This is a peer-reviewed, author's accepted manuscript of the following conference proceedings book chapter: Meggetto, F., & Moshfeghi, Y. (Accepted/In 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. 2023. ACM.
https://doi.org/10.1145/3539618.3591824

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

Francesco Meggetto
NeuraSearch Laboratory, University of Strathclyde
Glasgow, UK
francesco.meggetto@strath.ac.uk

Yashar Moshfeghi
NeuraSearch Laboratory, University of Strathclyde
Glasgow, UK
yashar.moshfeghi@strath.ac.uk

ABSTRACT

Podcasts are spoken documents that, in recent years, have gained widespread popularity. 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 issues, in this work, we present *Podify*. It is the first web-based platform for podcast streaming and consumption specifically designed for 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. Users can manually curate playlists of podcast episodes for consumption. With mechanisms to collect explicit feedback from users (i.e., liking and disliking behaviour), *Podify* also automatically collects implicit feedback (i.e., all user interactions). Users' behaviour can be easily exported to a readable format for subsequent experimental analysis. 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.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Podcast; Platform; Listening; Search; User Behaviour; Logging

ACM Reference Format:
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), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591824

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591824

1 INTRODUCTION

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., Spotify and Amazon Music). This poses a UI challenge on what and how much information has to be presented to the user [20]. Podcasts are spoken documents (they can be represented by their transcripts of their spoken content) that have gained significant interest in recent years. They have recently started a rapidly growing process and have swiftly become, although under-researched [12], an essential part of listening habits. As of 2023, there are over 1 million active podcasts and over 30 million episodes in over 100 languages. They have been sharply rising in popularity such that in the US alone, 75% of the entire population is familiar with the term "podcasting", 55% of the US population has listened to one, and 37% are monthly listeners [22].

In 2020, Spotify identified the podcast as an important research domain and released the Spotify Podcast Dataset [6]. 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. The dataset's release was in conjunction with the TREC Podcast Track [11], which ran in 2020 and 2021, where two shared tasks were defined: retrieval of fixed two-minute segments and episodes summarisation. 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 [15]. 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.

To address these needs, and given the considerable growing importance of conducting research in podcasts, in this paper, we release a new web-based platform named *Podify*. It is the first podcast streaming service specifically designed for academic research. With high resemblances to existing modern streaming services, and a scalable design to accommodate large-scale user studies, it implements a customisable catalogue search, with manual playlist creation and curation, podcast listening, and explicit and implicit feedback 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.

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.

This platform allows researchers to conduct user studies in the podcast domain to alleviate the lack of user interactions in the Spotify Podcast Dataset [15]. This work can foster academic research toward podcast consumption patterns and listening activities and thus elicit research in novel personalisation techniques.

2 RELATED WORK

Although podcasts have been around for decades, they have only gained significant research interest recently. Podcasts have their unique characteristics [12] that differentiate them from other spoken documents (e.g., news [8], or TED talks [9, 16]). For example, their length (on average between half an hour to an hour long), conversational nature (unscripted or spontaneously organised discourse), speakers count, presentation format (e.g., interviews or monologues), and extraneous content such as ads [17] make their analysis complex [12]. This is further exacerbated by its multimodal nature, making search and recommendation unique and different from the other domains [5].

To foster research in this domain, Spotify released the English Podcast Dataset [6] in 2020. In the later years, such a dataset has been augmented with precomputed audio features [1] and the Portuguese dataset [19]. The release of the English dataset was in conjunction with the TREC Podcast Track [11], where the two shared tasks of segment retrieval and summarisation were also released. The goal of these two tasks is to ultimately find relevant information in audio or noisy transcripts and generate audio trailers for listeners to get a preview of the podcast. Numerous works were submitted to the track in 2020, and 2021 [11]. To enable content-based search and indexing by standard information retrieval methods, a full textual representation as a transcript is valuable [4, 12]. Automatic speech recognition (ASR) systems are thus used to infer a textual representation from the audio stream. However, they pose significant challenges due to the podcasts' lengthy nature [7] (and thus the need for segmentation) and the potentially significant noise introduced by the ASR system (the error reported for the Spotify Podcast Dataset is 18%) [6]. Finally, a transcript also ignores the paralinguistic characteristics of spoken language [14].

Information access tools, such as search engines and recommender systems, are essential to finding and discovering podcasts [12]. As of 2021, most podcast search is done via catalogue match, using the metadata provided by podcast creators [4]. However, such metadata is of highly varying quality and may provide significant noise in the search process [4, 21]. As of 2019, recommendations from friends and family are in the top-three ways that people find podcasts [18], with very few works published in podcast recommendation since [2, 5, 10, 13]. Most recently, Liang et al. [13] explored users' personal goals for goal-focused recommendations. With the podcast medium's success and consistent increase of shows and episodes, there is a need to develop systems that can leverage this large amount of podcast content. They should offer academic and independent researchers the capability to perform user studies and advance the field. However, there is currently no podcast streaming platform with a content-based podcast search engine [4] for academic research and no data collection with logged user behaviour [15]. To address these limitations, in this paper, we present *Podify*. It is the first podcast streaming service that is specifically designed for academic research. In particular, it provides a large-scale search within its podcast episodes catalogue, a listening experience that highly resembles existing streaming services, and with all the users' interactions (implicit and explicit) being automatically collected and easily exported for processing and experimental analysis.

3 SYSTEM ARCHITECTURE

In this section we describe the *Podify*'s user interface, functionalities, and the user behaviour that is collected.

3.1 User Interface

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 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 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 1 (C). They can be inspected by clicking on individual episodes to view their metadata (Figure 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 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 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, systems, episodes, playlists, users and their logged interactions).
(2) All the collected users' behavioural data can be inspected and easily exported in a readable 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,

[1]Demonstration available in the provided YouTube video due to space limitations.

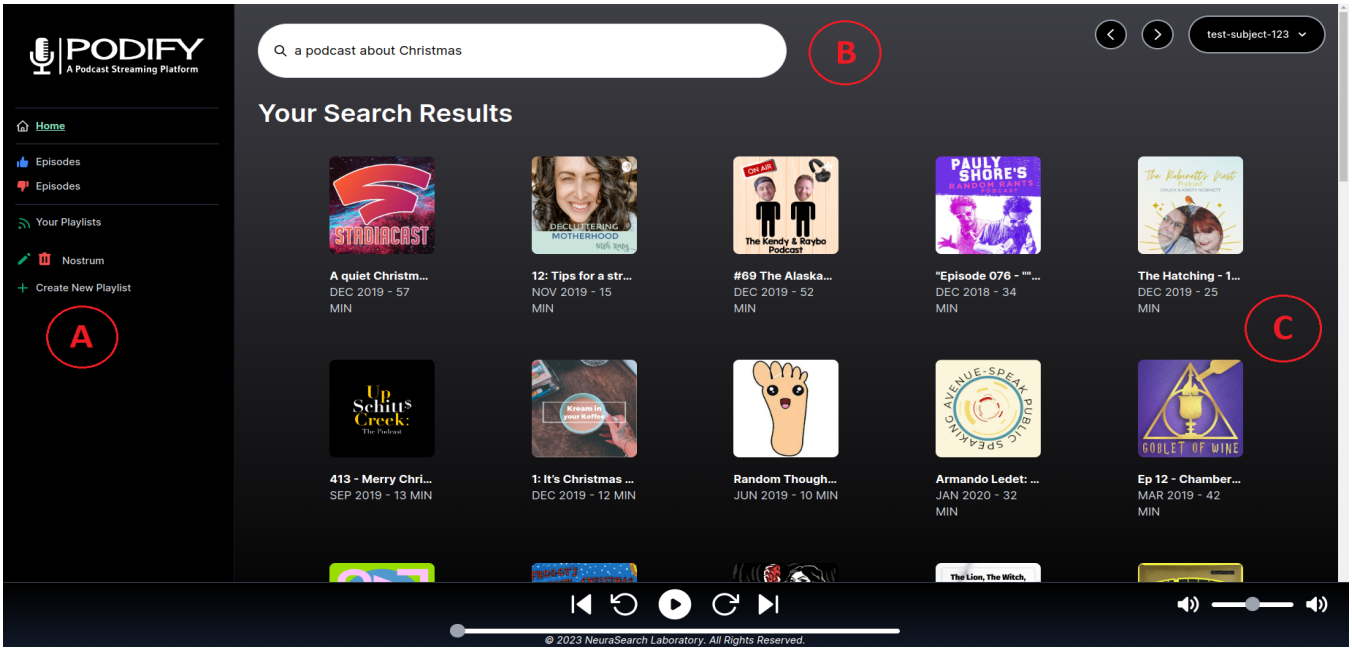

Figure 1: *Podify*'s user interface with an example of catalogue search. Top@50 results for the query "a podcast about Christmas".

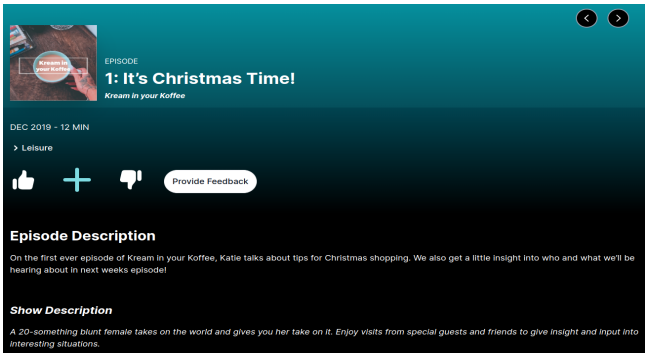

Figure 2: Episode's page with metadata (e.g., publication date), like, add to a playlist, dislike, and textual explicit feedback.

surveys. We expect future versions to extend this functionality to include more control variables (e.g., search intents or recommendation procedures).

3.2 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 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 [3, 4] 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.

3.3 Catalogue Creation Procedure

The procedure to create the *Podify*'s catalogue of podcast episodes is straightforward and based on the RSS feeds. A 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 3.2. 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 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 [6]. However, the RSS feeds originating from the Spotify Podcast Dataset were used for this demonstration. In particular, we selected the first 1,000 episodes that have a valid RSS feed.

3.4 User Behaviour

All user interactions with *Podify* are automatically logged, and they are easily exportable in a readable format (CSV). Each interaction

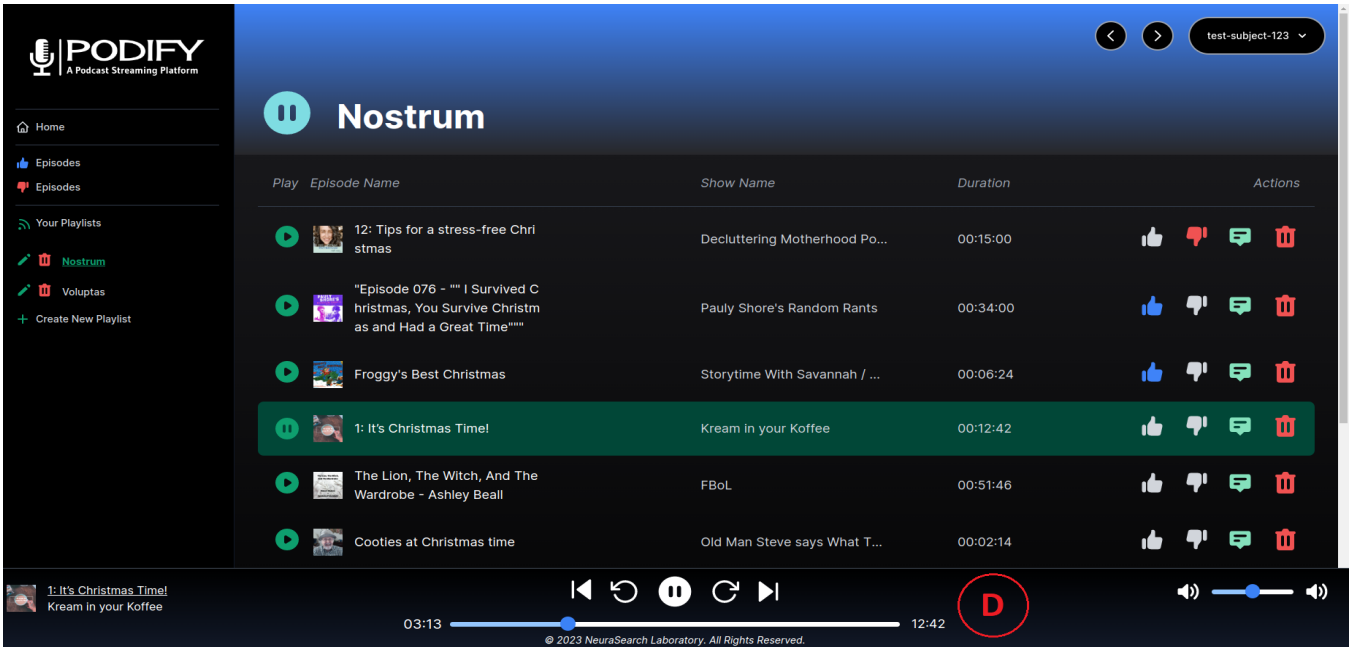

Figure 3: *Podify*'s user interface with an example of a manually curated playlist (i.e., "Nostrum") and episode consumption.

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 (GDPR compliant) about the traffic source (referrer, referring domain, landing page), country-level geocoding, and technology (browser, OS, device type). 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.

3.5 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. We 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 3.1), this cron schedule is also implemented to avoid any potential data loss. Given *Podify*'s modern and state-of-the-art architectural implementation, we expect its integration into modern pipelines for scaling and swift content delivery to conduct large-scale user studies.

4 CONCLUSIONS AND FUTURE WORK

In this work, we propose *Podify*, a web-based podcast streaming and consumption platform. With high resemblances to existing podcast streaming services, it is designed to support academic research in the podcast domain, specifically in the under-researched search and user behavioural analysis areas. *Podify* is the first available research-purpose podcast streaming service that also provides a large-scale and highly customisable search within its easy-to-create catalogue of podcast episodes. 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. Mechanisms to collect explicit feedback (i.e., liking and disliking of an episode) are also provided. We hope to integrate various recommendation and personalisation procedures and a show-level search and browsing experience in future work.

Acknowledgement: This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/R513349/1].

REFERENCES

[1] Abigail Alexander, Matthijs Mars, Josh C Tingey, Haoyue Yu, Chris Backhouse, Sravana Reddy, and Jussi Karlgren. 2021. Audio Features, Precomputed for Podcast Retrieval and Information Access Experiments. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 3–14.

[2] Greg Benton, Ghazal Fazelnia, Alice Wang, and Ben Carterette. 2020. Trajectory based podcast recommendation. *arXiv preprint arXiv:2009.03859* (2020).

[3] Jana Besser, Martha Larson, and Katja Hofmann. 2010. Podcast search: User goals and retrieval technologies. *Online information review* 34, 3 (2010), 395–419.

[4] Ben Carterette, Rosie Jones, Gareth F Jones, Maria Eskevich, Sravana Reddy, Ann Clifton, Yongze Yu, Jussi Karlgren, and Ian Soboroff. 2021. 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*. 2247–2251.

[5] Ching-Wei Chen, Rosie Jones, Zahra Nazari, Longqi Yang, Maria Eskevich, Gareth James Francis Jones, and Sergio Oramas. 2021. PodRecs 2021: 2nd Workshop on Podcast Recommendations. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 796–798.

[6] Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, et al. 2020. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*. 5903–5917.

[7] Maria Eskevich, Walid Magdy, and Gareth JF Jones. 2012. 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, 170–181.

[8] Marcello Federico and Gareth JF Jones. 2003. The CLEF 2003 cross-language spoken document retrieval track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 646–652.

[9] Yoichiro Hasebe. 2015. Design and implementation of an online corpus of presentation transcripts of TED talks. *Procedia-Social and Behavioral Sciences* 198 (2015), 174–182.

[10] Bernd Huber, Yixue Wang, Jean Garcia-Gathright, and Jenn Thom. 2022. Explaining Podcast Recommendations To Users with Content Diversity Labels. (2022).

[11] Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth JF Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. 2021. Trec 2020 podcasts track overview. *arXiv preprint arXiv:2103.15953* (2021).

[12] Rosie Jones, Hamed Zamani, Markus Schedl, Ching-Wei Chen, Sravana Reddy, Ann Clifton, Jussi Karlgren, Helia Hashemi, Aasish Pappu, Zahra Nazari, et al. 2021. 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*. 1554–1565.

[13] Yu Liang, Aditya Ponnada, Paul Lamere, and Nediyana Daskalova. 2023. Enabling Goal-Focused Exploration of Podcasts in Interactive Recommender Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 142–155.

[14] Katariina Martikainen, Jussi Karlgren, and Khiet Phuong Truong. 2022. Exploring audio-based stylistic variation in podcasts. In *INTERSPEECH 2022*.

[15] Francesco Meggetto, Yashar Moshfeghi, and Rosie Jones. 2021. On Building a Podcast Collection with User Interactions. (2021).

[16] Sravana Reddy, Mariya Lazarova, Yongze Yu, and Rosie Jones. 2021. 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)*. 632–643.

[17] Sravana Reddy, Yongze Yu, Aasish Pappu, Aswin Sivaraman, Rezvaneh Rezapour, and Rosie Jones. 2021. Detecting Extraneous Content in Podcasts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1166–1173.

[18] Edison Research. 2019. *The Podcast Consumer*. Retrieved Feb 6, 2023 from https://www.edisonresearch.com/the-podcast-consumer-2019/

[19] Edgar Tanaka, Ann Clifton, Joana Correia, Sharmistha Jat, Rosie Jones, Jussi Karlgren, and Winstead Zhu. 2022. Cem Mil Podcasts: A Spoken Portuguese Document Corpus. *arXiv preprint arXiv:2209.11871* (2022).

[20] Mi Tian, Claudia Hauff, and Praveen Chandar. 2022. On the Challenges of Podcast Search at Spotify. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 5098–5099.

[21] Francisco B Valero, Marion Baranes, and Elena V Epure. 2022. Topic Modeling on Podcast Short-Text Metadata. In *European Conference on Information Retrieval*. Springer, 472–486.

[22] Gavin Whitner. 2023. *The Meteoric Rise of Podcasting*. Retrieved Feb 6, 2023 from https://musicoomph.com/podcast-statistics/