, and X. Wang, "Deep reinforcement learning in recommender systems: A survey and new perspectives," *Knowl. Based Syst.*, vol. 264, p. 110335, 2023. [Online]. Available: https://doi.org/10.1016/j.knosys.2023.110335

[69] D. M. Greenberg and P. J. Rentfrow, "Music and big data: a new frontier," *Current opinion in behavioral sciences*, vol. 18, pp. 50–56, 2017.

[70] D. Afchar, A. B. Melchiorre, M. Schedl, R. Hennequin, E. V. Epure, and M. Moussallam, "Explainability in music recommender systems," *AI Mag.*, vol. 43, no. 2, pp. 190–208, 2022. [Online]. Available: https://doi.org/10.1002/aaai.12056

[71] G. Jawaheer, M. Szomszor, and P. Kostkova, "Characterisation of explicit feedback in an online music recommendation service," in *Proceedings of the*

*2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26-30, 2010*, X. Amatriain, M. Torrens, P. Resnick, and M. Zanker, Eds. ACM, 2010, pp. 317–320. [Online]. Available: https://doi.org/10.1145/1864708.1864776

[72] X. Amatriain, J. M. Pujol, and N. Oliver, "I like it... I like it not: Evaluating user ratings noise in recommender systems," in *User Modeling, Adaptation, and Personalization, 17th International Conference, UMAP 2009, formerly UM and AH, Trento, Italy, June 22-26, 2009. Proceedings*, ser. Lecture Notes in Computer Science, G. Houben, G. I. McCalla, F. Pianesi, and M. Zancanaro, Eds., vol. 5535. Springer, 2009, pp. 247–258. [Online]. Available: https://doi.org/10.1007/978-3-642-02247-0_24

[73] G. Jawaheer, M. Szomszor, and P. Kostkova, "Comparison of implicit and explicit feedback from an online music recommendation service," in *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems, HetRec '10, Barcelona, Spain, September 26, 2010*, P. Brusilovsky, I. Cantador, Y. Koren, T. Kuflik, and M. Weimer, Eds. ACM, 2010, pp. 47–51. [Online]. Available: https://doi.org/10.1145/1869446.1869453

[74] O. Barral, I. Kosunen, T. Ruotsalo, M. M. Spapé, M. J. Eugster, N. Ravaja, S. Kaski, and G. Jacucci, "Extracting relevance and affect information from physiological text annotation," *User Modeling and User-Adapted Interaction*, vol. 26, pp. 493–520, 2016.

[75] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White, "Evaluating implicit measures to improve web search," *ACM Transactions on Information Systems (TOIS)*, vol. 23, no. 2, pp. 147–168, 2005.

[76] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Infor-*

*mation Retrieval*, vol. 51, no. 1. Association for Computing Machinery, New York, NY, USA, 2005, pp. 4–11.

[77] G. Jacucci, O. Barral, P. Daee, M. Wenzel, B. Serim, T. Ruotsalo, P. Pluchino, J. Freeman, L. Gamberini, S. Kaski *et al.*, "Integrating neurophysiologic relevance feedback in intent modeling for information retrieval," *Journal of the Association for Information Science and Technology*, vol. 70, no. 9, pp. 917–930, 2019.

[78] Y. Moshfeghi and J. M. Jose, "An effective implicit relevance feedback technique using affective, physiological and behavioural features," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 2013, pp. 133–142.

[79] I. Arapakis, J. M. Jose, and P. D. Gray, "Affective feedback: an investigation into the role of emotions in the information seeking process," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. Association for Computing Machinery, New York, NY, USA, 2008, pp. 395–402.

[80] G. Buscher, A. Dengel, R. Biedert, and L. van Elst, "Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond," *ACM Trans. Interact. Intell. Syst.*, vol. 1, no. 2, pp. 9:1–9:30, 2012. [Online]. Available: https://doi.org/10.1145/2070719.2070722

[81] D. Kelly and J. Teevan, "Implicit feedback for inferring user preference: a bibliography," vol. 37, no. 2, 2003, pp. 18–28. [Online]. Available: https://doi.org/10.1145/959258.959260

[82] G. Dror, N. Koenigstein, Y. Koren, and M. Weimer, "The yahoo! music dataset and kdd-cup '11," in *Proceedings of KDD Cup 2011 competition, San Diego, CA, USA, 2011*, ser. JMLR Proceedings, G. Dror, Y. Koren, and M. Weimer, Eds., vol. 18. JMLR.org, 2012, pp. 8–18. [Online]. Available: http://proceedings.mlr.press/v18/dror12a.html

[83] W. Wang, F. Feng, X. He, L. Nie, and T.-S. Chua, "Denoising implicit feedback for recommendation," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 373–381.

[84] M. Schedl, P. Knees, B. McFee, and D. Bogdanov, "Music recommendation systems: Techniques, use cases, and challenges," in *Recommender Systems Handbook*. Springer, 2022, pp. 927–971.

[85] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, K. Järvelin, M. Beaulieu, R. A. Baeza-Yates, and S. Myaeng, Eds. ACM, 2002, pp. 253–260. [Online]. Available: https://doi.org/10.1145/564376.564421

[86] Y. Song, S. Dixon, and M. Pearce, "A survey of music recommendation systems and future perspectives," in *9th international symposium on computer music modeling and retrieval*, vol. 4. Citeseer, 2012, pp. 395–410.

[87] C. Hansen, C. Hansen, L. Maystre, R. Mehrotra, B. Brost, F. Tomasi, and M. Lalmas, "Contextual and sequential user embeddings for large-scale music recommendation," in *Proceedings of the Fourteenth ACM Conference on Recommender Systems*, 2020, pp. 53–62.

[88] H. Wen, L. Yang, and D. Estrin, "Leveraging post-click feedback for content recommendations," in *Proceedings of the 13th ACM Conference on Recommender Systems*, 2019, pp. 278–286.

[89] P. J. Rentfrow and S. D. Gosling, "The do re mi's of everyday life: the structure and personality correlates of music preferences." *Journal of personality and social psychology*, vol. 84, no. 6, p. 1236, 2003.

[90] ——, "Message in a ballad: The role of music preferences in interpersonal perception," *Psychological science*, vol. 17, no. 3, pp. 236–242, 2006.

[91] A. Langmeyer, A. Guglhör-Rudan, and C. Tarnai, "What do music preferences reveal about personality? a cross-cultural replication using self-ratings and ratings of music samples." *Journal of individual differences*, vol. 33, no. 2, p. 119, 2012.

[92] S. Volokhin and E. Agichtein, "Understanding music listening intents during daily activities with implications for contextual music recommendation," in *Proceedings of the 2018 conference on human information interaction and retrieval*, 2018, pp. 313–316.

[93] P. Lamere. (2015, jun) The drop machine. [Online]. Available: https://musicmachinery.com/2015/06/16/the-drop-machine/

[94] S. Banerjee and A. Pal, "Skipping skippable ads on youtube: How, when, why and why not?" in *Proceedings of the 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. IEEE, 2021, pp. 1–5.

[95] D. Belanche, C. Flavián, and A. Pérez-Rueda, "Brand recall of skippable vs non-skippable ads in youtube: Readapting information and arousal to active audiences," *Online Information Review*, vol. 44, no. 3, pp. 545–562, 2020.

[96] ——, "User adaptation to interactive advertising formats: The effect of previous exposure, habit and time urgency on ad skipping behaviors," *Telematics and Informatics*, vol. 34, no. 7, pp. 961–972, 2017.

[97] J. R. Taylor and R. T. Dean, "Influence of a continuous affect ratings task on listening time for unfamiliar art music," *Journal of New Music Research*, vol. 50, no. 3, pp. 242–258, 2021.

[98] E. Pampalk, T. Pohle, and G. Widmer, "Dynamic playlist generation based on skipping behavior," in *Proceedings of International Society for Music Information Retrieval Conference*, vol. 5, 2005, pp. 634–637.

[99] K. Bosteels, E. Pampalk, and E. E. Kerre, "Evaluating and analysing dynamic playlist generation heuristics using radio logs and fuzzy set theory." in *Proceedings*

*of International Society for Music Information Retrieval Conference*, vol. 9, 2009, pp. 351–356.

[100] B. Yang, S. Lee, S. Park, and S.-g. Lee, "Exploiting various implicit feedback for collaborative filtering," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12 Companion. New York, NY, USA: Association for Computing Machinery, 2012, p. 639–640. [Online]. Available: https://doi.org/10.1145/2187980.2188166

[101] C. Hansen, R. Mehrotra, C. Hansen, B. Brost, L. Maystre, and M. Lalmas, "Shifting consumption towards diverse content on music streaming platforms," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 238–246.

[102] J. McInerney, B. Brost, P. Chandar, R. Mehrotra, and B. Carterette, "Counterfactual evaluation of slate recommendations with sequential reward interactions," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020, pp. 1779–1788.

[103] H. Lu, M. Zhang, and S. Ma, "Between clicks and satisfaction: Study on multi-phase user preferences and satisfaction for online news reading," in *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 2018, pp. 435–444.

[104] X. Xin, A. Karatzoglou, I. Arapakis, and J. M. Jose, "Self-supervised reinforcement learning for recommender systems," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 2020, pp. 931–940.

[105] Y. Lyu, S. Dai, P. Wu, Q. Dai, Y. Deng, W. Hu, Z. Dong, J. Xu, S. Zhu, and X.-H. Zhou, "A semi-synthetic dataset generation framework for causal inference in recommender systems," *arXiv preprint arXiv:2202.11351*, 2022.

[106] O. A. Heggli, J. Stupacher, and P. Vuust, "Diurnal fluctuations in musical preference," *Royal Society open science*, vol. 8, no. 11, p. 210885, 2021.

[107] G. Fazelnia, E. Simon, I. Anderson, B. Carterette, and M. Lalmas, "Variational user modeling with slow and fast features," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, ser. WSDM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 271–279. [Online]. Available: https://doi.org/10.1145/3488560.3498477

[108] C. Chen, P. Lamere, M. Schedl, and H. Zamani, "Recsys challenge 2018: automatic music playlist continuation," in *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, S. Pera, M. D. Ekstrand, X. Amatriain, and J. O'Donovan, Eds. ACM, 2018, pp. 527–528. [Online]. Available: https://doi.org/10.1145/3240323.3240342

[109] M. Schedl, "The lfm-1b dataset for music retrieval and recommendation," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR 2016, New York, New York, USA, June 6-9, 2016*, J. R. Kender, J. R. Smith, J. Luo, S. Boll, and W. H. Hsu, Eds. ACM, 2016, pp. 103–110. [Online]. Available: https://doi.org/10.1145/2911996.2912004

[110] A. Ferraro, D. Bogdanov, and X. Serra, "Skip prediction using boosting trees based on acoustic features of tracks in sessions," in *Proceedings of the 2019 WSDM Cup Workshop, February 15th 2019, Melbourne, Australia*. ACM Press, 2019.

[111] D. Afchar and R. Hennequin, "Making neural networks interpretable with attribution: application to implicit signals prediction," in *Proceedings of the 14th ACM Conference on Recommender Systems*, 2020, pp. 220–229.

[112] G. Dulac-Arnold, L. Denoyer, P. Preux, and P. Gallinari, "Datum-wise classification: a sequential approach to sparsity," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2011, pp. 375–390.

[113] J. Janisch, T. Pevnỳ, and V. Lisỳ, "Classification with costly features using deep reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3959–3966.

[114] J. Feng, M. Huang, L. Zhao, Y. Yang, and X. Zhu, "Reinforcement learning for relation classification from noisy data," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[115] M. A. Wiering, H. Van Hasselt, A.-D. Pietersma, and L. Schomaker, "Reinforcement learning algorithms for solving classification problems," in *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*. IEEE, 2011, pp. 91–96.

[116] J. Janisch, T. Pevnỳ, and V. Lisỳ, "Classification with costly features as a sequential decision-making problem," *Machine Learning*, vol. 109, no. 8, pp. 1587–1615, 2020.

[117] C. Martinez, G. Perrin, E. Ramasso, and M. Rombaut, "A deep reinforcement learning approach for early classification of time series," in *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2030–2034.

[118] E. Lin, Q. Chen, and X. Qi, "Deep reinforcement learning for imbalanced classification," *Applied Intelligence*, pp. 1–15, 2020.

[119] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[120] R. Berry, "Will the ipod kill the radio star? profiling podcasting as radio," *Convergence*, vol. 12, no. 2, pp. 143–162, 2006.

[121] P. Research. (2021, jun) Audio and podcasting fact sheet. [Online]. Available: https://www.pewresearch.org/journalism/fact-sheet/audio-and-podcasting/

[122] Y. Liang, A. Ponnada, P. Lamere, and N. Daskalova, "Enabling goal-focused exploration of podcasts in interactive recommender systems," in *Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI 2023, Sydney, NSW, Australia, March 27-31, 2023*. ACM, 2023, pp. 142–155. [Online]. Available: https://doi.org/10.1145/3581641.3584032

[123] G. Whitner. (2023, jul) The meteoric rise of podcasting. [Online]. Available: https://musicoomph.com/podcast-statistics/

[124] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[125] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - RIAO 2000, 6th International Conference, College de France, France, April 12-14, 2000. Proceedings*, J. Mariani and D. Harman, Eds. CID, 2000, pp. 1–20. [Online]. Available: https://dl.acm.org/doi/10.5555/2835865.2835867

[126] M. Federico and G. J. F. Jones, "The CLEF 2003 cross-language spoken document retrieval track," in *Comparative Evaluation of Multilingual Information Access Systems, 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers*, ser. Lecture Notes in Computer Science, C. Peters, J. Gonzalo, M. Braschler, and M. Kluck, Eds., vol. 3237. Springer, 2003, p. 646. [Online]. Available: https://doi.org/10.1007/978-3-540-30222-3_61

[127] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study: Final report. in proc. darpa broadcast news transcription and understanding workshop," 1998.

[128] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. P. Bigham, "Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*

*2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 6798–6802. [Online]. Available: https://doi.org/10.1109/ICASSP39728.2021.9413520

[129] Apple. (2021) A podcaster's guide to rss. [Online]. Available: https://help.apple.com/itc/podcasts_connect/#/itcb54353390

[130] M. Sharpe, "A review of metadata fields associated with podcast RSS feeds," *CoRR*, vol. abs/2009.12298, 2020. [Online]. Available: https://arxiv.org/abs/2009.12298

[131] F. B. Valero, M. Baranes, and E. V. Epure, "Topic modeling on podcast short-text metadata," in *European Conference on Information Retrieval*. Springer, 2022, pp. 472–486.

[132] M. Federico and G. J. Jones, "The clef 2003 cross-language spoken document retrieval track," in *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 2003, pp. 646–652.

[133] Y. Hasebe, "Design and implementation of an online corpus of presentation transcripts of ted talks," *Procedia-Social and Behavioral Sciences*, vol. 198, pp. 174–182, 2015.

[134] S. Reddy, M. Lazarova, Y. Yu, and R. Jones, "Modeling language usage and listener engagement in podcasts," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 632–643.

[135] S. Reddy, Y. Yu, A. Pappu, A. Sivaraman, R. Rezapour, and R. Jones, "Detecting extraneous content in podcasts," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Association for Computational Linguistics, 2021, pp. 1166–1173. [Online]. Available: https://doi.org/10.18653/v1/2021.eacl-main.99

[136] J. Karlgren, "Lexical variation in english language podcasts, editorial media, and social media," *Northern European Journal of Language Technology*, vol. 8, no. 1, 2022.

[137] J. Besser, M. Larson, and K. Hofmann, "Podcast search: User goals and retrieval technologies," *Online information review*, vol. 34, no. 3, pp. 395–419, 2010.

[138] K. Martikainen, J. Karlgren, and K. Truong, "Exploring audio-based stylistic variation in podcasts," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2343–2347. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-10871

[139] M. Barthet, S. Hargreaves, and M. Sandler, "Speech/music discrimination in audio podcast using structural segmentation and timbre recognition," in *Exploring Music Contents: 7th International Symposium, CMMR 2010, Málaga, Spain, June 21-24, 2010. Revised Papers 7*. Springer, 2011, pp. 138–162.

[140] J. Ogata and M. Goto, "Podcastle: Collaborative training of language models on the basis of wisdom of crowds," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*. ISCA, 2012, pp. 2370–2373. [Online]. Available: https://doi.org/10.21437/Interspeech.2012-621

[141] L. Yang, Y. Wang, D. Dunne, M. Sobolev, M. Naaman, and D. Estrin, "More than just words: Modeling non-textual characteristics of podcasts," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, J. S. Culpepper, A. Moffat, P. N. Bennett, and K. Lerman, Eds. ACM, 2019, pp. 276–284. [Online]. Available: https://doi.org/10.1145/3289600.3290993

[142] E. Garmash, E. Tanaka, A. Clifton, J. Correia, S. Jat, W. Zhu, R. Jones, and J. Karlgren, "Cem mil podcasts: A spoken portuguese document corpus for multi-

modal, multi-lingual and multi-dialect information access research," pp. 48–59, 2023.

[143] A. Alexander, M. Mars, J. C. Tingey, H. Yu, C. Backhouse, S. Reddy, and J. Karlgren, "Audio features, precomputed for podcast retrieval and information access experiments," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2021, pp. 3–14.

[144] J. Besser, K. Hofmann, and M. A. Larson, "An exploratory study of user goals and strategies in podcast search," in *LWA 2008 - Workshop-Woche: Lernen, Wissen & Adaptivität, Würzburg, Deutschland, 6.-8. Oktober 2008, Proceedings*, ser. Technical Report, J. Baumeister and M. Atzmüller, Eds., vol. 448. Department of Computer Science, University of Würzburg, Germany, 2008, pp. 27–34.

[145] E. M. Voorhees, D. K. Harman *et al.*, *TREC: Experiment and evaluation in information retrieval*. Citeseer, 2005, vol. 63.

[146] Y. Yu, J. Karlgren, A. Clifton, M. I. Tanveer, R. Jones, and H. R. Bonab, "Spotify at the TREC 2020 podcasts track: Segment retrieval," in *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020*, ser. NIST Special Publication, E. M. Voorhees and A. Ellis, Eds., vol. 1266. National Institute of Standards and Technology (NIST), 2020. [Online]. Available: https://trec.nist.gov/pubs/trec29/papers/Spotify.P.pdf

[147] S. Hofstätter, M. Sertkan, and A. Hanbury, "TU wien at TREC DL and podcast 2021: Simple compression for dense retrieval," in *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021*, ser. NIST Special Publication, I. Soboroff and A. Ellis, Eds., vol. 500-335. National Institute of Standards and Technology (NIST), 2021. [Online]. Available: https://trec.nist.gov/pubs/trec30/papers/TU_Vienna-DL-Pod.pdf

[148] T. Ahmed and S. Bulathwela, "Towards proactive information retrieval in noisy text with wikipedia concepts," vol. 3318, 2022. [Online]. Available: https://ceur-ws.org/Vol-3318/paper10.pdf

[149] P. Galuscáková, S. Nair, and D. W. Oard, "Combine and re-rank: The university of maryland at the TREC 2020 podcasts track," vol. 1266, 2020. [Online]. Available: https://trec.nist.gov/pubs/trec29/papers/UMD_IR.P.pdf

[150] L. Vaiani, M. La Quatra, L. Cagliero, and P. Garza, "Leveraging multimodal content for podcast summarization," in *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 2022, pp. 863–870.

[151] H. Karlbom, "Abstractive summarization of podcast transcriptions," Ph.D. dissertation, Uppsala Universitet, 2021.

[152] A. Sharma and H. Pandey, "LRG at TREC 2020: Document ranking with xlnet-based models," *CoRR*, vol. abs/2103.00380, 2021. [Online]. Available: https://arxiv.org/abs/2103.00380

[153] B. Calik, "A multimodal approach: Acoustic-linguistic modelling for neural extractive speech summarisation on podcasts," Master's thesis, Utrecht University, 2023.

[154] P. Owoicho and J. Dalton, "Glasgow representation and information learning lab (GRILL) at TREC 2020 podcasts track," in *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020*, ser. NIST Special Publication, E. M. Voorhees and A. Ellis, Eds., vol. 1266. National Institute of Standards and Technology (NIST), 2020. [Online]. Available: https://trec.nist.gov/pubs/trec29/papers/uog_msc.P.pdf

[155] P. Manakul and M. Gales, "Cued_speech at trec 2020 podcast summarisation track," *arXiv preprint arXiv:2012.02535*, 2020.

[156] C. Zheng, H. J. Wang, K. Zhang, and L. Fan, "A two-phase approach for abstractive podcast summarization," vol. 1266, 2020. [Online]. Available: https://trec.nist.gov/pubs/trec29/papers/udel_wang_zheng.P.pdf

[157] ——, "A baseline analysis for podcast abstractive summarization," *CoRR*, vol. abs/2008.10648, 2020. [Online]. Available: https://arxiv.org/abs/2008.10648

[158] H. Karlbom and A. Clifton, "Abstractive podcast summarization using bart with longformer attention," in *Proceedings of the 29th Text Retrieval Conference (TREC) notebook. NIST*, 2020.

[159] K. Song, F. Liu, C. Li, X. Wang, and D. Yu, "Automatic summarization of open-domain podcast episodes," vol. 1266, 2020. [Online]. Available: https://trec.nist.gov/pubs/trec29/papers/UCF_NLP.P.pdf

[160] R. Rezapour, S. Reddy, R. Jones, and I. Soboroff, "What makes a good podcast summary?" in *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, and G. Kazai, Eds. ACM, 2022, pp. 2039–2046. [Online]. Available: https://doi.org/10.1145/3477495.3531802

[161] M. Tsagkias, M. A. Larson, W. Weerkamp, and M. de Rijke, "Podcred: a framework for analyzing podcast preference," in *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web, WICOW 2008, Napa Valley, California, USA, October 30, 2008*, K. Tanaka, T. Matsuyama, E. Lim, and A. Jatowt, Eds. ACM, 2008, pp. 67–74. [Online]. Available: https://doi.org/10.1145/1458527.1458545

[162] M. Tsagkias, M. Larson, and M. d. Rijke, "Exploiting surface features for the prediction of podcast preference," in *European Conference on Information Retrieval*. Springer, 2009, pp. 473–484.

[163] M. Tsagkias, M. Larson, and M. De Rijke, "Predicting podcast preference: An analysis framework and its application," *Journal of the American Society for information Science and Technology*, vol. 61, no. 2, pp. 374–391, 2010.

[164] M. Chadha, A. Avila, and H. Gil de Zúñiga, "Listening in: Building a profile of podcast users and analyzing their political participation," *Journal of Information Technology & Politics*, vol. 9, no. 4, pp. 388–401, 2012.

[165] S. McClung and K. Johnson, "Examining the motives of podcast users," *Journal of Radio & Audio Media*, vol. 17, no. 1, pp. 82–95, 2010.

[166] S. Fietze, "Podcast in higher education: Students usage behaviour," *Same places, different spaces. Proceedings ascilite Auckland 2009*, pp. 314–318, 2009.

[167] P. L. Gay, R. Bemrose-Fetter, G. Bracey, and F. Cain, "Astronomy cast: Evaluation of a podcast audience's content needs and listening habits," *Communicating Astronomy with the Public*, vol. 1, no. 1, pp. 24–29, 2007.

[168] C. Evans, "The effectiveness of m-learning in the form of podcast revision lectures in higher education," *Comput. Educ.*, vol. 50, no. 2, pp. 491–498, 2008. [Online]. Available: https://doi.org/10.1016/j.compedu.2007.09.016

[169] K. F. Hew, "Use of audio podcast in k-12 and higher education: A review of research topics and methodologies," *Educational Technology Research and Development*, vol. 57, pp. 333–357, 2009.

[170] E. Research. (2019, apr) The podcast consumer. [Online]. Available: https://www.edisonresearch.com/the-podcast-consumer-2019/

[171] H. Hashemi, A. Pappu, M. Tian, P. Chandar, M. Lalmas, and B. A. Carterette, "Neural instant search for music and podcast," in *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, F. Zhu, B. C. Ooi, and C. Miao, Eds. ACM, 2021, pp. 2984–2992. [Online]. Available: https://doi.org/10.1145/3447548.3467188

[172] Z. Nazari, C. Charbuillet, J. Pages, M. Laurent, D. Charrier, B. Vecchione, and B. Carterette, "Recommending podcasts for cold-start users based on music listening and taste," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, Eds. ACM, 2020, pp. 1041–1050. [Online]. Available: https://doi.org/10.1145/3397271.3401101

[173] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou, "Recommender systems," *Physics reports*, vol. 519, no. 1, pp. 1–49, 2012.

[174] S. Wang, L. Hu, Y. Wang, L. Cao, Q. Z. Sheng, and M. Orgun, "Sequential recommender systems: Challenges, progress and prospects," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 6332–6338. [Online]. Available: https://doi.org/10.24963/ijcai.2019/883

[175] C. Chen, R. Jones, Z. Nazari, L. Yang, M. Eskevich, G. J. F. Jones, and S. Oramas, "Podrecs 2021: 2nd workshop on podcast recommendations," in *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021*, H. J. C. Pampín, M. A. Larson, M. C. Willemsen, J. A. Konstan, J. J. McAuley, J. Garcia-Gathright, B. Huurnink, and E. Oldridge, Eds. ACM, 2021, pp. 796–798. [Online]. Available: https://doi.org/10.1145/3460231.3470931

[176] Y. Deldjoo, M. Schedl, P. Cremonesi, and G. Pasi, "Recommender systems leveraging multimedia content," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 106:1–106:38, 2021. [Online]. Available: https://doi.org/10.1145/3407190

[177] G. Benton, G. Fazelnia, A. Wang, and B. Carterette, "Trajectory based podcast recommendation," *arXiv preprint arXiv:2009.03859*, 2020.

[178] M. Aziz, A. Wang, A. Pappu, H. Bouchard, Y. Zhao, B. A. Carterette, and M. Lalmas, "Leveraging semantic information to facilitate the discovery of underserved podcasts," in *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, and H. Tong, Eds. ACM, 2021, pp. 3707–3716. [Online]. Available: https://doi.org/10.1145/3459637.3481934

[179] L. Yang, M. Sobolev, Y. Wang, J. Chen, D. Dunne, C. Tsangouri, N. Dell, M. Naaman, and D. Estrin, "How intention informed recommendations modulate choices: A field study of spoken word content," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, Eds. ACM, 2019, pp. 2169–2180. [Online]. Available: https://doi.org/10.1145/3308558.3313540

[180] T. M. McDonald, L. Maystre, M. Lalmas, D. Russo, and K. Ciosek, "Impatient bandits: Optimizing recommendations for the long-term without delay," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 1687–1697. [Online]. Available: https://doi.org/10.1145/3580305.3599386

[181] Z. Xing, M. Parandehgheibi, F. Xiao, N. Kulkarni, and C. Pouliot, "Content-based recommendation for podcast audio-items using natural language processing techniques," in *2016 IEEE International Conference on Big Data (IEEE BigData 2016), Washington DC, USA, December 5-8, 2016*, J. Joshi, G. Karypis, L. Liu, X. Hu, R. Ak, Y. Xia, W. Xu, A. Sato, S. Rachuri, L. H. Ungar, P. S. Yu, R. Govindaraju, and T. Suzumura, Eds. IEEE Computer Society, 2016, pp. 2378–2383. [Online]. Available: https://doi.org/10.1109/BigData.2016.7840872

[182] L. Yang, M. Sobolev, C. Tsangouri, and D. Estrin, "Understanding user interactions with podcast recommendations delivered via voice," in *Proceedings*

*of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, S. Pera, M. D. Ekstrand, X. Amatriain, and J. O'Donovan, Eds. ACM, 2018, pp. 190–194. [Online]. Available: https://doi.org/10.1145/3240323.3240389

[183] Z. Nazari, P. Chandar, G. Fazelnia, C. M. Edwards, B. A. Carterette, and M. Lalmas, "Choice of implicit signal matters: Accounting for user aspirations in podcast recommendations," in *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, and L. Médini, Eds. ACM, 2022, pp. 2433–2441. [Online]. Available: https://doi.org/10.1145/3485447.3512115

[184] B. Huber, Y. Wang, J. Garcia-Gathright, and J. Thom, "Explaining podcast recommendations to users with content diversity labels 181-186," vol. 3124, pp. 181–186, 2022. [Online]. Available: https://ceur-ws.org/Vol-3124/paper20.pdf

[185] K. Chen, S. Liu, B. Chen, and H. Wang, "Improved spoken document summarization with coverage modeling techniques," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. IEEE, 2016, pp. 6010–6014. [Online]. Available: https://doi.org/10.1109/ICASSP.2016.7472831

[186] M. Dupagne, D. M. Millette, and K. Grinfeder, "Effectiveness of video podcast use as a revision tool," *Journalism & Mass Communication Educator*, vol. 64, no. 1, pp. 54–70, 2009.

[187] R. H. Kay, "Exploring the use of video podcasts in education: A comprehensive review of the literature," *Computers in Human Behavior*, vol. 28, no. 3, pp. 820–831, 2012.

[188] N. Özdener and Y. Güngör, "Effects of video podcast technology on peer learning and project quality," *Procedia-Social and Behavioral Sciences*, vol. 2, no. 2, pp. 2217–2221, 2010.

[189] I. Mudra and M. Kitsa, "What, how and why? tiktok as a promising channel for media promotion," *Media Literacy and Academic Research*, vol. 5, no. 2, pp. 225–237, 2022.

[190] B. Agulló and A. Matamala, "Subtitling for the deaf and hard-of-hearing in immersive environments: results from a focus group," *The Journal of Specialised Translation*, vol. 32, pp. 217–235, 2019.

[191] M. Lalmas, H. O'Brien, and E. Yom-Tov, "Measuring user engagement," *Synthesis lectures on information concepts, retrieval, and services*, vol. 6, no. 4, pp. 1–132, 2014.

[192] S. Attfield, G. Kazai, M. Lalmas, and B. Piwowarski, "Towards a science of user engagement (position paper)," in *Proceedings of the WSDM workshop on user modelling for Web applications*, vol. 1, 2011.

[193] Y. Moshfeghi, M. Matthews, R. Blanco, and J. M. Jose, "Influence of timeline and named-entity components on user engagement," in *Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings*, ser. Lecture Notes in Computer Science, P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. M. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, Eds., vol. 7814. Springer, 2013, pp. 305–317. [Online]. Available: https://doi.org/10.1007/978-3-642-36973-5_26

[194] Y. Moshfeghi, B. Piwowarski, and J. M. Jose, "Handling data sparsity in collaborative filtering using emotion and semantic based features," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 625–634. [Online]. Available: https://doi.org/10.1145/2009916.2010001

[195] H. L. O'Brien and E. G. Toms, "The development and evaluation of a survey to measure user engagement," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 1, pp. 50–69, 2010.

[196] Y. Moshfeghi and J. M. Jose, "On cognition, emotion, and interaction aspects of search tasks with different search intentions," ser. WWW '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 931–942. [Online]. Available: https://doi.org/10.1145/2488388.2488469

[197] D. Holtz, B. Carterette, P. Chandar, Z. Nazari, H. Cramer, and S. Aral, "The engagement-diversity connection: Evidence from a field experiment on spotify," in *EC '20: The 21st ACM Conference on Economics and Computation, Virtual Event, Hungary, July 13-17, 2020*, P. Biró, J. D. Hartline, M. Ostrovsky, and A. D. Procaccia, Eds. ACM, 2020, pp. 75–76. [Online]. Available: https://doi.org/10.1145/3391403.3399532

[198] S. Chan-Olmsted and R. Wang, "Understanding podcast users: Consumption motives and behaviors," *New media & society*, vol. 24, no. 3, pp. 684–704, 2022.

[199] D. García-Marín, "Mapping the factors that determine engagement in podcasting: design from the users and podcasters' experience," *Communication & society*, pp. 49–63, 2020.

[200] P. Herrera, Z. Resa, and M. Sordo, "Rocking around the clock eight days a week: an exploration of temporal patterns of music listening," in *1st Workshop On Music Recommendation And Discovery (WOMRAD), ACM RecSys, 2010, Barcelona, Spain*, 2010.

[201] M. C. Hans and M. T. Smith, "Interacting with audio streams for entertainment and communication," in *Proceedings of the Eleventh ACM International Conference on Multimedia, Berkeley, CA, USA, November 2-8, 2003*, L. A. Rowe, H. M. Vin, T. Plagemann, P. J. Shenoy, and J. R. Smith, Eds. ACM, 2003, pp. 539–545. [Online]. Available: https://doi.org/10.1145/957013.957128

[202] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[203] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[204] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *WIREs Data Mining Knowl. Discov.*, vol. 2, no. 1, pp. 86–97, 2012. [Online]. Available: https://doi.org/10.1002/widm.53

[205] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[206] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," USA, Tech. Rep., 2007.

[207] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[208] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of intelligent information systems*, vol. 17, no. 2, pp. 107–145, 2001.

[209] S. Saitta, B. Raphael, and I. F. Smith, "A comprehensive validity index for clustering," *Intelligent Data Analysis*, vol. 12, no. 6, pp. 529–548, 2008.

[210] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Proceedings of the 2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 911–916.

[211] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[212] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[213] K. Kryszczuk and P. Hurley, "Estimation of the number of clusters using multiple clustering validity indices," in *Proceedings of the International Workshop on Multiple Classifier Systems*. Springer, 2010, pp. 114–123.

[214] S. Theodoridis and K. Koutroumbas, *Pattern recognition*, 3rd ed. Elsevier, 2006.

[215] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[216] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, 2nd ed. MIT press, 2018.

[217] S. Milano, M. Taddeo, and L. Floridi, "Recommender systems and their ethical challenges," *Ai & Society*, vol. 35, no. 4, pp. 957–967, 2020.

[218] C. Watkins and K. C. U. of Cambridge), *Learning from Delayed Rewards*. Cambridge University, 1989.

[219] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[220] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proceedings of International Conference on Machine Learning*, 2016, pp. 1995–2003.

[221] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proceedings of International Conference on Machine Learning*, 2016, pp. 1928–1937.

[222] S. Kapturowski, G. Ostrovski, J. Quan, R. Munos, and W. Dabney, "Recurrent experience replay in distributed reinforcement learning," in *Proceedings of the International conference on learning representations*, 2018.

[223] S. Lange, T. Gabel, and M. Riedmiller, "Batch reinforcement learning," in *Reinforcement learning*. Springer, 2012, pp. 45–73.

[224] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.

[225] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *Proceedings of International Conference on Machine Learning*. PMLR, 2019, pp. 2052–2062.

[226] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proceedings of International Conference on Machine Learning*. PMLR, 2018, pp. 1587–1596.

[227] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2016. [Online]. Available: http://arxiv.org/abs/1509.02971

[228] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[229] G. Dulac-Arnold, N. Levine, D. J. Mankowitz, J. Li, C. Paduraru, S. Gowal, and T. Hester, "Challenges of real-world reinforcement learning: definitions, benchmarks and analysis," *Machine Learning*, pp. 1–50, 2021.

[230] S. Fujimoto, E. Conti, M. Ghavamzadeh, and J. Pineau, "Benchmarking batch deep reinforcement learning algorithms," *arXiv preprint arXiv:1910.01708*, 2019.

[231] R. Agarwal, D. Schuurmans, and M. Norouzi, "An optimistic perspective on offline reinforcement learning," in *Proceedings of International Conference on Machine Learning*. PMLR, 2020, pp. 104–114.

[232] W. Dabney, M. Rowland, M. Bellemare, and R. Munos, "Distributional reinforcement learning with quantile regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[233] A. Kumar, J. Fu, M. Soh, G. Tucker, and S. Levine, "Stabilizing off-policy q-learning via bootstrapping error reduction," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[234] M. Hausknecht and P. Stone, "Deep recurrent q-learning for partially observable mdps," in *Proceedings of the 29th Association for the Advancement of Artificial Intelligence Symposium Series*, 2015.

[235] A. Kuhnle, M. Schaarschmidt, and K. Fricke, "Tensorforce: a tensorflow library for applied reinforcement learning," Web page, 2017. [Online]. Available: https://github.com/tensorforce/tensorforce

[236] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[237] W. Mason and S. Suri, "Conducting behavioral research on amazon's mechanical turk," *Behavior research methods*, vol. 44, no. 1, pp. 1–23, 2012.

[238] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance," *Journal of the American Society for information Science and Technology*, vol. 58, no. 13, pp. 2126–2144, 2007.

[239] Z. Pinkosova, W. J. McGeown, and Y. Moshfeghi, "Revisiting neurological aspects of relevance: An EEG study," in *Machine Learning, Optimization, and Data Science - 8th International Workshop, LOD 2022, Certosa di Pontignano, Italy, September 19-22, 2022, Revised Selected Papers, Part II*, ser. Lecture Notes in Computer Science, G. Nicosia, V. Ojha, E. L. Malfa, G. L. Malfa, P. M. Pardalos, G. D. Fatta, G. Giuffrida, and R. Umeton, Eds., vol. 13811. Springer, 2022, pp. 549–563. [Online]. Available: https://doi.org/10.1007/978-3-031-25891-6_41

[240] E. Sormunen, "Liberal relevance criteria of TREC -: counting on negligible documents?" in *SIGIR 2002: Proceedings of the 25th Annual International*

*ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, K. Järvelin, M. Beaulieu, R. A. Baeza-Yates, and S. Myaeng, Eds. ACM, 2002, pp. 324–330. [Online]. Available: https://doi.org/10.1145/564376.564433

[241] B. Arora and J. N. Giri, "Dual coding theory and its application in healthcare facility," *International Journal of Health Sciences*, no. II, pp. 5021–5025.

[242] W. Li, J. Yu, Z. Zhang, and X. Liu, "Dual coding or cognitive load? exploring the effect of multimodal input on english as a foreign language learners' vocabulary learning," *Frontiers in Psychology*, vol. 13, p. 834706, 2022.

[243] M. Janner, Q. Li, and S. Levine, "Offline reinforcement learning as one big sequence modeling problem," *Advances in neural information processing systems*, vol. 34, pp. 1273–1286, 2021.

# Appendix A

# Evaluation of the Proposed DQNs

This appendix contains supplementary evaluation results for the demonstration of the applicability and effectiveness of DRL in predicting user's music skipping behaviour from listening sessions (Chapter 4).

## A.1 Comparison of DQN Architectures

Table A.1 provides a comparison of the nine state-of-the-art DQN architectures that have been explored in the corresponding chapter. This analysis is carried out with the same evaluation process reported in Section 4.5.1 and Table 4.2. DQN6 corresponds to the observations stacking architecture, where six corresponds to the number of stacked observations. For RNN-based architectures, I explore Deep Recurrent Q-Network with Gated Recurrent Units (DRQN_GRU) and Long Short Term Memory (DRQN_LSTM) layers. To reflect the architecture proposed in [234], I employ a zero start, tabula rasa, state initialisation for the training procedure.

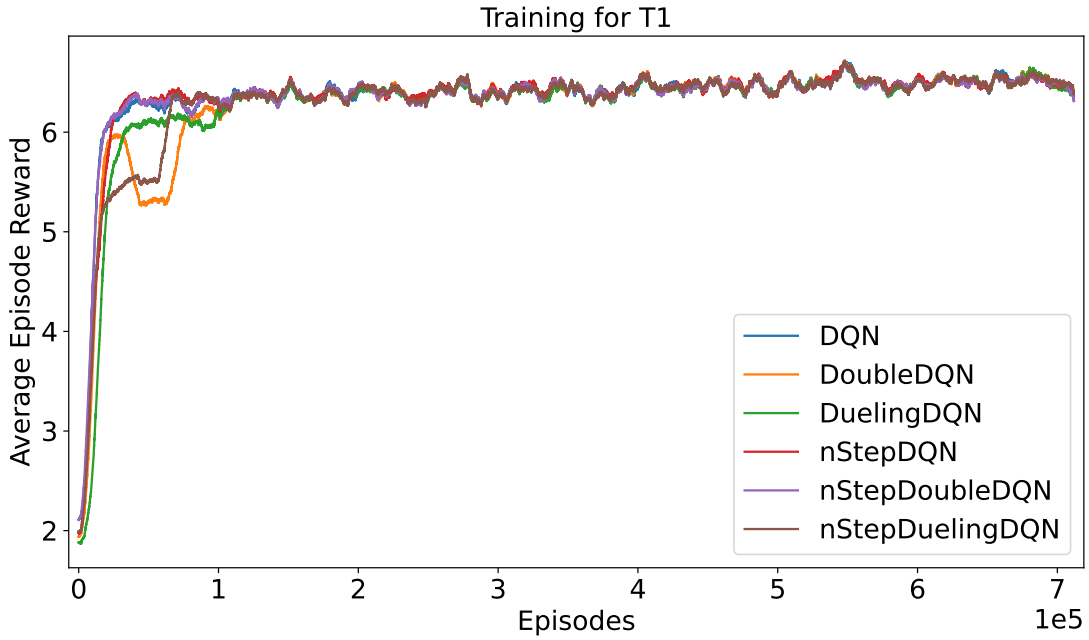
## A.2 Convergence Analysis of the DQNs

Figure A.1 provides the training curves for the fully and partially observable architectures. The sample standard deviation of the mean episode reward is not provided due

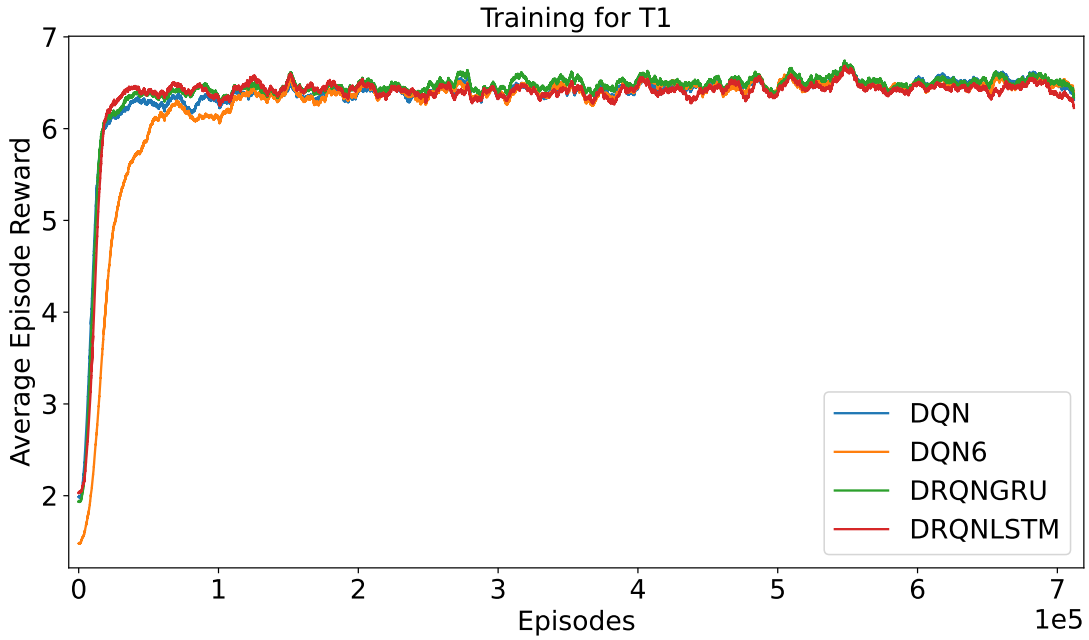
Table A.1: MAA and FPA results for the nine state-of-the-art proposed DQN architectures, categorised as fully ("*FU.Obs*") and partially observable ("*PA.Obs*"). The reported results are the averages across all test sets (with 95% CI). The best performing model is highlighted in bold.

| | | *MAA* | | *FPA* | |
|---|---|---|---|---|---|
| | | Mean | 95% CI | Mean | 95% CI |
| *FU.Obs* | DQN | **0.8201** | [0.819 - 0.821] | **0.8810** | [0.880 - 0.882] |
| | DDQN | 0.8195 | [0.818 - 0.821] | 0.8808 | [0.880 - 0.882] |
| | Dueling DQN | 0.8197 | [0.818 - 0.821] | 0.8807 | [0.880 - 0.882] |
| | n-step DQN | 0.8199 | [0.819 - 0.821] | 0.8806 | [0.880 - 0.881] |
| | n-step DDQN | 0.8198 | [0.818 - 0.821] | 0.8807 | [0.880 - 0.881] |
| | n-step Dueling DQN | 0.8193 | [0.818 - 0.821] | 0.8802 | [0.879 - 0.881] |
| *PA.Obs* | DQN6 | 0.8162 | [0.814 - 0.819] | 0.8780 | [0.876 - 0.880] |
| | DRQN_GRU | 0.8181 | [0.814 - 0.823] | 0.8792 | [0.876 - 0.882] |
| | DRQN_LSTM | 0.8178 | [0.813 - 0.822] | 0.8789 | [0.876 - 0.882] |

to the high number of architectures simultaneously explored and to ease the presentation of the results. With similar training curves, only the learning performance for architectures used for evaluation in test set T1 is reported.




(a) Fully Observable Architectures




(b) Partially Observable Architectures

Figure A.1: Learning performance for the state-of-the-art DQN architectures used for evaluation in test set T1. The x-axis reports episodes (listening sessions) in the order of $10^5$ and the y-axis is the average reward per episode. The mean of each episode for the 5 randomly-seeded runs is selected for plotting.

# Appendix B

# Participant Overview and Information Sheet, Consent Form, Task Execution Sheet, and Questionnaires

This appendix contains supplementary material for Chapter 6. It includes the overview sheet for the study (B.1), information sheets (B.2), consent form (B.3), entry questionnaires (B.4), task execution sheet (B.5), post-task questionnaires (B.6), and exit questionnaires (B.7).

## B.1 Overview Sheet for The Study

### B.1.1 Topical Session



## [Topical] Online Segment Retrieval Podcast User Study

We aim to investigate the influence of different modalities, in the context of podcast **topical** segment retrieval, for multi-modality podcast information access. In particular, the present study focuses on understanding users' behaviour during podcast listening and the level of relevance of the retrieved podcast segments. A segment is defined as a two-minute snippet of a podcast episode. The experiment will last approximately 120 minutes.

You are warmly invited to take part in the study. If you are interested in participating, please click on the "Information Sheet" button below. You will be presented with a participant information sheet which will provide all the required information about the study and it will give you time to consider your involvement and participation in the study. Clicking on "Information Sheet" does not impose any obligation. Therefore, you are not obliged to perform the study if you wish not to do so. It is also important to remember that you can withdraw from the study at any point of time. You can also have your data erased, and this will be done without giving a reason and without such decision having any adverse effects.

Thank you for your interest. Please feel free to contact us if you have any questions:

**Francesco Meggetto**
PhD Candidate
NeuraSearch Laboratory
Department of Computer & Information Sciences
University of Strathclyde
Livingston Tower
16 Richmond Street
Glasgow, G1 1XQ
francesco.meggetto@strath.ac.uk

**Dr Yashar Moshfeghi**
Senior Lecturer
NeuraSearch Laboratory
Department of Computer & Information Sciences
University of Strathclyde
Livingstone Tower
16 Richmond Street
Glasgow, G1 1XQ
yashar.moshfeghi@strath.ac.uk

Information Sheet

### B.1.2 Known-Item Session



## [Known-Item] Online Segment Retrieval Podcast User Study

We aim to investigate the influence of different modalities, in the context of podcast **known-item** segment retrieval, for multi-modality podcast information access. In particular, the present study focuses on understanding users' behaviour during podcast listening and the level of relevance of the retrieved podcast segments. A segment is defined as a two-minute snippet of a podcast episode. The experiment will last approximately 120 minutes.

You are warmly invited to take part in the study. If you are interested in participating, please click on the "Information Sheet" button below. You will be presented with a participant information sheet which will provide all the required information about the study and it will give you time to consider your involvement and participation in the study. Clicking on "Information Sheet" does not impose any obligation. Therefore, you are not obliged to perform the study if you wish not to do so. It is also important to remember that you can withdraw from the study at any point of time. You can also have your data erased, and this will be done without giving a reason and without such decision having any adverse effects.

Thank you for your interest. Please feel free to contact us if you have any questions:

**Francesco Meggetto**
PhD Candidate
NeuraSearch Laboratory
Department of Computer & Information Sciences
University of Strathclyde
Livingston Tower
16 Richmond Street
Glasgow, G1 1XQ
francesco.meggetto@strath.ac.uk

**Dr Yashar Moshfeghi**
Senior Lecturer
NeuraSearch Laboratory
Department of Computer & Information Sciences
University of Strathclyde
Livingstone Tower
16 Richmond Street
Glasgow, G1 1XQ
yashar.moshfeghi@strath.ac.uk

Information Sheet

## B.2 Information Sheet for The Study

### B.2.1 Topical Session



## Information Sheet

**Name of department**: Computer & Information Sciences, University of Strathclyde.
**Title of the study**: Online Segment Retrieval Podcast User Study

**Introduction**
Our research revolves around investigating the influence of different modalities, in the context of podcast segment retrieval, for multi-modality podcast information access. We are also interested in understanding users' behaviour during podcast listening and the level of relevance of the retrieved podcast segments. A segment is defined as a two-minute snippet of a podcast episode.

The experiment is entirely conducted through a questionnaire that will start after completion of a consent form in the next step. You are being invited to take part in the study. Before you decide to participate, it is important for you to understand why the research is being conducted and what will be involved during the procedure. Please take time to read the following information carefully and ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part. Thank you for reading this.

**What is the purpose of this research?**
The aim of the present study is to investigate the role of user behaviour in the context of assessing the relevance (from high-relevant to non-relevant) of the retrieved podcast segments to answer a **topical information need (i.e., finding general information about a topic).**

By examining various modalities and the relevance judgements, we would like to understand (i) the effect of different modalities on task completion and (ii) the user behaviour as well as textual and audio characteristics of podcasts that allow for the defining of relevant from non-relevant segments.

**Do you have to take part?**
You do not have to take part in the study, and participation is voluntary. Should you consent to participate in the study, you still have the right to withdraw at any time without providing any explanation.

**What will you do in the project?**
The experiment will take place online and entirely through a questionnaire. Whilst you perform the experimental tasks, we will continuously record your interactions within the podcast system (e.g., which segment you listen to, the listening activity such as play/pause, etc.). This will give us the necessary information to understand the role of user behaviour in the context of assessing the level of relevance for the retrieved segments.

# Appendix B. Participant Overview and Information Sheet, Consent Form, Task Execution Sheet, and Questionnaires

The experiment will last approximately 120 minutes and it is completely voluntary. Before introducing the survey and starting the experiment, a consent form will be presented. After agreeing to the stated conditions and signing it (where you will be asked to provide your randomly assigned participant's ID for validation purposes only), you will then start the experiment with an entry questionnaire. The entry questionnaire will contain demographic questions (i.e., age, education level) and the level of experience and familiarity with podcasts.

There are a total of four tasks to perform, each taking 15 minutes to complete. Each task will have a different topic for a topical type of search intent (i.e., you want to find information about a topic), and a variation of the podcast streaming platform (standard or enriched system). Prior to the start of the tasks, explicit instructions and a training video will be provided.

The specifics of each task will be provided before the start of the task. After completing each task, in the respective post-task questionnaire there are questions about the participant's views on the given task, their level of interest and engagement, and the task's perceived difficulty. After all tasks have been completed, there is a final exit questionnaire. This final stage includes questions about the tasks, the participant's overall experience with the experiment, and system and task preference.

Please note that we strongly encourage participants to assess the relevance of the segments to the best of their abilities and to provide explicit feedback (like / dislike or textual feedback with explicit rating). The participant's participation might be declined if the performance is too low, i.e. they misjudge the relevance level of too many segments. Please feel free to pose any questions during the session. It is also important to remember that a participant can withdraw from the study at any point in time. Upon participant's request, all their data can be erased, and this shall be done without giving a reason and without such a decision having any adverse effects.

**Why have you been invited to take part?**
You have been invited to take part in the experiment because you possess an advanced level of the English language (listening, reading, and writing).

**What information is being collected in the project?**
Standard demographic information, participant's views, and experience with podcasts will be collected through the questionnaire. The participant's randomly assigned ID will also be collected, and solely for validation purposes. Researchers will collect the participant's interactions with the podcast streaming platform as the experimental tasks are performed. All the data that is being collected and stored is complaint with the General Data Protection Regulation (GDPR). Once the participation is validated, the respective participant's ID will be permanently and securely deleted from the collected data. No other personal or identifiable information is collected nor will be used.

**What happens to the information collected from the project?**
All information and data collected during the experiment will be anonymised to the best of our abilities. No personal details will be collected, except for the potentially identifiable participant's randomly assigned ID. This will be securely stored in digital format, encrypted, and permanently deleted as soon as the data is collected and the corresponding participant's participation is validated. The interaction logs and survey data we collect will be retained by the organisation's researchers and may be used in future project publications, following similar ethically approved research protocols. Your participation will remain confidential. Any directly identifiable information (i.e. the participant's assigned ID) will NOT appear in any published documents relating to the research conducted.

**Who will have access to the information?**
Only the organisation's researchers will have access to the data. It is possible that the data may be used by the below-mentioned researchers for other similar ethically approved research protocols, where the same standards of confidentiality will apply. Due to the

sensitive nature of the data, the data will not be shared (unless approved by the Principal Investigator: Dr Yashar Moshfeghi).

**Where will the information be stored and how long will it be kept for?**
Only the organisation's researchers will have access to the data. It is possible that the data may be used by the below-mentioned researchers for other similar ethically approved research protocols, where the same standards of confidentiality will apply. The collected data will be stored in a secured and encrypted location within the department of Computer & Information Sciences at the University of Strathclyde. The location will be password protected, and under the management of the University of Strathclyde's network and data protection team. As per UK Research and Innovation (UKRI)'s requirements, the data will be kept for ten years, after which it will be securely disposed.

**What happens next?**
Please, take time to consider your involvement and participation in the present study. If you are happy to be involved in the study, please click on "Next Step", where you will be presented with a consent form. After agreeing to all stated conditions, the experiment will begin. Otherwise, if you do not want to take part in the study, we thank you for your time and attention.

***Thank you for reading this information. Please feel free to contact the researchers if you are unsure about the study.***
***Researchers contact details:***

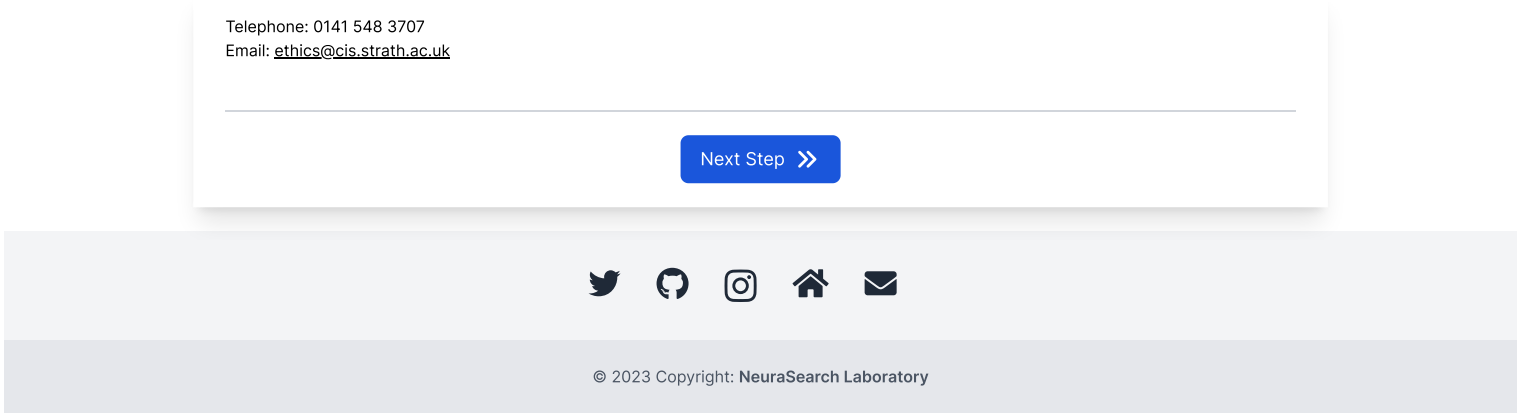
**Francesco Meggetto**
PhD Candidate
NeuraSearch Laboratory
Department of Computer & Information Sciences
University of Strathclyde
Livingston Tower
16 Richmond Street
Glasgow, G1 1XQ
francesco.meggetto@strath.ac.uk

**Dr Yashar Moshfeghi**
Senior Lecturer
NeuraSearch Laboratory
Department of Computer & Information Sciences
University of Strathclyde
Livingstone Tower
16 Richmond Street
Glasgow, G1 1XQ
yashar.moshfeghi@strath.ac.uk

**Chief Investigator details:**
This research was granted ethical approval by the Computer & Information Sciences Departmental Ethics Committee (University of Strathclyde) under application number 2077. If you have any questions or concerns, before, 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, contact details are provided below:

Departmental Ethics Committee
Department of Computer & Information Sciences
University of Strathclyde
Livingstone Tower

26 Richmond Street
Glasgow
G1 1XH
Scotland, United Kingdom


Telephone: 0141 548 3707
Email: ethics@cis.strath.ac.uk
Next Step
© 2023 Copyright: NeuraSearch Laboratory

### B.2.2 Known-Item Session



## Information Sheet

**Name of department**: Computer & Information Sciences, University of Strathclyde.
**Title of the study**: Online Segment Retrieval Podcast User Study

**Introduction**
Our research revolves around investigating the influence of different modalities, in the context of podcast segment retrieval, for multi-modality podcast information access. We are also interested in understanding users' behaviour during podcast listening and the level of relevance of the retrieved podcast segments. A segment is defined as a two-minute snippet of a podcast episode.

The experiment is entirely conducted through a questionnaire that will start after completion of a consent form in the next step. You are being invited to take part in the study. Before you decide to participate, it is important for you to understand why the research is being conducted and what will be involved during the procedure. Please take time to read the following information carefully and ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part. Thank you for reading this.

**What is the purpose of this research?**
The aim of the present study is to investigate the role of user behaviour in the context of assessing the relevance (from high-relevant to non-relevant) of the retrieved podcast segments to answer a **known-item information need (i.e., finding something that is known to exist but under an unknown name).**

By examining various modalities and the relevance judgements, we would like to understand (i) the effect of different modalities on task completion and (ii) the user behaviour as well as textual and audio characteristics of podcasts that allow for the defining of relevant from non-relevant segments.

**Do you have to take part?**
You do not have to take part in the study, and participation is voluntary. Should you consent to participate in the study, you still have the right to withdraw at any time without providing any explanation.

**What will you do in the project?**
The experiment will take place online and entirely through a questionnaire. Whilst you perform the experimental tasks, we will continuously record your interactions within the podcast system (e.g., which segment you listen to, the listening activity such as play/pause, etc.). This will

give us the necessary information to understand the role of user behaviour in the context of assessing the level of relevance for the retrieved segments.

The experiment will last approximately 120 minutes and it is completely voluntary. Before introducing the survey and starting the experiment, a consent form will be presented. After agreeing to the stated conditions and signing it (where you will be asked to provide your randomly assigned participant's ID for validation purposes only), you will then start the experiment with an entry questionnaire. The entry questionnaire will contain demographic questions (i.e., age, education level) and the level of experience and familiarity with podcasts.

There are a total of four tasks to perform, each taking 15 minutes to complete. Each task will have a different topic for a known-item type of search intent (i.e., finding something that is known to exist but under an unknown name), and a variation of the podcast streaming platform (standard or enriched system). Prior to the start of the tasks, explicit instructions and a training video will be provided.

The specifics of each task will be provided before the start of the task. After completing each task, in the respective post-task questionnaire there are questions about the participant's views on the given task, their level of interest and engagement, and the task's perceived difficulty. After all tasks have been completed, there is a final exit questionnaire. This final stage includes questions about the tasks, the participant's overall experience with the experiment, and system and task preference.

Please note that we strongly encourage participants to assess the relevance of the segments to the best of their abilities and to provide explicit feedback (like / dislike or textual feedback with explicit rating). The participant's participation might be declined if the performance is too low, i.e. they misjudge the relevance level of too many segments. Please feel free to pose any questions during the session. It is also important to remember that a participant can withdraw from the study at any point in time. Upon participant's request, all their data can be erased, and this shall be done without giving a reason and without such a decision having any adverse effects.

**Why have you been invited to take part?**
You have been invited to take part in the experiment because you possess an advanced level of the English language (listening, reading, and writing).

**What information is being collected in the project?**
Standard demographic information, participant's views, and experience with podcasts will be collected through the questionnaire. The participant's randomly assigned ID will also be collected, and solely for validation purposes. Researchers will collect the participant's interactions with the podcast streaming platform as the experimental tasks are performed. All the data that is being collected and stored is complaint with the General Data Protection Regulation (GDPR). Once the participation is validated, the respective participant's ID will be permanently and securely deleted from the collected data. No other personal or identifiable information is collected nor will be used.

**What happens to the information collected from the project?**
All information and data collected during the experiment will be anonymised to the best of our abilities. No personal details will be collected, except for the potentially identifiable participant's randomly assigned ID. This will be securely stored in digital format, encrypted, and permanently deleted as soon as the data is collected and the corresponding participant's participation is validated. The interaction logs and survey data we collect will be retained by the organisation's researchers and may be used in future project publications, following similar ethically approved research protocols. Your participation will remain confidential. Any directly identifiable information (i.e. the participant's assigned ID) will NOT appear in any published documents relating to the research conducted.

**Who will have access to the information?**

Only the organisation's researchers will have access to the data. It is possible that the data may be used by the below-mentioned researchers for other similar ethically approved research protocols, where the same standards of confidentiality will apply. Due to the sensitive nature of the data, the data will not be shared (unless approved by the Principal Investigator: Dr Yashar Moshfeghi).

**Where will the information be stored and how long will it be kept for?**
Only the organisation's researchers will have access to the data. It is possible that the data may be used by the below-mentioned researchers for other similar ethically approved research protocols, where the same standards of confidentiality will apply. The collected data will be stored in a secured and encrypted location within the department of Computer & Information Sciences at the University of Strathclyde. The location will be password protected, and under the management of the University of Strathclyde's network and data protection team. As per UK Research and Innovation (UKRI)'s requirements, the data will be kept for ten years, after which it will be securely disposed.

**What happens next?**
Please, take time to consider your involvement and participation in the present study. If you are happy to be involved in the study, please click on "Next Step", where you will be presented with a consent form. After agreeing to all stated conditions, the experiment will begin. Otherwise, if you do not want to take part in the study, we thank you for your time and attention.

***Thank you for reading this information. Please feel free to contact the researchers if you are unsure about the study.***
***Researchers contact details:***

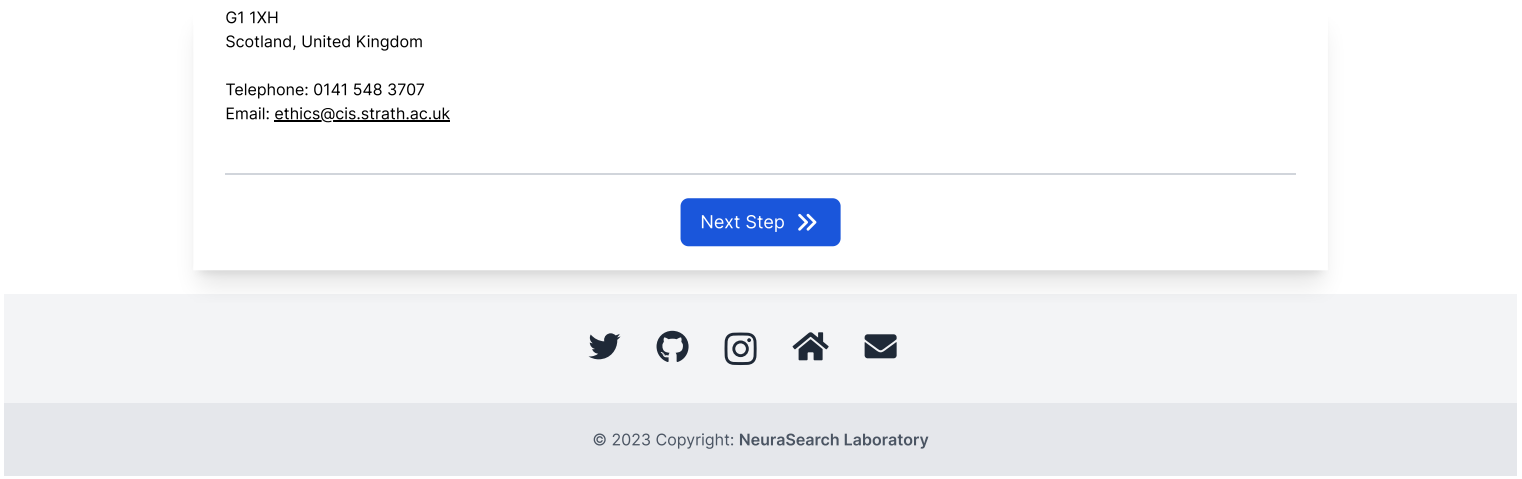
**Francesco Meggetto**
PhD Candidate
NeuraSearch Laboratory
Department of Computer & Information Sciences
University of Strathclyde
Livingston Tower
16 Richmond Street
Glasgow, G1 1XQ
francesco.meggetto@strath.ac.uk

**Dr Yashar Moshfeghi**
Senior Lecturer
NeuraSearch Laboratory
Department of Computer & Information Sciences
University of Strathclyde
Livingstone Tower
16 Richmond Street
Glasgow, G1 1XQ
yashar.moshfeghi@strath.ac.uk

**Chief Investigator details:**
This research was granted ethical approval by the Computer & Information Sciences Departmental Ethics Committee (University of Strathclyde) under application number 2077. If you have any questions or concerns, before, 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, contact details are provided below:

Departmental Ethics Committee
Department of Computer & Information Sciences
University of Strathclyde
Livingstone Tower

26 Richmond Street
Glasgow


G1 1XH
Scotland, United Kingdom
Telephone: 0141 548 3707
Email: ethics@cis.strath.ac.uk
Next Step
© 2023 Copyright: NeuraSearch Laboratory

## B.3 Consent Form



### Consent Form

**Name of department**: Computer & Information Sciences, University of Strathclyde.
**Title of the study**: Online Segment Retrieval Podcast User Study
**Ethics Approval No.**: 2077

Researcher's name: Francesco Meggetto
Researcher's email: francesco.meggetto@strath.ac.uk

If you would like a copy of the consent form to keep, please ask the researcher. If you have any complaints or concerns about this research, you can direct these to Departmental Ethics Committee, in writing by email at: ethics@cis.strath.ac.uk

Please read the following statements and indicate whether you agree by ticking the box for each statement:

Please provide your Participant ID: *

☐ 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 my personal information will be used and what will happen to it (i.e. how it will be stored and for how long). *

☐ 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 completion, without having to give a reason and without any consequences. *

I understand that I can request the withdrawal from the study of some personal information and that whenever possible researchers will comply with my request. *

I understand that anonymised data (i.e. data that do not identify me personally) cannot be withdrawn once they have been included in the study. *

I understand that any information recorded in the research will remain confidential and no information that identifies me will be made publicly available. *

I understand that my randomly assigned Participant ID will be collected. The Participant ID is solely used to verify and validate my participation in the study. It will securely be deleted immediately after my participation is validated, thus resulting in no personal identifiable information being collected, stored, and used by the researchers. *

I consent to being a participant in the project. *

I consent to be a participant in this study. *

* *Required field*

Sign Consent Form

## B.4 Entry Questionnaire

### B.4.1 Part A



## Entry Questionnaire

Researcher's name: Francesco Meggetto
Researcher's email: francesco.meggetto@strath.ac.uk

What gender do you identify as? *

Is your gender identity the same as the sex you were assigned at birth? *

- ○ Yes
- ○ No
- ○ Prefer Not To Say

What is your age? *

What is your nationality? *

How many years of formal education do you have? *

What is the highest degree or level of school you have completed? If currently enrolled, highest degree received. *

What is your current occupation? *

How would you rate your english level skills? *

* *Required field*

Next Step

### B.4.2 Part B

NeuraSearch Laboratory

University of Strathclyde Science

## Entry Questionnaire

What are your main reasons for listening to podcasts? Please select all that apply.

- [ ] For entertainment / to relax
- [ ] For emotional companionship
- [ ] To learn practical knowledge and skills
- [ ] For hobby and interested
- [ ] To find like-minded people or make new friends
- [ ] To learn new things and explore more
- [ ] To stay up-to-date with the trends
- [ ] I follow the hosts or guests

What are the activities you do while listening to podcasts? Please select all that apply.

- [ ] Not doing anything else, focused on listening
- [ ] Indoor exercising
- [ ] Driving
- [ ] Dining

- [ ] Studying or working
- [ ] Relaxing before going to sleep
- [ ] Walking or riding a bike
- [ ] Doing housework or chores (including Cooking)
- [ ] Outdoors walking, running, or walking dogs
- [ ] During leisure time
- [ ] Getting ready for breakfast or other morning routine
- [ ] Riding public transportation

---

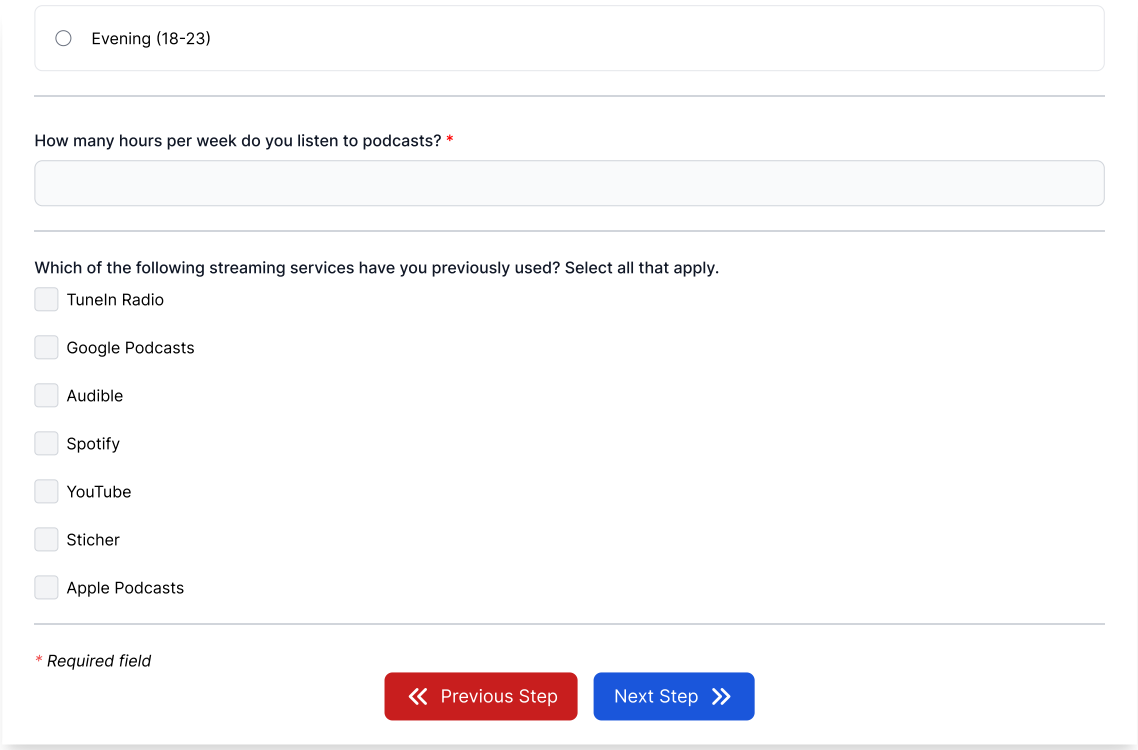
When considering a new podcast, is important that the podcast shows (select all that apply):

- [ ] Episode title
- [ ] Frequency of new episode releases
- [ ] That you have heard of the presenter
- [ ] The podcast's ratings and reviews
- [ ] That you have heard of the interview guest
- [ ] Episode description

---

During what time of the day do you usually listen to podcasts? *

- ○ Afternoon (12-17)
- ○ Morning (6-11)
- ○ Night (0-5)


Evening (18-23)
How many hours per week do you listen to podcasts? *
Which of the following streaming services have you previously used? Select all that apply.
TuneIn Radio
Google Podcasts
Audible
Spotify
YouTube
Sticher
Apple Podcasts
* Required field
Previous Step
Next Step
© 2023 Copyright: NeuraSearch Laboratory

### B.4.3 Part C



## Entry Questionnaire

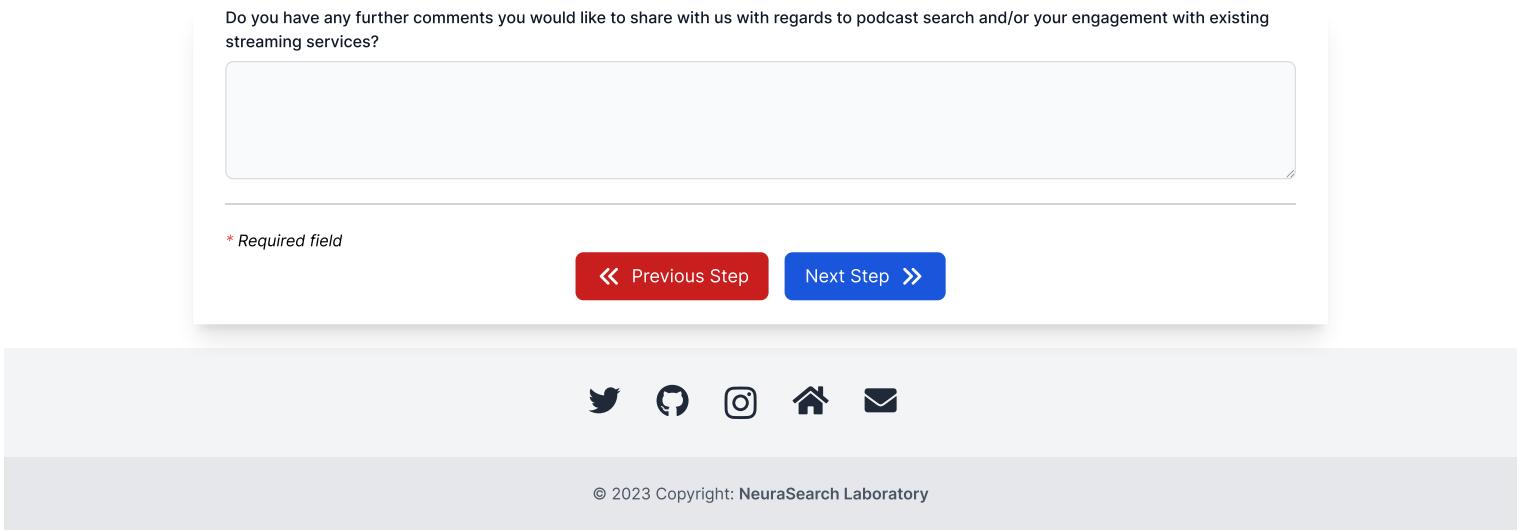
How would you rate your experience in finding general information about a topic? *

How would you rate your experience in finding something that you know exists but under an unknown name (i.e., known-item)? *

How would you rate your experience of searching for podcasts? *

How would you rate your experience in finding information from a podcast (e.g., when listening to one)? *

How would you rate your experience in navigating the catalogue of podcasts? *


Do you have any further comments you would like to share with us with regards to podcast search and/or your engagement with existing streaming services?
* Required field
Previous Step
Next Step
© 2023 Copyright: NeuraSearch Laboratory

## B.5 Task Execution



### Task Execution

Please read carefully the instructions below. It is important that you do NOT outsource any external information about the topic presented below (i.e., do not perform a web search to learn more about this topic). The topic and your task description for this task are:

**Topic**: "**causes and prevention of wildfires**"

**Task Description**: *2019 saw a large number of wildfires, in Australia, California, and the Amazon. What were people saying about them? What caused them? How could they be prevented? I am interested in news reports but also speculation, rumor, and unverified information.*

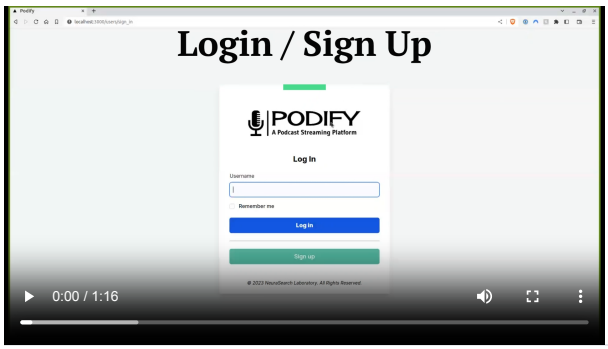
#### Instructions

Please, make sure you read and understand the topic and task presented above. You will then be asked to do the following, and in this order:

1. **Click on the button "Perform Task"**. This will open a new tab on your browser, redirecting you to the podcast streaming platform. IMPORTANT: do NOT close the survey tab!
2. **Login / Sign up** to the podcast streaming service **using the Participant ID** that you have provided in the consent form. You will be reminded of your ID below.
3. **You will be presented with an automatically generated playlist of ten podcast segments.** They are 2-minute long snippets of podcast episodes. You can learn more about these segments by clicking on them and inspecting their episode's metadata (e.g., description).
4. **Listen to as many segments as necessary** to complete the task.
    1. Listen to segments in any order you wish.
    2. Feel free to use any of the features provided by the podcast streaming platform in order to best complete the task (e.g., the listening controls).
    3. **Mandatory:** please use thumbs up/down to provide relevance and provide feedback for all segments you have listened.
5. Once you have completed your task, **close the podcast streaming system and continue with the survey.**

### Training Video

Please note that this video is only showcasing the podcast streaming platform and how to follow the instructions. It does not illustrate how to complete the task successfully!




As a reminder, this is your Participant ID: **test-subject-123**

☐ I confirm that this is my Participant ID. *

☐ I understand that I will not close this window whilst performing the task. *

Perform Task

☐ I have completed the task and I am ready to proceed to the next stage *

* *Required field*

Next Step »

Appendix B. Participant Overview and Information Sheet, Consent Form, Task Execution Sheet, and Questionnaires

## B.6 Post-Task Questionnaire



### Post Task Questionnaire

The podcast system that I used was:

| | Strongly Disagree | Somewhat Disagree | Neither Agree Nor Disagree | Somewhat Agree | Strongly Agree |
|---|---|---|---|---|---|
| Enjoyable * | ○ | ○ | ○ | ○ | ○ |
| Annoying * | ○ | ○ | ○ | ○ | ○ |
| Frustrating * | ○ | ○ | ○ | ○ | ○ |
| Confusing * | ○ | ○ | ○ | ○ | ○ |
| Visually Appealing * | ○ | ○ | ○ | ○ | ○ |
| Attractive * | ○ | ○ | ○ | ○ | ○ |

The task that I performed was:

| | Strongly Disagree | Somewhat Disagree | Neither Agree Nor Disagree | Somewhat Agree | Strongly Agree |
|---|---|---|---|---|---|
| Tiring * | ○ | ○ | ○ | ○ | ○ |

| | | | | | |
|---|---|---|---|---|---|
| Enjoyable * | | | | | |
| Involving * | | | | | |
| Interesting * | | | | | |
| Difficult * | | | | | |
| Easy To Perform * | | | | | |

With regards to completing the task, I feel that the segments in the playlist were:

| | Strongly Disagree | Somewhat Disagree | Neither Agree Nor Disagree | Somewhat Agree | Strongly Agree |
|---|---|---|---|---|---|
| Ranked Correctly (i.e., most relevant first, least relevant last) * | | | | | |
| Relevant * | | | | | |
| Helpful * | | | | | |

* *Required field*

Next Step

## B.7 Exit Questionnaire

### B.7.1 Part A

NeuraSearch Laboratory

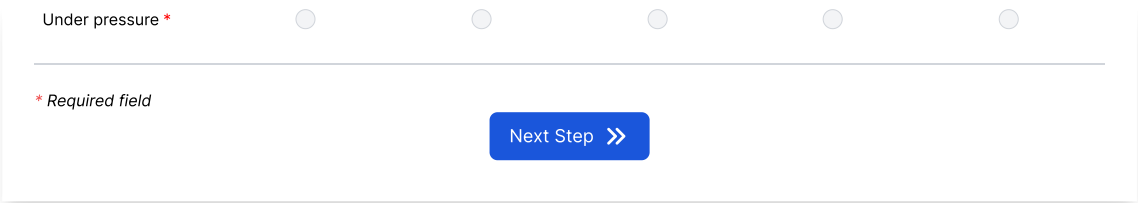
University of Strathclyde Science

## Exit Questionnaire

In this final questionnaire, we are going to ask you a few questions with regards to your performance, your view of the experiment, and which system and task you preferred and why.

I feel that, during the study, I was:

| | Strongly Disagree | Somewhat Disagree | Neither Agree Nor Disagree | Somewhat Agree | Strongly Agree |
|---|---|---|---|---|---|
| Given clear instructions * | ○ | ○ | ○ | ○ | ○ |
| Motivated * | ○ | ○ | ○ | ○ | ○ |
| Bored * | ○ | ○ | ○ | ○ | ○ |
| Comfortable * | ○ | ○ | ○ | ○ | ○ |
| Interested * | ○ | ○ | ○ | ○ | ○ |
| Tired * | ○ | ○ | ○ | ○ | ○ |
| Given enough time * | ○ | ○ | ○ | ○ | ○ |
| At ease with the procedure * | ○ | ○ | ○ | ○ | ○ |




### B.7.2 Part B



## Exit Questionnaire

1. My effort was constant throughout all four tasks: *

- ○ Yes
- ○ No

2. If you answered "no" to question 1, please identify the task which you think you have put more effort in:

3. I believe my performance was constant throughout all tasks: *

- ○ Yes
- ○ No

4. If you answered "no" to question 3, please identify the task which you think you have performed better:

5. Did you find some tasks harder than others? *

- ○ Yes
- ○ No

6. Please describe your answer for question 5. *

7. Which system did you prefer? *

- ○ Baseline
- ○ Enriched (with transcript)

8. Please describe your answer for question 7. *

9. With regards to completing the task, having the transcript was:

| | Strongly Disagree | Somewhat Disagree | Neither Agree Nor Disagree | Somewhat Agree | Strongly Agree |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| Informative * | ○ | ○ | ○ | ○ | ○ |
| Unhelpful * | ○ | ○ | ○ | ○ | ○ |
| Easier * | ○ | ○ | ○ | ○ | ○ |
| Useful * | ○ | ○ | ○ | ○ | ○ |
| Irrelevant * | ○ | ○ | ○ | ○ | ○ |
| Engaging * | ○ | ○ | ○ | ○ | ○ |
| Undesirable * | ○ | ○ | ○ | ○ | ○ |

10. Please tell us what you thought of this study and how we could improve it and/or any other general comments.

* *Required field*

« Previous Step  Next Step »