From Clicks to Cues:

Exploring User Behaviour As a Language in Music Video Consumption

by

Vishruty Mittal

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on June 14, 2023 at 02:00 PM (CET).

Student Number:	5584825	
Project Duration:	October 2022 - June 2023	
Thesis Committee:	Dr. Ujwal Gadiraju	TU Delft, chair and supervisor
	Dr. Maria Soledad Pera	TU Delft
	Dr. Megha Khosla	TU Delft
Industry Supervisor:	Dr. Zoltán Szlávik	XITE

TU Delft XITE

Preface

This thesis report presents a culmination of my work towards obtaining the degree of Master of Science in Computer Science, in the Data Science track at the Delft University of Technology, The Netherlands. This project began with a discussion with my university supervisor Dr. Ujwal Gadiraju and my industry supervisor, Dr. Zoltán Szlávik revolving around the interpretation of user actions while consuming music videos.

This report introduces the reader to the problem domain, highlights existing research gaps, and delineates the motivation behind adopting language models for decoding user behavioural actions. It introduces an embedding and clustering methodology to discern and characterize distinct user behaviour patterns during music video consumption. Further, this research dives into examining the impact of contextual factors on user behaviour.

I am deeply thankful to Dr. Ujwal Gadiraju, for his guidance and constant encouragement throughout the process. I am equally grateful to Dr. Zoltán Szlávik whose valuable insights and critical feedback have significantly shaped this work. I further extend my gratitude to my mentors, Ioannis Petros Samiotis and Garrett Allen, for the numerous brain storming sessions, their availability, insights, and encouraging nature. I also wish to acknowledge the data science team at XITE, who, with their perpetual readiness to help and ideas, enriched this project. I am also thankful to my family and friends for their unwavering support and belief in me and my work. Lastly, I would like to acknowledge the role of ChatGPT in assisting with the writing process and helping me with paraphrasing certain sections.

I hope you enjoy reading this research as much as I enjoyed working on it.

Vishruty Mittal
Delft, June 2023

Abstract

As music video streaming occupies a significant market share in how people consume music, gaining an understanding of user behavioural patterns becomes increasingly crucial. This understanding can enable better music video streaming experiences by tailoring them towards more personalized and user-centric designs. Though prior works have emphasized user behaviour during solely listening to music, understanding user actions/clicks while consuming music videos remains largely unexplored. Given the unique experience offered by the combination of audio and visual elements, there is a need for focused research in this area.

Therefore this study attempts to bridge this research gap by collecting and analysing a large dataset of streaming sessions from a music video streaming company - XITE. In total, we analyzed 1.8 million sessions from approximately 270,000 unique users. The behaviour exhibited during those sessions is interpreted as a language and modelled using the Language Model - Doc2Vec. This facilitated the conversion of session action sequences into embeddings. Our findings suggest that music video streaming sessions exhibit cohesive user interaction patterns, which can be grouped into distinct clusters, thereby enabling the detection of distinct behavioural patterns across user sessions.

Furthermore, previous studies have indicated that user interactions with multimedia streaming platforms can be influenced by the context in which content is consumed. Extending these findings, our analysis of behavioural clusters revealed that certain user behaviours while consuming music videos are associated with specific music video genres and temporal factors. For instance, we discovered that passive sessions predominantly commence around 10 am, while sessions requiring more active engagement typically start in the evening. The insights derived from this study are valuable for improving user-centric design in music video streaming platforms and providing businesses with data-driven recommendations for strategic planning.

A research paper based on parts of this thesis (interpreting user actions as a language by employing Doc2vec, clustering and cluster interpretation via correlation analysis) has been submitted to the 2023 edition of the International Society for Music Information Retrieval (ISMIR) conference. The paper can be found in Appendix A.

Contents

List of Figures

List of Tables

1 Introduction

> Music is a moral law. It gives a soul to the Universe, wings to the mind, flight to the imagination, a charm to sadness, gaiety and life to everything. It is the essence of order, and leads to all that is good and just and beautiful. —Plato

Music, a universal form of expression, has been savoured by humans across the centuries. It has resonated with us from the time of Aristotle, reverberated through the writings of Marsilio Ficino and echoed in the poems of Rabindranath Tagore. In our contemporary world, music is omnipresent. Even at this present moment, people across the world are listening to music in various settings such as offices, restaurants, shopping complexes, cars and bars. Due to its pervasive presence in daily life, the role and significance of music as well as the reasons why humans listen to it, has been long speculated among philosophers and has inspired multiple studies [10, 30]. Moreover, understanding user behaviour and distinguishing patterns in music consumption has become a critical area of research [67], as unravelling these insights holds the potential to comprehend user preferences and requirements, ultimately facilitating the development of more personalized recommender systems and streaming platforms.

1.1. Music Videos: A Visual Form of Musical Expression

1.1.1. Popularity of Music Videos

Music videos, as a medium of musical expression, have been widely popular for decades. The emergence of Music Television (MTV) in the United States in 1981 heralded a significant change in the music industry. By broadcasting music videos around the clock, MTV transformed the way in which music was consumed and experienced by audiences. As an assortment of American music videos gradually became available internationally in the early 1980s, this shift in the medium of music distribution quickly radiated to other parts of the world. Consequently, this worldwide exposure led to a ripple effect, inspiring other nations to produce their own music videos showcasing local artists. The international popularity of music videos further catalyzed the establishment of MTV Europe in 1987 [23].

Additionally, a cursory exploration of major online video hubs exhibits a clear indication of the extensive popularity and abundance of music videos today. Music video-related content is one of the most significant and prevalent content categories on YouTube. These videos consistently secure the top positions on the site's charts for the most viewed, most popular, and most discussed videos [7, 15]. After the inception of VEVO in 2009, by the end of 2011, the VEVO channel on YouTube had become the most viewed channel worldwide, garnering a staggering 56 billion views. Further, VEVO reported delivering approximately 3.3 billion views per month in 2012 [62, 15]. Considering the continuing popularity of music video content, it's unsurprising that these videos remain dominant in the most viewed content on YouTube. According to Statista [39], music videos have overwhelmingly dominated the top-viewed videos on YouTube since 2010, with just one exception. Furthermore, as of 2022, it is estimated that approximately 40% of the digital music audience in the United States consumes music through YouTube Music. Therefore, with the rise of online streaming platforms like YouTube, which have exponentially increased the accessibility of music videos to audiences worldwide, it is clear that music videos have become a vital part of today's music culture.

1.1.2. Unlocking the Power of Combined Modalities

The long-standing theory, proposing that intertwining multiple modalities results in superior cognitive outcomes compared to singular modalities [5], has garnered wide-ranging support within the realm of psychology [69]. This long-standing theoretical perspective highlights the cognitive benefits of utilizing multiple modalities simultaneously, suggesting that the confluence of sensory inputs fosters more effective information processing and retention [16].

Numerous studies across various domains have demonstrated the impact of audio-video content on users compared to audio-only [37, 56]. The inclusion of video enhances user interaction by providing non-verbal information and bolstering verbal descriptions, as compared to audio-only content [28]. In the realm of music, there is evidence that the combination of audio and video presents comprehensive images that result in the content being perceived as "more" creative, hot, bold, dynamic, and loud [69]. While music alone can stir emotions and convey messages, music videos grant artists the chance to tell the story behind their music. This combination of sensory inputs provides a more comprehensive experience by promoting a deeper emotional connection with the music [81].

Hence, the blending of the video modality into auditory music can significantly influence users' perception of music. As a result, user interaction with a streaming platform while viewing music videos can be influenced by the amalgamation of messages from multiple sensory channels, the music-audio and the visual music elements.

1.2. Research Gap

Numerous research efforts have been undertaken to understand the motivations behind why individuals listen to music, its role in day-to-day life, and its overall impact on them. Gantz et al. [18] discovered that individuals listen to music with the intention of filling uncomfortable silences, passing the time, mitigating feelings of solitude and managing their mood. In a similar study, Premuzic et al. [10] identified three primary motives for listening to music: cognitive/rational appreciation, emotional regulation, or as background sound while engaging in other activities such as studying, socializing, or working. Hargreaves et al. [55] also found that individuals often use their musical preferences as a "badge" to convey their views and values to others. Given its strong ability to influence memories and emotions, music has also been recognized for its therapeutic effects in helping individuals suffering from dementia in care homes [51].

Therefore given that music is a ubiquitous aspect of human culture and has been associated with broad psychological functions, it becomes important to know whether different people listen to music in similar or disparate ways [67]. Accordingly, comprehending the factors that contribute to variations in music listening behaviour is a critical inquiry in the field of music, as it has the potential to shape the music industry towards a more user-centric approach. Customizing recommender systems based on individual listening behaviours can optimize music discovery and recommendation algorithms, resulting in more personalized and gratifying user experiences. Furthermore, insights into user behaviours can inform design decisions pertaining to user interfaces and user services, leading to more engaging and user-centric music platforms.

Despite music videos being an integral part of our society, research centred around them is limited. While there have been some studies on the psychological effects of watching music videos and why people watch them [11, 6], to the best of our knowledge, there has been limited investigation into whether individuals exhibit diverse interaction patterns, based on the user actions, during music video consumption. Furthermore, the role of user context in shaping these viewing patterns remains largely unexplored.

It is important to note that past research has explored user interaction patterns by analyzing user actions within the auditory music sphere. For instance, Meggetto et al. [52] in their study scrutinized the user skipping behaviour during music streaming sessions. However, this research primarily focused on skip actions and did not encompass other user actions such as playlist changes, likes, and searches. Furthermore, as discussed previously, auditory music and music videos offer distinct experiences to users, which can essentially influence user behaviour. Therefore, further research specifically focused on music videos is needed to better understand and characterize user interaction patterns during music video streaming sessions.

1.3. Research Questions

With the goal of understanding user behaviour while consuming music videos, we address the following research questions:

- **RQ1: How can we characterize the interaction patterns that emerge during user interactions while streaming music videos?**

  In this study, we view user interaction signals within a session as a form of behavioural language. Consequently, by employing a Language Model, we generate embeddings pertaining to each session. Upon clustering these session embeddings, we identify and characterize distinct interaction patterns. This method allows us to gain insights into the diverse behavioural patterns which users exhibit while consuming music videos.

- **RQ2: What is the impact of temporal and genre-related contextual features on the interaction patterns?**

  The influence of temporal context and music genre on user preferences has been established and is further discussed in Chapter 2. To address this research question, we investigate the interaction patterns per session in relation to music genre, time of day, day of the week, and month. This analysis provides valuable insights into how these contextual variables may influence user behaviour in the domain of music video consumption.

1.4. Contributions

Through the answers provided for the above-mentioned research questions, the contributions of this thesis can be summarized as follows:

- **C1:** This study offers an analysis of user behaviour data, as gathered on an interactive music video platform - XITE. The primary focus lies in identifying and interpreting distinct user behavioural patterns during music video streaming sessions.
- **C2:** Our research expands upon the understanding of how contextual elements such as the time of day, month, and music genre affect user behavioural patterns. This serves to deepen our comprehension of the varied user behaviours arising while consuming music videos.
- **C3:** We introduce a unique methodological contribution by conceptualizing user actions as language and leveraging a Language Model for their analysis. While such techniques have been employed across different domains, their application within the sphere of music video interaction is unique. Moreover, we enhance this approach by integrating clustering methodologies and data visualizations, effectively addressing our posed research questions.
- **C4:** Furthermore, we establish and draw parallels between our interpretation of user session interactions in the music video streaming domain and the user behaviours observed in past studies related to auditory music.

1.5. Thesis outline

This document is organized as follows: Chapter 2 provides a comprehensive review of past studies, focusing on user feedback modelling, the application of Language Models in user action analysis, and the influence of context on user music preferences. The methodology employed in this research to interpret user action data is meticulously described in Chapter 3. In Chapter 4, we implement this methodology on user data and present the findings. These outcomes are subsequently discussed in Chapter 5, where we also reflect on our research questions, relate our work to prior studies, and discuss potential limitations and future directions. Lastly, Chapter 6 concludes this study.

2

Related Work

This chapter outlines the previous works and related background which forms the foundation of this study. It highlights the significance of user feedback signals and elucidates different methods for modelling these signals. It also describes the impact of context on the music preference of the users.

2.1. User feedback signals

In previous studies on recommender systems, both explicit and implicit user feedback have been utilized [88, 48, 29]. Explicit feedback generally involves user ratings, while implicit feedback refers to user session history, which may be less directly interpretable [91]. For instance implicit feedback includes user behaviours such as item views, link clicks, purchases, and duration of stay on a web page. Given the limited availability of explicit feedback and the abundance of implicit feedback [29], numerous studies have leveraged implicit feedback signals from users as crucial features in order to deliver more personalized and effective recommendations [3, 87]. Advances in Recommender System research also suggest that implicit feedback often provides more comprehensive and extensive insights into user behaviour than explicit ratings [14].

Further, studies have shown that it is crucial to model user interactions to comprehend user intent and provide more relevant search results [86, 33]. User interactions in terms of clicks (*e.g.,* the number of pauses or forward seeks) have been utilised to model video click behaviours in the e-learning domain to understand student behaviour and cater to their learning needs [43, 34]. In the field of information retrieval, researchers have utilised clustering methods to model user interactions and identify user groups that share similar browsing behaviour by analysing the clickstream data [20, 75, 79]. In the music domain, the research by Wen et al. [85] shows that skip and song completion are strong indicators of user preferences. Their research shows that the recommender systems that take skip information have a better performance in comparison to baseline approaches. Further, implicit user feedback in terms of skip and song completion have also been utilized to model user behaviour during the music listening sessions [52].

2.2. Modelling user Feedback/Interaction signals

2.2.1. Comparative Analysis of Sequence Modelling Techniques

The significance of sequentiality in clicks for characterizing short-term session behaviour has been established in previous research [61, 24]. Therefore, a variety of techniques have been utilized in past studies to manage sequential data. Markov-based methods have proven to be effective in representing and analyzing sequences of data due to their ability to accurately depict interconnected information [36]. Zhang et al. [90] employed a Hidden Markov Model (HMM) approach to identify patterns in user mobile usage data sequences. They computed a distance matrix between users and applied k-medoid clustering algorithm to extract sequential user behaviour patterns.

However, Markov-based methods have limitations in terms of generalizability, as explicit rules for all potential user action sequences must be provided to the system [46]. Additionally, these methods require significant storage and run-time resources in the given context, owing to the presence of multiple sequential patterns and potential states within the data [20]. Markov-based methods are also dependent on the assumption that a variable at time t relies exclusively on the values of the variable at time t-1. In our particular case, this assumption may not be applicable, as user actions during interactions with such platforms can be driven by user intent [53]. Further, it has been observed that user intent can persist throughout an entire session or undergo changes within the session [71]. Consequently, user interaction signals at time t may not be entirely independent of those at time t-n. This necessitates the adoption of higher-order Markov Chains, which would subsequently lead to an increase in computational complexity [45].

Furthermore, n-grams have been employed for modelling sequential user activities in various contexts. Lin et al. [47] utilized n-grams to characterize user action sequences during interactions with search engines. In a similar vein, Li et al. [44] implemented n-grams to model user clickstream data within MOOC platforms, with the objective of predicting future learning achievements. As previously discussed, user behaviour during music video consumption is contingent upon user intent, which can subsequently influence a sequence of user actions. Consequently, to discern distinct interaction patterns in our research, the utilization of longer n-grams is necessary. However, a key limitation of n-gram-based methods is the exponential growth of vocabulary size as n increases, resulting in a reduced number of users sharing identical tokens and, consequently, limiting the extraction of commonalities [21].

In order to encapsulate the session dynamics, a multitude of studies have demonstrated the suitability of models based on Recurrent Neural Networks (RNNs), as these networks are recognized for their proficiency in encoding temporal information [76]. Donkers et al. [14] used RNN to make next item recommendations on the basis of user interactions and the items they had consumed in the past. Hansen et al. [22] introduced a novel methodology to model sequential music skip behaviour within a session of streamed music content. Their model consisted of two stacked RNNs which functioned as encoding and predictor network, where the encoder network focused on encoding the first half of the user streaming session. Consequently, the decoder network utilised the received encodings to make sequential skip predictions. Nonetheless, RNN's suffer from vanishing gradient issue, [25] which make them ineffective for encoding longer sessions. This limitation in RNN's has been overcome by using Long Short-Term Memory (LSTM's) [26]. However, another constraint of RNN-based models is the stringent order assumption inherent in RNNs [82, 84]. This restricts the applicability of RNNs and LSTMs to sequences exhibiting a flexible order. Although user actions during music video consumption may display a certain order, acquiring a rigid ordering is unattainable in real-world scenarios.

2.2.2. Language Models

Natural Language Processing (NLP) techniques have been extensively employed in unsupervised topic modelling, wherein linguistic documents are clustered to identify similar topics [17, 8, 80]. Furthermore, prior research has successfully applied language models to generate embeddings for non-linguistic sequential data. For instance, Russac et al. [70] utilized `word2vec` to create embeddings for sequential data in an unsupervised manner, addressing credit card fraud detection. Prior studies have also adopted the `doc2vec` model, as proposed by Mikolov et al. [42], to represent user behaviour within a session. Phi et al. and Ludewig and Jannach [60, 50] employed `Doc2vec` to model user behavioural data in e-commerce and hotel ranking contexts, respectively. In these applications, user actions were regarded as words, while sequential actions in a session were interpreted as sentences. Drawing inspiration from such prior research, in our study, we consider user behaviour as a sequence of intra-session activities.

Word2Vec

Mikolov et al. [54] proposed Word2Vec which is an unsupervised learning algorithm designed to learn distributed representations of words by capturing their semantic and syntactic properties. Thereby it converts words to vector representations, called word embeddings. The model consists of two architectures: Continuous Bag of Words (CBOW) and Skip-gram as shown in Figure 2.1. Both of the architectures employ shallow neural networks, trained to reconstruct the linguistic context of words. The CBOW architecture predicts the target word given its context, while Skip-gram predicts context words given a target word. Furthermore, training the Word2Vec algorithm involves maximizing the likelihood of observing actual context words, using optimization techniques like hierarchical softmax and negative sampling for computational efficiency. Thereby, the learned word embeddings capture meaningful relationships between words, benefiting a range of NLP tasks.

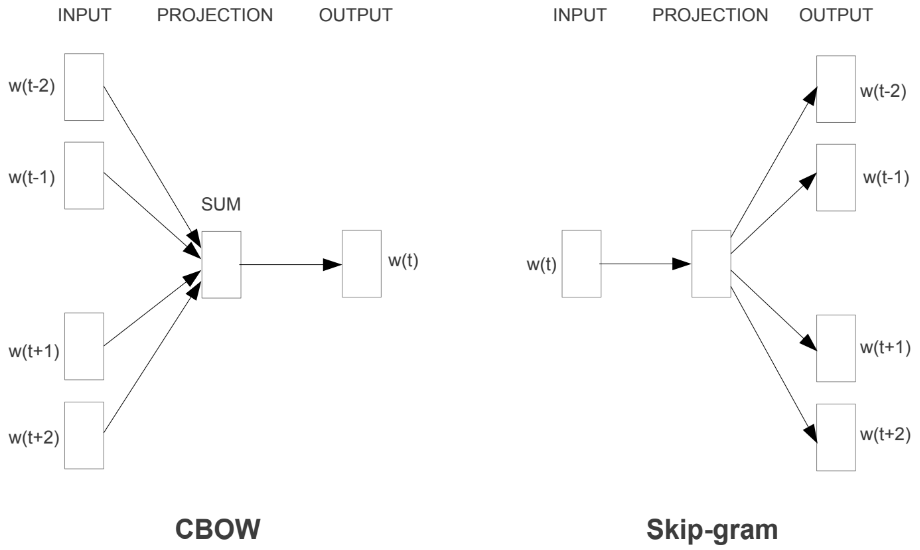

Figure 2.1: Word2Vec architectures - CBOW and Skip-gram [54]

Sent2Vec
Pagliardini et al. [57] introduced Sent2Vec, an unsupervised learning algorithm for producing fixed-dimensional sentence embeddings. This method expands upon the Continuous Bag-of-Words (CBOW) model utilized in Word2Vec by incorporating compositional n-gram features instead of merely unigram features. The model is designed to predict target words based on the contextual information of adjacent n-grams. Moreover, Sent2Vec implements positional weighting, attributing varying weights to n-grams contingent upon their placement within the sentence. This strategy facilitates the acquisition of both local word order information and global sentence-level semantics.

Doc2Vec
Le and Mikolov [42] presented the Paragraph Vector, or Doc2Vec, an unsupervised learning technique aimed at obtaining continuous distributed vector representations for pieces of text, by building on their prior research on Word2Vec for learning word embeddings. A key benefit of the Doc2Vec approach lies in its capacity to accommodate text documents of disparate lengths (from a phrase or sentence to a large document) while maintaining the sequence and semantic properties of the documents when representing them in a continuous vector space. The authors have proposed two distinct architectures for training the Paragraph Vector: Distributed Memory (PV-DM) and Distributed Bag of Words (PV-DBOW).

In the PV-DM architecture, the paragraph vector is concatenated with word vectors to predict the next word in the context window. This approach aims to preserve the order of words and learn the contextual relationship between them. During training, the paragraph vectors and word vectors are updated via stochastic gradient descent and back propagation. Further, during inference step, paragraph vector representations for previously unseen documents is generated by keeping the learned word vectors and other model parameters fixed, while optimizing the paragraph vector using gradient descent. The PV-DBOW architecture, on the other hand, is similar to Skip-gram model in Word2vec and ignores the context words and focuses solely on learning the paragraph vector by predicting the words in the document.

Considering the inherent ability of the Doc2Vec model to adeptly handle text documents with varying lengths, from brief phrases to comprehensive documents, it stands out as the most appropriate choice for this study, especially in the light of potential variability in session lengths that may be encountered. Furthermore, the proven efficacy of the Doc2Vec model in encoding non-linguistic sequences, as demonstrated in prior works, bolsters the rationale behind opting this method in the current study.

2.3. Context and music preference

Numerous studies have emphasized the importance of understanding the user's context when providing services, such as music recommendations, as their preferences may be influenced by their current context [27] [58]. On a music video streaming platform, users' behaviour, such as liking or skipping videos, is expected to be affected by their music preferences [85]. Prior research has demonstrated that these preferences can be shaped by the users' context [38, 32]. Thus, it is crucial to consider the interplay between users' behaviour, music preferences, and context, to effectively cater to their needs and expectations on such platforms.

2.3.1. What is context?

One of the initial works on context aware computing Schilit et al. [74] defined context in terms of *"where you are, who you are with, and what resources are nearby."* As context awareness gained more traction, Dey et al. [13] defined context as *"any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves."*
Therefore, all the factors which influence the interaction between user and the application can be combined as context.

Multiple researchers have incorporated user context for effective music recommendations to the user. For instance, Park et al. in [58] utilized sensors like - temperature, humidity, noise sensors; user profile and system information to infer the user context. They exploited fuzzy Bayesian networks to obtain the context and utility theory to inspect user music preference given the context. Kaminskas et al. [32] used the place of interest to user to select music relevant to the user.

2.3.2. Types of context

Abowd et al. and Razzaque et al. [2, 65], in their work have proposed various categories of context. Goker et al. [19] also presented five distinct context categories: (1) Environmental context, which includes the entities that surround the user, such as objects, services, and temperature; (2) Personal context, which comprises the user's physiological and mental states; (3) Task context, which describes the activities being undertaken by individuals within the user context; (4) Social context, which characterizes the social aspects of the current user context, including information about friends, neighbors, and colleagues; and (5) Spatio-temporal context, which pertains to the temporal and spatial dimensions of the user context.

However, for our study, we find the context classifications proposed by Kaminskas and Ricci [31] to be more befitting, given that their work is focused on the music domain and shares some overlap with the previously mentioned categories. Kaminskas and Ricci [31] introduce a schema consisting of three main categories: (1) environment-related context, (2) user-related context, and (3) multimedia context. The environment-related context captures user location, time-related information such as the time of the day and day of the week, and weather-related data. The user-related context contains details about the user's current activity, emotional state, and demographic information. Multimedia context refers to other multimedia sources (apart from music) that users might be exposed to during their interaction with the platform. This classification enables a more nuanced understanding of user contexts, particularly within the music domain.

Schedl *et al.* [73] also mention music context as an important factor that influences human music perception which for *e.g.,* encapsulates information on song lyrics, music video clips, and artists' background.

Since the platform considered in our work does not retain any data regarding user demographics, and due to the opacity of information in real life regarding user activity, emotional state, and other multimedia platforms the user may be engaged with, we limit the focus of our study to the environmental (*temporal*) and music-related (*genre*) contexts.

2.4. Summary

In this chapter, we reviewed previous research and observed that implicit user feedback signals have been extensively utilized to model user behaviour and interaction patterns. We also analyzed multiple methodologies for modelling user action sequences within a session. Furthermore, we familiarized ourselves with the association between contextual factors and user behaviour.

3

Methodology

The Research Questions that this study seeks to address have been outlined in Chapter 1. Consequently, this chapter elucidates the methodology employed to address both Research Questions, drawing upon the insights obtained from the previous works discussed in Chapter 2. An overview of the followed methodology is presented in Figure 3.1.

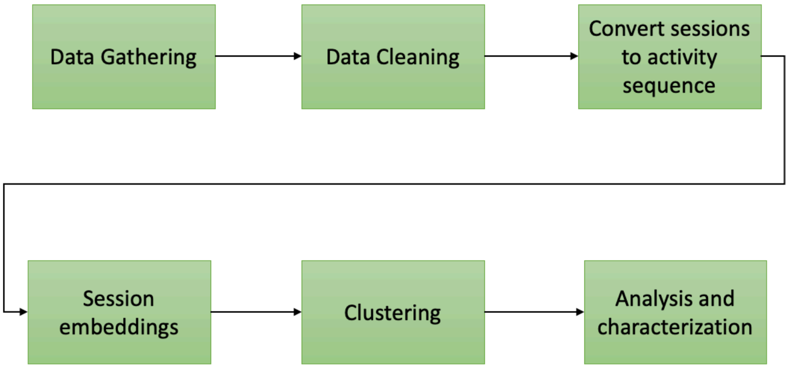
Data Gathering
Data Cleaning
Convert sessions to activity sequence
Session embeddings
Clustering
Analysis and characterization

Figure 3.1: Overview of experimental method

The following sections delve into the process in greater detail, providing a comprehensive account of the approach taken to answer the Research Questions.

3.1. Data Gathering

This study collected user activity data per music video streaming session, as logged by XITE. The data collection was exclusively focused on users located within the United States of America (USA), Canada, and United Kingdom (UK) during the period of March 2022 to January 2023. An overview of the session attributes that were retrieved and processed in this study can be found in Table 3.1. Further the sessions in which users consume curated playlists on their televisions, through their selected digital media players have been considered. These playlists tend to have cohesive content, and they auto-start once selected. Hence, a typical user flow is expected to contain content-based interactions with those playlists.

User Interaction Data

To study user behaviour during their streaming sessions, the "`event`" attribute from the session data has been utilized (see Table 3.1). This attribute encompasses all possible interactions a user can perform, including: 1) "`video completed`" when a music video has been played until its end, 2) "`skip`", 3) "`like`", 4) "`unlike`", 5) "`playlist change`", 6) content "`search`", 7) "`pause`", and 8) "`unpause`". Even though

Attribute	Description
session_id	Identifier of streaming session on the platform
video_id	Identifier of music video completed
playlist_id	Identifier of playlist curated by experts (optional)
video_type	Type of the video (music/ad)
event	The action performed by the user
timestamp	The local timestamp of the event's occurrence
genre	Genre of the specific music video

Table 3.1: Overview of the data schema used to analyse user interaction patterns on the music video streaming platform - XITE

the pause and unpause features were released on the platform during the period of the study, they were retained in the data. The decision was primarily driven by the lack of information regarding their potential impact on behavioural patterns. A list of all events (actions) along with their descriptions has been outlined in Table 3.2.

Event	Description
Playlist change	Changed the music video playlist
Video completed	The entire music video was played
Search	User searched for a particular video
Skip	The particular video was not completed, and the user switched to the next video
Like	Like button was pressed
Unlike	Unlike button was pressed
Pause	Music video was paused
Unpause	Music video was unpaused

Table 3.2: Events (user actions) logged on the considered music video platform along with their descriptions

Contextual Data

To conduct a contextual analysis in subsequent parts of this study, in addition to the attributes mentioned in the Table 3.1, the timestamp attribute was utilized, which encapsulates the local timestamp of the user. Additionally, the genre attribute has been employed to identify the genre of each music video.

3.2. Data Preparation

3.2.1. Data Cleaning

Similar to other streaming services, ads are shown by XITE during the streaming sessions. We filtered out ads from session data by using the attribute video_type as watching ads is not a user-initiated action. In subsequent stages of the study, expire and close events were also excluded from the sessions. They have been removed since it cannot be ascertained whether these events were user-initiated or happened due to external factors like hardware problems or power outages.

3.2.2. Sessions as Activity Sequences

As every session contains a series of video plays with associated user actions, user behaviour in every session has been represented as a sequence of user actions grouped by associated videos. For instance, consider a hypothetical scenario for a user, User A, as depicted in Table 3.3. User A watched

Video_id	Actions
1	skip
2	like, like, skip
3	video completed
4	pause, unpause, video completed
5	like, video completed

Table 3.3: Toy Example for user actions in a session

five videos in this session and performed a series of actions. The first video was skipped, and then the user liked the second video twice before ultimately skipping it. The third video was played to its entirety, and so on for subsequent videos in the session. By grouping the subsequent actions per video, the behavioural nuances are retained, which otherwise may had been lost by solely analyzing the session as a sequence of actions. Since our current interest is purely in user action sequences characterising each session, we decided not to encode which exact video and playlist was played, and for how long. This resulted in each session being represented as a sequence of activity sequences, such as, the toy example in Table 3.3 has been represented as : *[[skip], [like, like, skip], [video completed], [pause, unpause, video completed], [like, video completed]*. Our approach aligns with previous practices in literature where the sequentiality in clicks has been described as an important aspect while characterising short-term session behaviour [61, 24].

After analyzing the generated action sequences, it was observed that duplicates of the same action were occasionally logged on the same music video. This could have resulted from a slow internet connection, a false understanding of the interaction (e.g., multiple presses on ”like” do not yield different results), or miscommunications between the platform’s asynchronous processes. To tackle this semantically uninformative user input, only the first instance of such actions within each video activity sequence was retained. Therefore, consecutive like actions on a video were combined into a single action. Similarly, instances where a user repeatedly clicked on the skip button were also merged, as it mostly indicated the delay in loading the next video, presumably due to slow internet connection. Consequently, the running example depicted in Table 3.3 would be transformed into the action sequence: *[[skip], [like, skip], [video completed], [pause, unpause, video completed], [like, video completed]]*, where the consecutive likes on video 2 in the session are merged.

Furthermore, to ensure data quality, action sequences per video which may have arisen due to hardware or logging issues and were not logically possible were eliminated from the dataset. For instance, sequences such as ”like, skip, video completed” were removed as it is not feasible for a user to skip a video and still continue watching it until the end.

3.3. User Behaviour Characterization

3.3.1. Session Embeddings

In the context of examining user behaviour on a music video streaming platform, it is crucial to acknowledge the variability in duration of sessions. This variability poses a challenge when comparing activity sequences, as the length of each session may differ substantially. Simple padding techniques could adversely impact computational performance; thus, there is a necessity to construct representations of these sessions that yield equal length without compromising the information encapsulated within each session.

As delineated in Section 2.2.2, previous research has effectively utilized Language models to generate embeddings to represent user behavioural sequences. Therefore we employ Language models in this study to capture a user’s interaction with the platform as they tend to handle sequenced tokens well.

Consequently, adhering to the structure of natural language, each user action was treated as a letter, each activity sequence within a video as a word, and a session as a sentence. As a result, each ’word’ in a ’sentence’ mirrors the actions performed by a user on a specific video. Our methodology is informed by Phi et al. [60] and Ludewig et al. [50], who employed a similar representation in their studies to model user behaviour data in e-commerce and hotel ranking domains respectively, as discussed in Section 2.2.2. Consequently, each value of the `event` attribute is represented in our generated action sequence as illustrated in Table 3.4. Therefore, for instance, considering our ongoing example in Table 3.3, the resulting sequence would be represented as <*s ls p hkp lp*>, consistent with the provided action representations.

We further employ the Gensim implementation [66] of PV-DM (Distributed Memory version of Paragraph Vector), `Doc2Vec`, to transform session-based action sequences into a more compact representation. We chose PV-DM over the Distributed Bag of Words Paragraph Vector owing to its innate ability to comprehend word semantics and consider word order, as outlined in Section 2.2.2.

The goal of this study is to find different behavioural patterns in a session. Therefore, sessions which

Event	Representation
`playlist change`	c
`like`	l
`video completed`	p
`skip`	s
`unlike`	u
`search`	f
`pause`	h
`unpause`	k

Table 3.4: Representation of user actions in session activity sequences

HyperParameter	Value
`vector_size`	300
`epochs`	400
`dm`	1
`min_count`	1
`window`	15
`sample`	1e-5
`negative`	5

Table 3.5: Doc2Vec HyperParameter values

exhibit similar action patterns should intuitively be very similar to each other. This draws parallels with the text similarity task, which suggests that our `Doc2vec` embeddings can be trained using a similarity-based approach. As our study is completely based on unsupervised learning and has no predefined labels, there is no definitive method to evaluate which hyperparameters perform better. This provides us with a good rationale to train our `Doc2vec` model with hyperparameters similar to [41], which have been found to be the most effective for document similarity works. Therefore the consolidated list of hyperparameters used in this study have been listed in Table 3.5.

Furthermore, the `Doc2vec` model has been exclusively trained on unique action sequences. Utilizing unique sequences circumvents overfitting on prevalent sequences and underfitting on less common ones. Thus, ensuring that all the sequences are well represented. Nonetheless, this method may introduce the potential pitfall of over-emphasizing exceptionally rare sequences. The model's attempts to accommodate these sequences could potentially impair the quality of the embeddings for more frequently seen action sequences. However, as our goal is to comprehend the clickstream behaviours exhibited by users while consuming music videos, we treated all sequences equally during the training process. Subsequently, on training the `Doc2Vec` model with the hyperparameters outlined in Table 3.5, a 300-dimensional document vector `doc-vector` is acquired for each distinct action sequence.

3.3.2. Clustering

To characterise the behaviours exhibited in each session, we clustered the inferred embeddings for all sessions, using the K-Means++ implementation from Scikit-learn [9].

The "Elbow method" and "Silhouette score" are the two most widely used methods to determine the number of clusters [72]. The Silhouette score is known to be suitable for clearly separable clusters [78]. However, we found that the Silhouette score was not an ideal metric for assessing the number of clusters within our dataset. Our Silhouette scores ranged from 0.18 to 0.2 for varying values of k, indicating that the behavioural clusters in our data were not distinctly separable (as the score was closer to 0 and farther from 1). Consequently, we employed the Elbow method [4], acknowledging its subjective interpretation due to the potential absence of a clear elbow, to ascertain the optimal number of clusters.

3.4. Analysis

Cluster Characterization

To evaluate our inferred clusters, we conducted a correlation analysis between the different values of the `event` attribute of sessions (refer to Table 3.1) and the cluster category as the target variable. Utilizing one-hot encoding, we obtained boolean columns for each cluster category. Subsequently, we employed the Point Biserial correlation, which is mathematically equivalent to Pearson's correlation, to determine the correlation between them (as implemented in SciPy [83]). We selected the Point Biserial correlation method because it is suitable for assessing the correlation between a continuous variable (number of each interaction per session) and a dichotomous variable (cluster labels) [77]. A t-test with n-1 degrees of freedom was applied for statistical significance.

To gain further insight into the user behaviour represented by each cluster, we analyzed user interactions during streaming sessions within each cluster. Based on the different values for the attribute `event` (e.g., `video completed` and `skips`), we identified the most prominent user action within each behavioural cluster and qualitatively deduced the cluster characteristics. Drawing inspiration from Eren et al. [17], we further visualized the clusters using t-SNE by converting 300-dimensional embeddings per session to 2 dimensions. While we recognize that the process of dimension reduction might result in potential information loss, it's important to note that these reduced dimensions are primarily employed for the purpose of visual representation of our high dimensional clusters. This facilitates a more accessible understanding of the high-level trends and patterns within the data.

Behavioural n-gram Sequences

The behaviour sequences have been further analysed as n-grams, similar to [63]. This n-gram based approach was employed to identify and comprehend behavioural patterns unique to each cluster, using both the distribution of sessions containing the particular n-gram across the clusters and by calculating the Term Frequency (*tf*) of each n-gram within all the clusters. The *tf* calculation was subsequently normalized within each cluster using the Scikit-Learn's MinMaxScaler, a technique preferred for its ability to preserve the shape of the original distribution.

The Term Frequency *tf* quantifies the occurrence of an n-gram within a specific cluster. A high *tf* score within a cluster suggests that the corresponding n-gram behaviour is common and possibly indicative of the cluster's behaviour. Furthermore, if we observe that sessions containing a particular n-gram are predominantly linked to a specific cluster, it could imply that this behavioural pattern might serve as a distinctive characteristic of the behaviour displayed by that cluster. Thereby, by incorporating both, the normalized *tf* score and the distribution of sessions containing specific n-grams within clusters facilitates the development of a metric that allows to assess the relative association of each n-gram to every cluster. Therefore for an n-gram n_i present in a session grouped under cluster c_j, the relative association of n-gram n_i with cluster c_j, A_{n_i,c_j} has been computed as:

$$A_{n_i,c_j} = tf_{n_i,c_j} * F_{n_i,c_j}$$

where $tf_{n_i,c}$ represents the normalized term frequency of n-gram n_i and F_{n_i,c_j} represents the fraction of all sessions containing n-gram n_i grouped in cluster c_j.

Note that the association metric I_{n_i,c_j} ranges from 0 to 1 where values closer to 1 denote higher association of n-gram n_i with behavioural cluster c_j.

In our analysis, we refrained from employing the Inverse Document Frequency (*idf*). This strategic choice was driven by the consideration that *idf* would penalize an n-gram should it occur ubiquitously across most sessions within a specific cluster [64]. Contrary to this, our objective posits that frequent occurrence of an n-gram within a cluster might actually denote a characteristic behaviour pattern of that cluster. As a result, we assert that these recurrent n-grams might be reflective of the behaviour characterized by a cluster and should thereby maintain high scores in order to precisely portray their significance within the corresponding clusters.

Moreover, we opted to use fraction of all sessions containing given n-gram per cluster over the use of Document Frequency. The rationale behind this choice is that using a fraction provides a comparative measure of the number of sessions within a cluster containing a specific n-gram against all sessions that feature this n-gram. This approach provides more information than merely counting the number of sessions in a cluster that contain the n-gram.

The association metric is therefore employed to identify n-grams that may hold relevance to distinct clusters. This is achieved by deploying a box-plot to analyze the distribution of scores and thereby, discern n-grams that are potentially informative. Subsequently, these selected n-grams have been visualized using a heatmap to examine their relationships with different behavioural clusters, providing an intuitive and graphical representation of the association patterns.

Context based analysis

Ultimately, we sought to comprehend the impact of a session's temporal elements, and the predominant genre enjoyed during the session, on user's streaming behaviours. As the majority of our data pertained to users from the USA, we chose to focus our contextual analysis specifically on this region.

We selected sessions with duration exceeding one minute to ensure the analysis of only meaningful user engagement. Sessions longer than 16 hours were also removed to approximate the maximum length of a typical streaming session, assuming the user's recommended 8 hours of sleep [68].

We extracted temporal information from the starting time of each session using the `timestamp` attribute, encompassing: a) *time of the day*, b) *day of the week*, and c) *month*. We examined the distribution of sessions within each cluster with regards to these temporal attributes. With respect to genre, we determined the most frequent genre per session, based on the music videos viewed during it. Upon obtaining the most frequent genre per session, we computed the mode for each cluster to discern the genre predominantly associated with a behavioural cluster in our contextual analysis.

4

Results

This chapter describes the results achieved by implementing the methodology outlined in Chapter 3. These results help us address the aforementioned Research Questions by discussing the user behavioural patterns while streaming music videos and the impact of context on these behavioural clusters.

4.1. Data Overview

This study collected an extensive set of data, amassing about 1.8 million sessions from almost 270,000 users, mostly based in the United States. The data spans across the years 2022 and 2023, thus encompassing various time periods, including major holiday seasons such as Christmas and summer. These periods typically show variations in user behaviour and web traffic compared to other times of the year. Notably, the traffic was the highest during popular vacation months, specifically August and December where these two months accounted for around 20% of all the recorded sessions.

The sessions were generally evenly spread with in the day, with a noticeable dip in traffic between 11 pm and 4 am. Also, approximately a quarter of the sessions in our dataset occurred during weekends. Further, it is important to note that no demographic information about the users was available to take into account for this research.

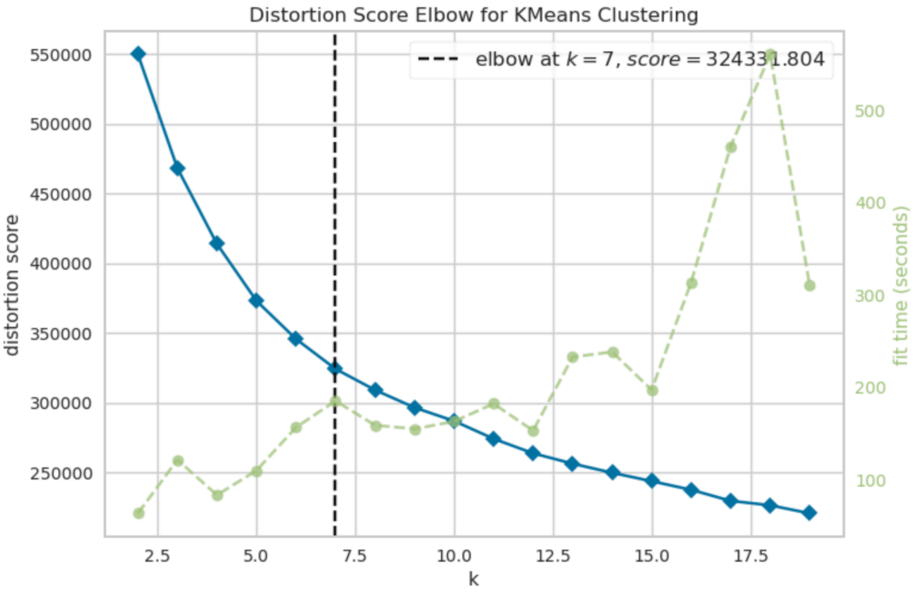

Figure 4.1: Number of Clusters

4.2. Clusters of Embedded Behavioural Sequences

As elaborated in Section 3.3.2, the Elbow method has been employed to ascertain the optimal count of clusters pertaining to the 300-dimensional session embeddings. The evaluation revealed the most suitable number of clusters to be seven, as depicted in Figure 4.1.

Further analysing each K-Means++ cluster by examining the point biserial correlation between the user actions, and the cluster each action sequence was assigned, unveil homogeneity in behavioural traits, as characterized by the actions performed by users within sessions. The p-value's corresponding to almost all of the point biserial correlation coefficients, associated with the different `event` attributes were found to be significant at the $p < 0.05$ level (see Table 4.1). Therefore, it is worth noting that each cluster is distinctly correlated to different user actions, implying that our clustering method successfully revealed patterns within the user behaviour embeddings.

Cluster	`video completed`	`like`	`skip`	`playlist change`	`search`	`pause`	`unpause`
C0	**-0.12***	-0.01*	-0.05*	0.01*	-0.03*	0.00*	0.00*
C1	**-0.15***	-0.03*	-0.14*	-0.11*	-0.06*	-0.01*	-0.01*
C2	-0.1*	0.00	**0.44***	0.02*	-0.02*	0.00	0.00
C3	-0.16*	-0.01*	-0.07*	**0.24***	-0.03*	0.00*	0.00*
C4	**0.54***	-0.02*	-0.10*	-0.06*	-0.04*	0.00*	0.00*
C5	-0.05*	0.01*	0.02*	-0.03*	**-0.25***	0.05*	0.04*
C6	-0.03*	**0.13***	0.06*	0.00*	-0.02*	0.00	0.00

Table 4.1: Point Biserial correlation and p-values for each cluster, where statistical significance ($p < 0.05$) is marked using an asterisk (*). Bold values indicate action types with the highest correlations.

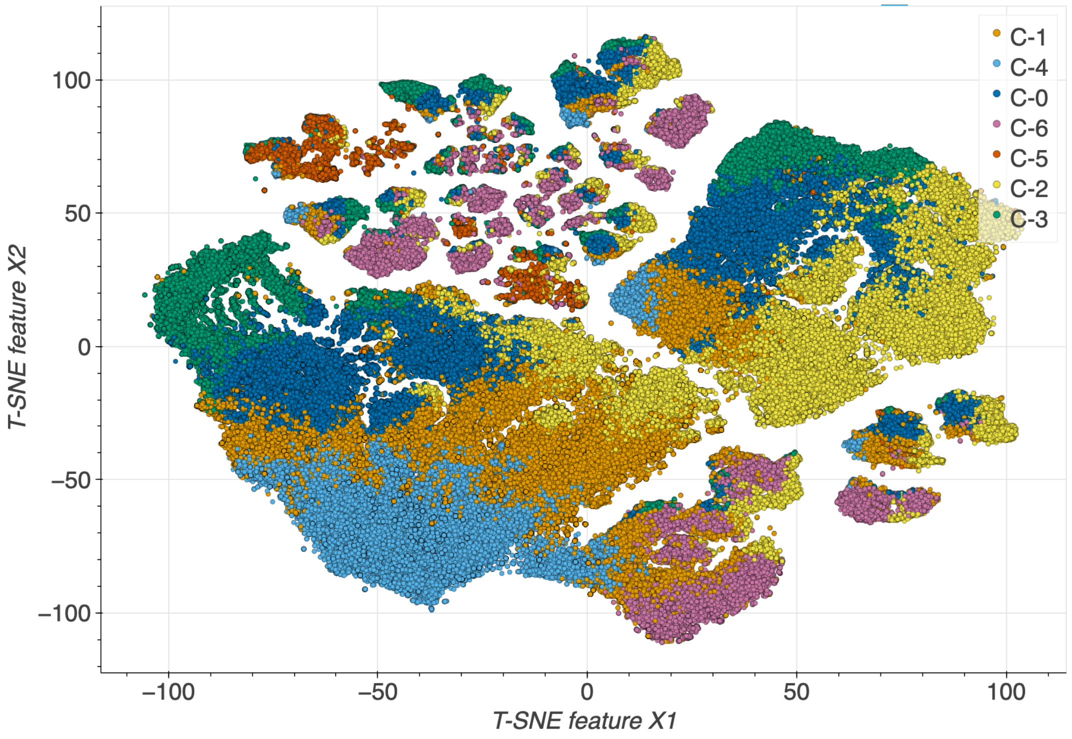

Figure 4.2: 'Cluster map' visualisation of embedded action sequences using t-SNE, where same-coloured dots represent a single behavioural cluster.

Qualitative analysis of the properties of found behavioural patterns further expand our understanding on the within-session behaviours. The two-dimensional features obtained via t-SNE, though not intrinsically interpretable, when used to visualise the clusters, clearly illustrate the relations of the clusters to each other. As shown in Figure 4.2, the clusters, denoted by differently colored dots, are distributed across the graph (due to projecting the high-dimensional vectors to a plane). This figure illustrates a

Cluster	video completed	like	skip	playlist change	search	pause	unpause	Session Characteristics
C0	27	0.16	0.9	2.9	0.15	0	0	Short sessions, many videos completed, few channel changes
C1	36	0.06	0.4	1.3	0	0	0	Shorter, passive listening sessions
C2	22	0.3	**9.6**	3.3	0.01	0	0	Exploration sessions
C3	9.9	0.2	0.8	**14**	0.01	0	0	Playlist juggling sessions
C4	**175**	0.6	0.26	1.4	0	0	0	Long, passive listening sessions
C5	28	1.4	2.2	1.4	**3**	0.1	0.1	Targeted listening sessions
C6	43.7	**7.8**	3.5	2.5	0.1	0	0	Selective listening sessions

Table 4.2: Mean values of user actions per behaviour cluster; the highest value per action is highlighted in bold.

pattern where each cluster tends to adjoin only a select group of other clusters. For instance, Cluster C4 adjoins cluster C1 in different sections of the figure. This qualitative observation intimates that the generated embeddings are informative and effectively encapsulate diverse user behaviours. This has been further discussed in further in this section.

Table 4.2 illustrates the mean action counts (video completed, like, skip, playlist change, search, pause, unpause) per cluster. Our analysis indicates that the sessions within behavioural clusters C0, C1, C2, and C5 have similar averages for videos played to completion. On the other hand, cluster C3 displays a considerably lower average of completed videos, whereas cluster C4 exceeds the rest with a higher average.

Behavioural cluster C4 is most correlated with video completed (Table 4.1) and also records an extraordinarily high count of fully watched videos (Table 4.2). This distinction suggests that sessions within cluster C4 typically exhibit a pattern of extended, passive engagement. This pattern is evidenced by a lower mean value for other user activities and the intuitive understanding that high video completion rates naturally extend the duration of sessions.

Behavioural cluster C6 is observed to have the second highest average video completed count, coupled with the strongest correlation with like actions according to the correlation analysis (Table 4.1) and also exhibit the highest average like count (Table 4.2). Furthermore, this cluster demonstrates a high mean for other user actions, including skip and playlist change. The combination of these traits reflects a more interactive and selective user engagement within the sessions encompassed in behavioural cluster C6.

Behavioural cluster C1 ranks third in terms of video completion rate and exhibits the highest correlation with video completion relative to other actions (Table 4.1). This cluster also shows lower average values for other user actions per session (Table 4.2), suggesting streaming sessions marked by lower engagement. Thus, we can infer these are shorter, passive sessions, especially when compared to those within cluster C4. It is intriguing to note the spatial positioning of sessions from clusters C4 and C1 in Figure 4.2, where they are often adjacent, hinting at behavioural similarity and transition between these clusters. On average, both C4 and C1 exhibit low user activity, indicating a preference for “sitting back and watching” content rather than interacting much with the platform.

Additionally, behavioural clusters C0 and C5 present comparable video completion rates. Cluster C5 exhibits the highest correlation with search (Table 4.1) and concurrently holds the highest average for search actions per session (Table 4.2). The sessions within cluster C5 also demonstrate high counts of skip and like actions, indicating more targeted streaming sessions. The highest average count of searches in C5, compared to other clusters, is potentially indicative of a greater degree of purposeful and targeted listening. Although few sessions exhibit such behaviour, they seem to neighbour most other clusters, but primarily those from C2, C3, C6, and to a much lesser degree, those from C1 and C4 (Figure 4.2). This can be attributed to C2, C3 and C6 showcasing selective behaviour (highest average on skip, playlist change and likes), which can be associated with the targeted experience that searching content can provide.

On the other hand, sessions within behavioural cluster C0 show a stronger correlation with video completion relative to other factors similar to sessions in clusters C1 and C4 (Table 4.1). Notably, sessions grouped in cluster C0 register lower counts for like and skip actions but exhibit a higher frequency of playlist change in comparison to clusters C1 and C4. This suggests that sessions in behavioural cluster C0 tend to be of shorter duration (as it has lower video completion rates in comparison to C1 and C4), with many videos reaching completion primarily. They also exhibit frequent `playlist change` actions. Further, the recurring proximity of clusters C0 and C1 in Figure 4.2 is intriguing, suggesting potential behavioural transitions or similarities.

Moreover, sessions in behavioural clusters C3 and C2 are the lowest and second lowest in terms of `video completed`, respectively. Sessions in cluster C2 are characterized by the strongest correlation with `skip` and highest average `skip` count, intriguingly coupled with the second highest average `playlist change` (Table 4.1 and Table 4.2). The elevated frequencies of `skips` and `playlist changes` suggest a pattern of exploratory user behaviour within these sessions. Furthermore, sessions belonging to behavioural cluster C3 demonstrate the most significant correlation and the highest mean with `playlist change` action. Thereby indicating that sessions symbolize browsing behaviour, characterized by changing the music playlist very frequently.

Clusters C0, C1, C3, and C4 appear to be closely aligned, as visualized in Figure 4.2. The blending of clusters into one another is a compelling observation. For instance, the lower left section of the Cluster Map (Figure 4.2) displays an intriguing merge of cluster C4, denoting long passive sessions, and cluster C1, signifying shorter passive listening sessions. Cluster C1 then fuses into C0, which typifies passive listening sessions with sporadic playlist changes. C0 then blends into cluster C3 which marks sessions with more frequent playlist changes. This sequential fusion pattern among behavioural clusters C4, C1, C0, and C3 is mirrored in various other areas of the map, reinforcing their behavioural interconnections.

Furthermore, Figure 4.2 reveals the presence of the same behavioural cluster at distinct spatial locations. A closer quantitative analysis divulges intriguing behavioural variations within sessions of the same cluster. For instance, behavioural cluster C3 is discernible in two primary sections. One section, located towards the left, is characterized by sessions with numerous `playlist change` actions and few video completions. Conversely, the section on the right exhibits sessions with frequent `playlist changes` and sporadic `skips`. Notably, the video completion rate is strikingly lower in the sessions from the left section of C3, averaging at 3.9 completed video streams per session, compared to an average of 15.2 video completions per session in the right section.

Likewise, behavioural cluster C0 appears in two primary sections, both demonstrating similar values of `video completed` and `playlist changes`. Nonetheless, a marked difference is observed in the number of skips: sessions in the right section of C0 exhibit a significantly higher number of mean skips (2.5) compared to the left section (0.37). A parallel trend is also observed in behavioural cluster C2, where sessions in the left section register fewer skips than the right. This underscores the presence of subtle behavioural disparities even within the same cluster and the effectiveness of session-level embeddings to capture the same.

4.3. Behavioural Analysis via Action Sequence N-grams

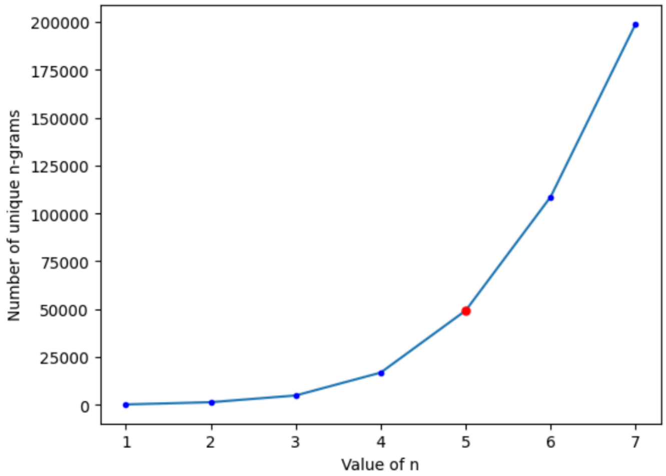

Figure 4.3: Number of unique n-grams for different values of n

We further scrutinized behavioural clusters by examining action sequences as n-grams. As illustrated in Figure 4.3, unique n-grams recorded in our dataset, increase exponentially as the value of n increases. Further, the frequency of recurrence of a specific n-gram correspondingly decreases, as expected. As a result, the n-gram transitions towards being a unique instance, undermining its potential for generalization as discussed in Section 2.2.1. Therefore, we have curtailed the maximum 'n' value to 5 in this study, considering the swift surge in the count of unique sequences.

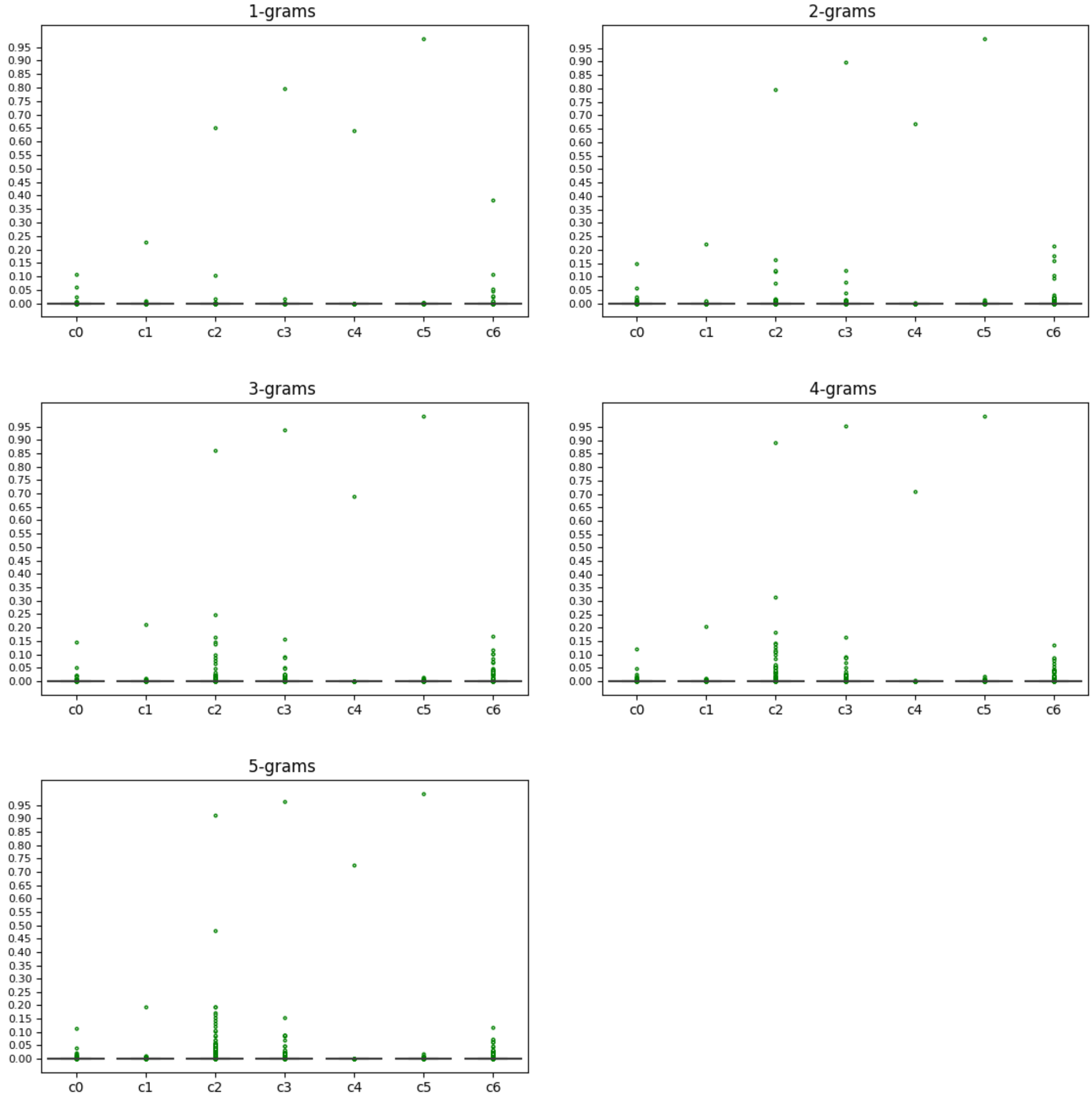

Figure 4.4: Boxplot visualization demonstrating the association between n-grams of various lengths and distinct behavioural clusters

Subsequently, the entire dataset embodies over 50,000 unique n-grams when considering sequential n-grams throughout all sessions. It is reasonably expected that among the extensive list of possible n-grams, only certain behavioural n-grams might act as a specific behaviour indicator. Indeed, as Figure 4.4 illustrates, the association measure for the majority of n-grams within each behavioural cluster tends to center around zero, with only a small fraction of n-grams per cluster registering high association scores. Therefore, these high-scoring outliers denote the n-grams with stronger affiliation to a specific cluster. For instance, cluster C2 in the 2-grams subplot of Figure 4.4 identifies four bigrams as outliers. Consequently, all the outlier n-grams in the Figure 4.4, with association value greater than 0.05 have been identified for a further nuanced analysis detailed in the subsequent paragraph. This criterion delivered 9 unigrams (1-grams), 16 bigrams (2-grams), 24 trigrams (3-grams), 33 quadgrams (4-grams), and 39 pentagrams (5-grams).

The above identified outlier n-grams have been represented in Figure 4.5. The action n-grams containing the `like` action display a stronger association with behavioural cluster C6 compared to other clusters, irrespective of the n-gram length. Notably, even a single instance of the `like` action within an n-gram elevates its association with cluster C6. Similarly, n-grams containing a `skip` action generally show a stronger link with behavioural cluster C2. However, this association shifts towards cluster C6 when the n-grams include both `skip` and `like` actions. For instance the n-grams <s, lp> and <lp, s> have a higher association with C6 than C2.

N-grams containing the `search` action consistently associate with behavioural cluster C5, irrespective of n-gram length. Similarly, all the n-grams which contain `like` action are associated with cluster C6. Additionally, n-grams beginning with the action pair `channel change, video completed` (`cp`), followed by successive `video completed` actions, are distinctively characteristic of cluster C0. Conversely, when `cp` precedes multiple `playlist changes`, it aligns the n-gram more with cluster C3. Furthermore, n-grams following a `playlist change` action with sequential `skips` tend to associate most with cluster C2. These observations further strengthen our earlier interpretation outlined in Section 4.2 where behavioural cluster C5 has been found to be correlated with `searches`, cluster C6 with `likes`, cluster C0 with `video completed and playlist changes`, cluster C3 with `playlist changes` and C2 with `skips`.

Interestingly, n-grams composed of `video completed` actions show stronger links with clusters C4, C1, and C0, in that order. This pattern aligns with our expectations, as these clusters show the highest correlation with the `video completed` action, as seen in Table 4.1.

Furthermore, we see the n-grams comprising of repeated `skips`, `playlist changes`, `videos completed` and `searches`, are strongly associated with behavioural clusters C2, C3, C4 and C5 respectively. This pattern, however, does not hold for the `like` action. This discrepancy can be attributed to the sporadic appearance of the n-gram <lp lp lp lp lp> within the sessions. Despite the fact that 99.38% of sessions containing this n-gram fall under behavioural cluster C6, its normalized term frequency is notably low, as outlined in Table 4.3. This observation implies that this particular action sequence is not as common compared to the other recurrent action sequences and has been further discussed in the Chapter 5.

Additionally, as seen in Figure 4.5, a considerable number of n-grams encompass the *skip* action as compared to n-grams featuring `search` or `like` actions. This observation is likely due to the predominant use of the `skip` functionality among XITE users, as evidenced by the frequency of skips being 47 times higher than likes in our dataset. Further, Figure 4.5 also suggests that varying permutations of actions within the same behavioural sub-sequences associate with the same behavioural cluster. For instance both the sub-sequences <p s s> and <s s p> are associated with the cluster C2. This observation, along with action recurrence within behavioural sub-sequences (<s s s s>, <lp lp lp lp>) has been further discussed in Section 5.1.

N-gram	tf C0	tf C1	tf C2	tf C3	tf C4	tf C5	tf C6	% C0	% C1	% C2	% C3	% C4	% C5	% C6
s s s s s	0	0	**1**	0.01	0	0.01	0.04	1.06	1.33	**91.23**	1.43	0.4	0.97	3.57
c c c c c	0	0	0.01	**1**	0	0.01	0.01	0.26	0.1	1.01	**96.48**	0.06	1.54	0.55
p p p p p	**1**	**1**	0.83	0.28	**1**	0.16	**1**	4.13	19.42	1.47	0.41	**72.43**	0.52	1.61
f f f f f	0	0	0	0	0	**1**	0	0.14	0.01	0.29	0.06	0.01	**99.37**	0.11
lp lp lp lp lp	0	0	0	0	0	0	**0.05**	0	0.01	0	0	0.13	0.47	**99.38**

Table 4.3: Distribution of the normalized Term Frequency(*tf*) of particular n-grams within each behavioural cluster, and the spread of sessions featuring these n-grams across the clusters (the values are rounded to two digits and the highest values per n-gram are highlighted in bold).

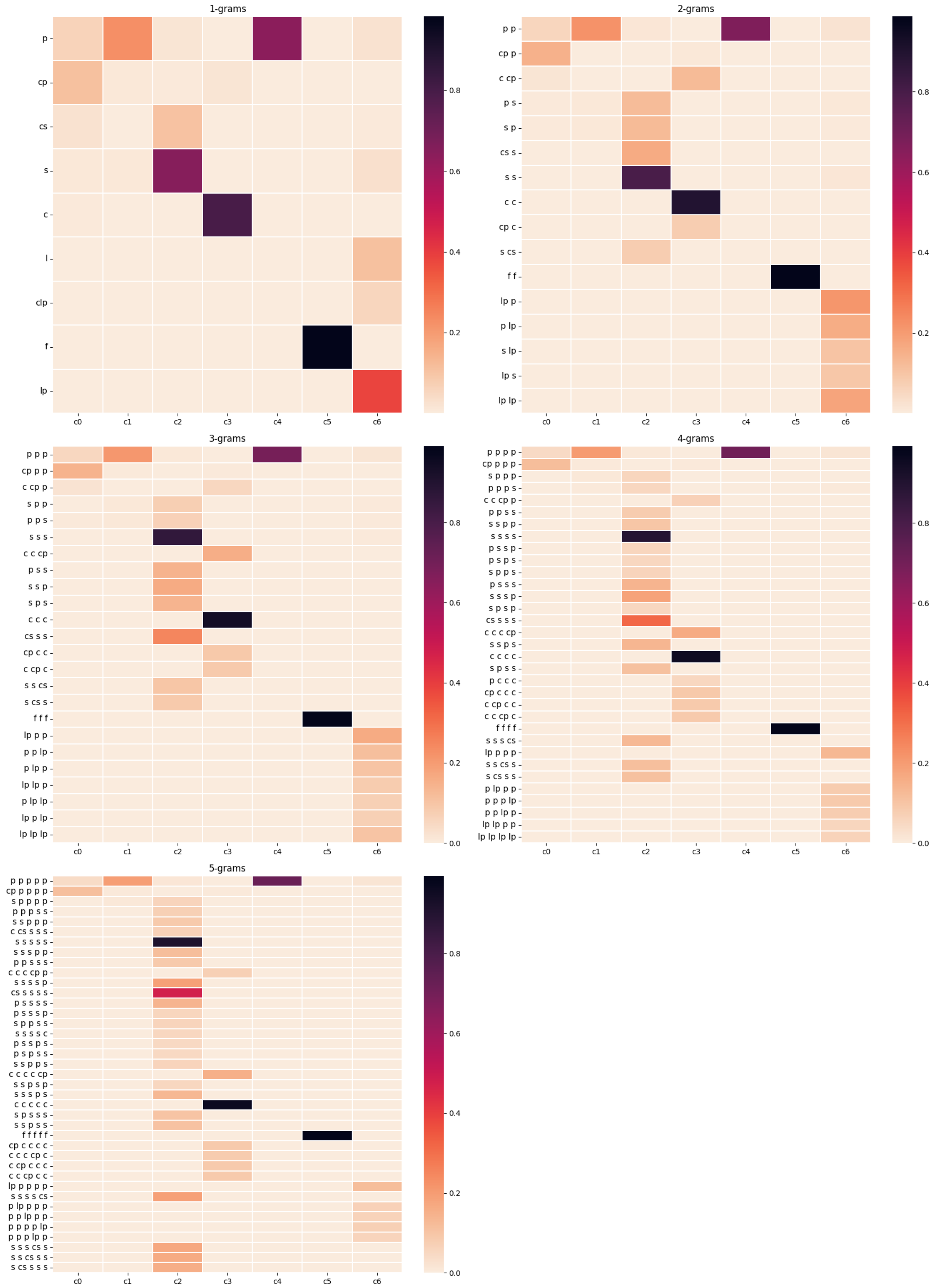

Figure 4.5: Heatmap illustration of the relationship between action sequence n-grams of different lengths and their association with distinct behavioural clusters

4.4. Contextual Influence on Behavioural Patterns

As discussed in Section 3.4, the contextual analysis conducted in this study primarily focuses on users residing in the United States of America. Moreover, only sessions with a duration ranging from more than a minute to less than 16 hours have been considered for the contextual analysis. Post this data filtering process, a total of 1.58 million sessions have been retained for further examination.

An examination of clusters relative to environmental context reveals discernible patterns: notably, long passive listening sessions—such as those in behavioural cluster C4—typically commence in the morning and peak around 10 a.m., as depicted in Figure 4.6. This distinct onset time, when contrasted with that of other clusters, suggests a tendency for users to engage in more interactive sessions as the day advances.

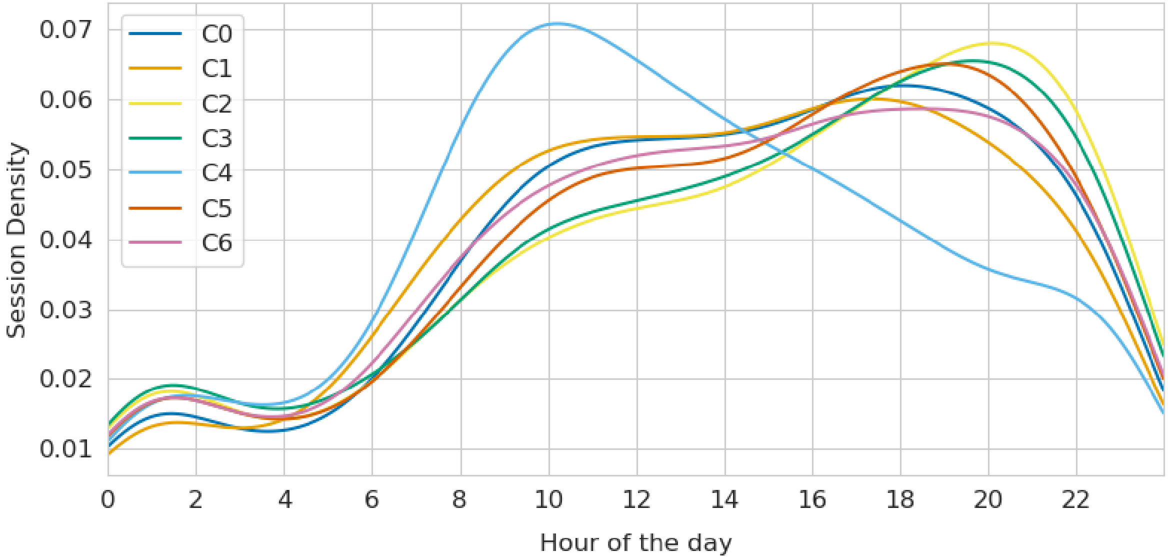

Figure 4.6: Session start density in relation to time of the day

Table 4.4 shows that clusters linked with more active behaviours (C0, C2, C3, C6) tend to peak in August, while the clusters associated with more passive behaviours (C1 and C4) peak in December. The December peak can reasonably be attributed to the holiday season in the USA, a period during which users may be more likely to engage in passive consumption of music videos, relative to other periods of the year.

The August peak is less immediately explainable with the data at hand. However, further consultation with the streaming company XITE, revealed a marketing campaign during that month which promoted the platform's interactive features. This could could conceivably elucidate the increase in more active session engagement during this timeframe. While this is not anticipated to impact the behavioural clusters or influence our other interpretations, a broader examination over an extended timescale could potentially enhance our understanding of whether these behaviours actually surge in August.

Further investigation into the influence of the day of the week on music consumption patterns revealed that the occurrence of all behavioural clusters peaked on Saturday. This could be a result of heightened user engagement with the XITE music streaming platform on Saturdays similar to heightened engagement on YouTube on Saturday's [1]. This can also possibly indicate a general preference for consuming more music videos on this particular day of the week. However, a definitive conclusion requires analysis of data across different platforms, spanning multiple years to account for potential variability and trends.

Exploring the relationship between behavioural tendencies and music video genres, we found that 'Pop,' 'Country,' and 'Rap/Hip-Hop' prevail as the most popular across the different behavioural clusters. 'Pop' and 'Rap/Hip-Hop' genres primarily correlate with active and exploratory behavioural clusters (C0, C2, C3, C5, C6). 'Pop' is mostly associated with relatively brief sessions, while 'Rap/Hip-Hop' appears

Cluster	Month	Genre
C0	Aug	Pop
C1	Dec	Country
C2	Aug	Rap/Hip-Hop
C3	Aug	Pop
C4	Dec	Country
C5	Jan	Rap/Hip-Hop
C6	Aug	Rap/Hip-Hop

Table 4.4: Most active months and most popular genres per behavioural cluster.

to connect more with sessions exhibiting selective behaviours, evidenced by higher counts of `skips`, `likes`, and `searches`. Conversely, ’Country’ emerges as the top genre for sessions in the clusters C1 and C4, both of which are characterized as more ’passive’, as they display significantly less interactive actions compared to other behavioural clusters.

5 Discussion

This chapter provides an interpretation of the findings presented in Chapter 4, effectively contextualizing the current study within the wider body of related research. Further, section 5.2 highlights the constraints inherent to this study, accompanied by suggestions for improvement and avenues for future research.

5.1. Closing the Loop: Revisiting our Research Questions

In this study, we aimed to unravel distinct patterns of user interaction while streaming music videos and interpret the implications of these patterns. Our structured analysis of user interactions, represented as embeddings, yielded discernible behavioural patterns across user sessions. While some of these patterns overlap, as visualized in Figure 4.2, there are also clear differences. Sessions grouped under behavioural clusters C2 and C3 characterize repeated skips and playlist changes, possibly signifying more exploratory user behaviour. In contrast, the sessions grouped in C4 indicate periods of more passive listening. The clear difference between passive and various active user behaviour should be taken into account when developing a video streaming service that aims at catering distinct audience types.

Furthermore, the characterized behaviours also align with the prior works. Previous studies have noted that individuals often play music and music videos as a background [49, 40], potentially explaining behavioural clusters C0, C1 and C4. Other research has shown that people frequently employ music for mood regulation [49, 35], potentially instigating targeted and selective listening sessions observed in clusters C5 and C6. Additionally, Lonsdale *et al.* [49] also established exploration and surveillance as motivations behind listening to music, which could explain behavioural clusters C2 and C3.

Moreover, we also found a pattern of recurrence for `skip, playlist change` and `search` actions. This, however, is a rare occurrence for the `like` action. This discrepancy may be because people sometimes do not remember the names of the songs and associated artists of the songs they wish to listen to [12] which might facilitate a sequence of `skips` or `playlist changes` or `searches`. It might also be the result of users' curiosity or desire to explore, resulting in successive `skips` or `playlist changes`. `Like` on the other hand, is implicit positive feedback, and people tend to press like when they genuinely like the song. Therefore, a succession of likes is intuitively less probable. These behaviour characterizations effectively address the **RQ1** of our research.

Additionally, in section 4.3, we observed that different permutations within an action sequence were associated with the same behavioural cluster. This suggests that the presence of an action in a sequence might be more significant than the particular order of the actions. This could potentially be a result of our study identifying higher-level behavioural clusters. For instance, we identified behavioural cluster C2 as sessions marked by exploratory behaviour, where users frequently `skip` tracks. Under this scenario, intuitively, the sequence of `skips` and `video completed` actions might not be important, as the primary observation is the high frequency of `skips` within this cluster. However, for a study targeting a more granular understanding - such as, delving deeper into exploratory behaviour by examining how many tracks users `skip` before settling on a video - the sequential ordering of actions could gain increased significance.

Moreover, the patterns identified in this study raise crucial questions from a user experience perspective. Questions such as, “how much time are users devoting to navigating between playlists?” and “how can this process be optimized to facilitate users connecting with their preferred playlists more efficiently?” are essential to investigate. These questions become even more significant for music video streaming platforms like XITE, which adopt a curated playlist approach. Frequent `playlist change` and `skip` sequences can provide valuable feedback to curators, helping to align playlists with user expectations and interests. This alignment will promote a user-centric platform where individuals can readily engage with their preferred music videos. It is also noteworthy, that sessions characterized by frequent `playlist changes` (cluster C3) demonstrate a lower video completion rate and consequently shorter sessions, as compared to sessions where skips are more frequent (cluster C2). This observation could suggest that frequent `playlist changes` may indicate a stronger (potentially negative) implicit user feedback than `skips`.

Upon discerning identifiable patterns of interaction, we delved deeper to explore whether such behaviour could be associated with contextual aspects, such as the time of day, the month of the year, or the genre of music listened to (addressing **RQ2**). Notably, sessions categorized within C4, indicative of long passive listening periods, were predominantly initiated during the morning hours. A peak in the commencement of such sessions was notably observed around 10 am (refer Figure 4.6). This finding is in line with the findings by Zhang et al. [89] who reported a similar trend of prolonged listening sessions on Spotify, primarily occurring in the morning on desktop devices. Moreover, our findings also suggest that the sessions with higher user interaction generally start towards the later part of the day. For instance, sessions with a high frequency of skips, categorized under the behavioural cluster C2, typically start in the evening. This observation parallels findings from a study conducted by Meggetto et al. [52] within the domain of audio-only music consumption. They reported that passive music listening is usually prevalent during the morning hours and that the frequency of skip actions surges in the evening. This pattern has been attributed to users consuming music videos more passively in the morning, possibly while engaging in other activities, and then becoming more actively involved by the evening. The congruity of these findings hints towards similar consumption patterns between the audio-only music and the music videos.

Additionally, certain patterns occur more frequently at specific times of the year, *e.g.,* C1 and C4 happen most frequently in December. Insights such as these can have enduring implications for music video streaming platforms. For instance, this knowledge can guide content selection by advocating for the promotion of longer playlists during December on the main platform homepage. Additionally, it can inform strategic business decisions for streaming platforms, such as offering complimentary premium subscriptions for the month of December or charging more, potentially catalyzing a surge in holiday peak (i.e., increase in user engagement) or adjusting the volume of advertisements considering the probable shift towards more passive content consumption during this period. Nevertheless, to establish a robust conclusion regarding user behaviour based on temporal attributes, a dataset spanning across years and sourced from various platforms is required, for the purposes of generalization.

Our analysis thus unveils opportunities for adaptation and personalization centered on user behaviour patterns. For instance, platforms could scrutinize user actions within a session and if the action sequence (n-grams) displays strong association with a characteristic behaviour, corresponding music video recommendations or other strategic interventions could be deployed. This process could be iteratively implemented at different times within a session to align with current user intentions. This approach could offer an adaptable, behaviour-centric user experience, potentially enhancing user engagement and satisfaction.

Finally, the user behaviours elucidated in this study can offer a foundational framework for emerging music platforms as they strategize to cater a diverse user base. In essence, our findings underscore the importance and potential of comprehending user behaviour patterns, thus enabling platforms to tailor their services based on user actions, consequently augmenting the user experience effectively.

5.2. Limitations and Future Works

This research, although illuminating in its findings, also surfaces several limitations that require consideration. Accordingly, this section delves into a discussion of these limitations and outlines prospective avenues for future research. In addition, we propose recommendations for music streaming platforms based on the insights drawn from our study.

5.2.1. Limitations

Since the intention of users can change within a session, this could modify the interaction signals and subsequently cause the session to embody various behaviours. Despite this, these sessions are typically classified into a cluster that aligns with their primary behaviour, potentially obscuring finer behavioural nuances. Additionally, while Doc2Vec is inherently agnostic to document length, it may not yield optimal embeddings for short sentences due to the presence of limited context [59]. This could be a focus for future work, refining and developing techniques for better handling of such scenarios.

Moreover, although K-Means++ selects initial cluster centroids following an empirical probability distribution, the grouping of sessions might vary to a certain degree across reruns. While this variability is unlikely to significantly alter the behavioural interpretation and overall outcomes of the study, the inherent stochasticity of K-Means++ could lead to changes in the classification of certain sessions, particularly those near cluster boundaries. This observation underlines the necessity for future research to investigate the potential impact of such variations.

Furthermore, the nature of the content and curated playlists within our collected data may incite interaction behaviours that are unique to the platform, the digital media player (Roku in this study) and to a certain degree, the demographics of our sample. Given the limited analogous studies in this domain, our findings are currently specific to music video streaming sessions on XITE and may not be generalizable beyond this platform. Therefore, we encourage further research in the domain of music video streaming to foster a more generalized understanding of user interaction behaviours across a larger variety of platforms.

5.2.2. Future Works

Further exploratory research into user behaviours during music video streaming sessions could delve into smaller, distinct patterns within individual sessions. This might identify various sub-patterns within a session that, when considered independently, gravitate towards different behavioural clusters i.e behavioural sequences corresponding to different user intentions. Consequently, the optimal length of behavioural sequences which can reflect user intention can also be determined. Such an analysis could refine our understanding of the change in user intention within a single streaming session, enhancing the efficacy of personalized recommendations and fostering an enhanced user experience.

Additional research could further examine the contextual factors influencing user behaviour during music video streaming sessions. In particular, our analysis revealed a peak in sessions exhibiting passive interaction during December. An intriguing aspect to further probe could be whether this surge is attributable to increased activity from existing users, seasonal users who primarily engage with the platform during this period, or new users whose sessions tend to be more passive. Consequently, an inter-session analysis per user could be instrumental in uncovering these patterns. A more profound understanding of user behaviours and their variations with subject to contextual factors, could inform the creation of more accurate user profiles, subsequently enhancing the personalization of recommendation strategies.

Another promising avenue for future work lies in predicting future user actions both within and beyond a session. Anticipating the user's immediate next action, or even predicting longer-term behaviours such as user return or churn, can be invaluable for tailoring the user experience and enriching user engagement. Moreover, examining behaviours from session to session within the context of the same user can shed light on short-term and long-term behavioural transitions. These insights could enhance our understanding of user engagement over time, informing strategic approaches for sustained user engagement with the platform.

Lastly, it would be beneficial to extend this analysis into the realm of advertising and other types of platform interventions, such as prompts. Evaluating the impact of advertisements and other interventions on subsequent user actions and behaviour patterns could reveal additional facets of user engagement. These findings could enhance the effectiveness of ad placements and other platform interventions, providing strategic insights that can bolster the overall user experience on music video streaming platforms.

5.2.3. Recommendations

Beyond the strategies outlined in Section 5.2.2, this section further explores potential next steps for the music video streaming platform, XITE, guided by the analysis and insights garnered from this study. These additional recommendations aim to further enrich the user experience.

- In behavioural clusters where Pop is the predominant genre, there is a higher frequency of `playlist changes`. This could suggest that Pop listeners are either more inquisitive, seeking to browse through various playlists, or possibly dissatisfied with the current playlist structure.

  To understand this behaviour further, an internal study could be initiated to trace the trajectory of users from one playlist to another. For instance, if a common pattern emerges where users typically navigate from playlist 'A' to playlist 'K' through multiple changes, this could suggest a thematic or stylistic link between these playlists. Consequently, it may be beneficial to merge these playlists to better align with user preferences.

  Furthermore, if such transitions are infrequent and no clear correlation is found between playlist transitions, it might be indicative of users' inherent inquisitiveness and desire to explore different playlists. In such a scenario, interventions such as prompts suggesting alternative playlists could be implemented. This approach could enhance user experience and satisfaction by making navigation and discovery more streamlined. The same can be extended to skips, where skip sequences can be analyzed for music video reordering within the playlists or for personalised interventions.

- Additionally, consecutive `search` action by the users (cluster C5) may suggest that the user does not recall the exact name of the music video or the artist. As such, by concatenating sequential search queries, the platform can gain a better understanding of what user is looking for. This improved insight can then be utilized to deliver more relevant search results, thereby enhancing user satisfaction and engagement.

- In the observed behavioural clusters, sessions with Rap/Hip-Hop as the predominant genre, had more video playbacks, thus leading to longer sessions in comparison to clusters dominated by Pop music. As a result, if the platform's objective is to prolong session duration, emphasizing or prioritizing Rap/Hip-Hop playlists within the user interface (UI) could be beneficial. Conversely, if the platform aims to foster exploration across various playlists, it might consider accentuating Pop music, which according to the current data encourages such user behaviour. Furthermore, as per this study, Country genre is associated with longer passive listening sessions. Therefore if the platform wants to encourage such engagement or maximize session lengths, the positioning of Country playlists within the UI could be adjusted to facilitate this objective.

- Moreover, incorporating user behaviour sequences into advertisement display strategies can enhance the effectiveness of platform's business model. For instance, during passive listening sessions, users may be less likely to skip ads, which could increase ad impressions and revenue. Conversely, since passive listeners may pay less attention to ads, limiting ad displays during these periods and focusing on active sessions might improve ad engagement and impact. Therefore ads on the platform can be placed in congruence with user behaviour and business strategy.

6
Conclusion

Humans share a timeless bond with music, a connection that surpasses cultural, chronological, and geographical barriers. The presence of music in various facets of our lives reinforces its relevance and importance. In the modern world, this connection has been further enriched by integrating music with the visual medium through music videos, shaping a new dimension of music consumption and appreciation. The prominence of music videos in today's digital age emphasizes the importance of understanding user behaviour in this domain for shaping a more user-centered music industry.

Despite the substantial role that music videos play in our society, the area of research focusing on user behaviours while consuming these videos remains significantly under-explored. Therefore, our study takes a step in that direction, providing insights by modelling user actions, that could potentially inform the design and development of more personalized and engaging recommender systems and music platforms. In this work, we sought to elucidate the patterns of user behaviour during music video consumption on the interactive platform - XITE. We modelled user actions as a language and employed a Language Model to model the user session behaviour. The user action sequences per session were thereby converted to embeddings and have been further clustered. In our work, we identified seven different behavioural groups which were analyzed and characterized to gain insights into the behavioural patterns. We examined these behavioural patterns in relation to contextual elements, such as time of day, month, and music genre, to gain a better understanding of the exhibited behaviours.

Interestingly, our analysis revealed that sessions with higher user interaction generally commence towards the later part of the day whereas more passive sessions tend to start in the morning, peaking at around 10 a.m. Moreover, in this research we also identified parallels between the observed user behaviour in the domain of music video streaming and established user behaviours documented in auditory music studies. This comparison facilitated a more comprehensive understanding of user behaviour spanning across different modes of music consumption.

We believe our findings offer valuable insights into user interaction behaviours, which can significantly contribute to the personalization and enhancement of user experiences on music streaming platforms. This study also identifies potential opportunities and limitations and outlines practical applications, serving as a valuable resource for future academic research and strategic decisions within the company. We encourage further research to expand upon our findings and hope that our work will be a foundation for subsequent explorations in this area.

References

[1] July 2021. URL: https://boosted.lightricks.com/when-is-the-best-time-to-post-videos-on-youtube/.

[2] Gregory D Abowd et al. “Towards a better understanding of context and context-awareness”. In: *International symposium on handheld and ubiquitous computing*. Springer. 1999, pp. 304–307.

[3] Linas Baltrunas and Xavier Amatriain. “Towards time-dependant recommendation based on implicit feedback”. In: *Workshop on context-aware recommender systems (CARS’09)*. 2009, pp. 25–30.

[4] Benjamin Bengfort et al. *Yellowbrick*. Version 0.9.1. Nov. 14, 2018. DOI: 10.5281/zenodo.1206264. URL: http://www.scikit-yb.org/en/latest/.

[5] Donald Eric Broadbent. “Successive responses to simultaneous stimuli”. In: *Quarterly Journal of Experimental Psychology* 8.4 (1956), pp. 145–152.

[6] Jane D Brown, Kenneth Campbell, and Lynn Fischer. “American adolescents and music videos: Why do they watch?” In: *Gazette (Leiden, Netherlands)* 37.1-2 (1986), pp. 19–32.

[7] Antony Bruno. “Fully Loaded Clip: Major Labels Exert Greater Control over Monetizing Music Videos”. In: *Billboard* 24 (2009), p. 7.

[8] Arif Budiarto et al. “Unsupervised news topic modelling with Doc2Vec and spherical clustering”. In: *Procedia Computer Science* 179 (2021), pp. 40–46.

[9] Lars Buitinck et al. “API design for machine learning software: experiences from the scikit-learn project”. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 2013, pp. 108–122.

[10] Tomas Chamorro-Premuzic and Adrian Furnham. “Personality and music: Can traits explain how people use music in everyday life?” In: *British journal of psychology* 98.2 (2007), pp. 175–185.

[11] Council on Communications and Media. “Impact of music, music lyrics, and music videos on children and youth”. In: *Pediatrics* 124.5 (2009), pp. 1488–1494.

[12] Sally Jo Cunningham and David M Nichols. “Exploring social music behaviour: An investigation of music selection at parties”. In: Ismir. 2009.

[13] Anind K Dey. “Understanding and using context”. In: *Personal and ubiquitous computing* 5.1 (2001), pp. 4–7.

[14] Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. “Sequential user-based recurrent neural network recommendations”. In: *Proceedings of the eleventh ACM conference on recommender systems*. 2017, pp. 152–160.

[15] Maura Edmond. “Here we go again: Music videos after YouTube”. In: *Television & New Media* 15.4 (2014), pp. 305–320.

[16] Mickie Edwardson, Donald Grooms, and Susanne Proudlove. “Television news information gain from interesting video vs. talking heads”. In: *Journal of Broadcasting & Electronic Media* 25.1 (1981), pp. 15–24.

[17] Maksim Ekin Eren et al. “COVID-19 kaggle literature organization”. In: *Proceedings of the ACM Symposium on Document Engineering 2020*. 2020, pp. 1–4.

[18] Walter Gantz et al. “Gratifications and expectations associated with pop music among adolescents”. In: *Popular Music & Society* 6.1 (1978), pp. 81–89.

[19] Ayse Göker and Hans I Myrhaug. “User Context and Personalisation.” In: *Eccbr workshops*. Vol. 2002. 2002, pp. 1–7.

[20] Şule Gündüz and M Tamer Özsu. "A web page prediction model based on click-stream tree representation of user behavior". In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, pp. 535–540.

[21] Lei Han et al. "Modelling user behavior dynamics with embeddings". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 445–454.

[22] Christian Hansen et al. "Modelling sequential music track skips using a multi-rnn approach". In: *arXiv preprint arXiv:1903.08408* (2019).

[23] Christine H Hansen and Ranald D Hansen. "Music and music videos". In: *Media entertainment: The psychology of its appeal* (2000), pp. 175–196.

[24] Balázs Hidasi et al. "Session-based recommendations with recurrent neural networks". In: *arXiv preprint arXiv:1511.06939* (2015).

[25] Sepp Hochreiter. "The vanishing gradient problem during learning recurrent neural nets and problem solutions". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6.02 (1998), pp. 107–116.

[26] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[27] Eric Horvitz et al. "Models of attention in computing and communication: from principles to applications". In: *Communications of the ACM* 46.3 (2003), pp. 52–59.

[28] Ellen A Isaacs and John C Tang. "What video can and can't do for collaboration: a case study". In: *Proceedings of the first ACM International Conference on Multimedia*. 1993, pp. 199–206.

[29] Gawesh Jawaheer, Martin Szomszor, and Patty 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*. 2010, pp. 47–51.

[30] Robert Jourdain and Dinesh Mehta. *Music, the brain, and ecstasy: How music captures our imagination*. W. Morrow New York, 1997.

[31] Marius Kaminskas and Francesco Ricci. "Contextual music information retrieval and recommendation: State of the art and challenges". In: *Computer Science Review* 6.2-3 (2012), pp. 89–119.

[32] Marius Kaminskas and Francesco Ricci. "Location-adapted music recommendation using tags". In: *International conference on user modeling, adaptation, and personalization*. Springer. 2011, pp. 183–194.

[33] Ashish Kathuria et al. "Classifying the user intent of web queries using k-means clustering". In: *Internet Research* (2010).

[34] Juho Kim et al. "Understanding in-video dropouts and interaction peaks in online lecture videos". In: *Proceedings of the first ACM conference on Learning@ scale conference*. 2014, pp. 31–40.

[35] Silvia Knobloch and Dolf Zillmann. "Mood management via the digital jukebox". In: *Journal of communication* 52.2 (2002), pp. 351–366.

[36] Mirjam Köck and Alexandros Paramythis. "Activity sequence modelling and dynamic clustering for personalized e-learning". In: *User Modeling and User-Adapted Interaction* 21.1 (2011), pp. 51–97.

[37] Armin Kohlrausch and Steven van de Par. "Audio—visual interaction in the context of multi-media applications". In: *Communication acoustics* (2005), pp. 109–138.

[38] Panu Korpipää et al. "Bayesian approach to sensor-based context awareness". In: *Personal and Ubiquitous Computing* 7 (2003), pp. 113–124.

[39] Published by L. Ceci and Feb 7. *Most viewed YouTube videos worldwide 2023*. Feb. 2023. URL: `https://www.statista.com/statistics/249396/top-youtube-videos-views/`.

[40] Christoph Lagger, Mathias Lux, and Oge Marques. "What makes people watch online videos: An exploratory study". In: *Computers in Entertainment (CIE)* 15.2 (2017), pp. 1–31.

[41] Jey Lau and Timothy Baldwin. "An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation". In: July 2016, pp. 78–86. DOI: `10.18653/v1/W16-1609`.

[42] Quoc Le and Tomas Mikolov. “Distributed representations of sentences and documents”. In: *International conference on machine learning*. PMLR. 2014, pp. 1188–1196.

[43] Nan Li et al. “MOOC video interaction patterns: What do they tell us?” In: *European Conference on Technology Enhanced Learning*. Springer. 2015, pp. 197–210.

[44] Xiao Li, Ting Wang, and Huaimin Wang. “Exploring n-gram features in clickstream data for MOOC learning achievement prediction”. In: *Database Systems for Advanced Applications: DASFAA 2017 International Workshops: BDMS, BDQM, SeCoP, and DMMOOC, Suzhou, China, March 27-30, 2017, Proceedings 22*. Springer. 2017, pp. 328–339.

[45] T Warren Liao. “Clustering of time series data—a survey”. In: *Pattern recognition* 38.11 (2005), pp. 1857–1874.

[46] Ryan Lichtenwalter, Katerina Lichtenwalter, and Nitesh V Chawla. “Applying Learning Algorithms to Music Generation.” In: *IICAI*. Citeseer. 2009, pp. 483–502.

[47] Jimmy Lin and W John Wilbur. “Modeling actions of PubMed users with n-gram language models”. In: *Information retrieval* 12 (2009), pp. 487–503.

[48] Nathan N Liu et al. “Unifying explicit and implicit feedback for collaborative filtering”. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010, pp. 1445–1448.

[49] Adam J Lonsdale and Adrian C North. “Why do we listen to music? A uses and gratifications analysis”. In: *British journal of psychology* 102.1 (2011), pp. 108–134.

[50] Malte Ludewig and Dietmar Jannach. “Learning to rank hotels for search and recommendation from session-based interaction logs and meta data”. In: *Proceedings of the Workshop on ACM Recommender Systems Challenge*. 2019, pp. 1–5.

[51] Orii McDermott, Martin Orrell, and Hanne Mette Ridder. “The importance of music for people with dementia: the perspectives of people with dementia, family carers, staff and music therapists”. In: *Aging & mental health* 18.6 (2014), pp. 706–716.

[52] Francesco Meggetto et al. “On skipping behaviour types in music streaming sessions”. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021, pp. 3333–3337.

[53] Rishabh Mehrotra et al. “Jointly leveraging intent and interaction signals to predict user satisfaction with slate recommendations”. In: *The World Wide Web Conference*. 2019, pp. 1256–1267.

[54] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).

[55] Adrian C North and David J Hargreaves. “Music and adolescent identity”. In: *Music education research* 1.1 (1999), pp. 75–92.

[56] Elijah O Ode. “Impact of audio-visual (AVS) resources on teaching and learning in some selected private secondary schools in Makurdi”. In: *International journal of Research in humanities, arts and literature* 2.5 (2014), pp. 195–202.

[57] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. “Unsupervised learning of sentence embeddings using compositional n-gram features”. In: *arXiv preprint arXiv:1703.02507* (2017).

[58] Han-Saem Park, Ji-Oh Yoo, and Sung-Bae Cho. “A context-aware music recommendation system using fuzzy bayesian networks with utility theory”. In: *International conference on Fuzzy systems and knowledge discovery*. Springer. 2006, pp. 970–979.

[59] Stefan Paun. “Parallel text alignment and monolingual parallel corpus creation from philosophical texts for text simplification”. In: *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: student research workshop*. 2021, pp. 40–46.

[60] Van-Thuy Phi, Liu Chen, and Yu Hirate. “Distributed representation based recommender systems in e-commerce”. In: *DEIM Forum*. 2016.

[61] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. “Sequence-aware recommender systems”. In: *ACM Computing Surveys (CSUR)* 51.4 (2018), pp. 1–36.

[62] Karl Quinn. *At long last, Vevo is music to our ears*. Apr. 2012. URL: `https://www.smh.com.au/entertainment/music/at-long-last-vevo-is-music-to-our-ears-20120415-1x1o6.html`.

[63] Santosh Raju, Prasad Pingali, and Vasudeva Varma. "An unsupervised approach to product attribute extraction". In: *Advances in Information Retrieval: 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings 31*. Springer. 2009, pp. 796–800.

[64] Juan Ramos et al. "Using tf-idf to determine word relevance in document queries". In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. 1. Citeseer. 2003, pp. 29–48.

[65] Mohammed Abhur Razzaque, Simon Dobson, and Paddy Nixon. "Categorization and modelling of quality in context information". In: (2006).

[66] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[67] Peter J Rentfrow and Samuel D Gosling. "The do re mi's of everyday life: the structure and personality correlates of music preferences." In: *Journal of personality and social psychology* 84.6 (2003), p. 1236.

[68] Rebecca Robbins et al. "Examining sleep deficiency and disturbance and their risk for incident dementia and all-cause mortality in older adults across 5 years in the United States". In: *Aging (Albany NY)* 13.3 (2021), p. 3254.

[69] Rebecca B Rubin et al. "Media use and meaning of music video". In: *Journalism Quarterly* 63.2 (1986), pp. 353–359.

[70] Yoan Russac, Olivier Caelen, and Liyun He-Guelton. "Embeddings of categorical variables for sequential data in fraud context". In: *International conference on advanced machine learning technologies and applications*. Springer. 2018, pp. 542–552.

[71] Eldar Sadikov et al. "Clustering query refinements by user intent". In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 841–850.

[72] Danny Matthew Saputra, Daniel Saputra, and Liniyanti D Oswari. "Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method". In: *Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*. Atlantis Press. 2020, pp. 341–346.

[73] Markus Schedl, Arthur Flexer, and Julián Urbano. "The neglected user in music information retrieval research". In: *Journal of Intelligent Information Systems* 41.3 (2013), pp. 523–539.

[74] Bill Schilit, Norman Adams, and Roy Want. "Context-aware computing applications". In: *1994 first workshop on mobile computing systems and applications*. IEEE. 1994, pp. 85–90.

[75] Qiang Su and Lu Chen. "A method for discovering clusters of e-commerce interest patterns using click-stream data". In: *electronic commerce research and applications* 14.1 (2015), pp. 1–13.

[76] Yong Kiam Tan, Xinxing Xu, and Yong Liu. "Improved recurrent neural networks for session-based recommendations". In: *Proceedings of the 1st workshop on deep learning for recommender systems*. 2016, pp. 17–22.

[77] Robert F Tate. "Correlation between a discrete and a continuous variable. Point-biserial correlation". In: *The Annals of mathematical statistics* 25.3 (1954), pp. 603–607.

[78] Tippaya Thinsungnoena et al. "The clustering validity with silhouette and sum of squared errors". In: *learning* 3.7 (2015).

[79] I-Hsien Ting, Chris Kimble, and Daniel Kudenko. "UBB mining: finding unexpected browsing behaviour in clickstream data to improve a Web site's design". In: *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. IEEE. 2005, pp. 179–185.

[80] JW Uys, ND Du Preez, and EW Uys. "Leveraging unstructured information using topic modelling". In: *PICMET'08-2008 Portland International Conference on Management of Engineering & Technology*. IEEE. 2008, pp. 955–961.

[81] Carol Vernallis. *Experiencing music video: Aesthetics and cultural context*. Columbia University Press, 2004.

[82] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. “Order matters: Sequence to sequence for sets”. In: *arXiv preprint arXiv:1511.06391* (2015).

[83] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: `10.1038/s41592-019-0686-2`.

[84] Shoujin Wang et al. “Sequential recommender systems: challenges, progress and prospects”. In: *arXiv preprint arXiv:2001.04830* (2019).

[85] Hongyi Wen, Longqi Yang, and Deborah Estrin. “Leveraging post-click feedback for content recommendations”. In: *Proceedings of the 13th ACM Conference on Recommender Systems*. 2019, pp. 278–286.

[86] Gui-Rong Xue et al. “Optimizing web search using web click-through data”. In: *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. 2004, pp. 118–126.

[87] Diyi Yang et al. “Local implicit feedback mining for music recommendation”. In: *Proceedings of the sixth ACM conference on Recommender systems*. 2012, pp. 91–98.

[88] Xiao Yu et al. “Recommendation in heterogeneous information networks with implicit user feedback”. In: *Proceedings of the 7th ACM conference on Recommender systems*. 2013, pp. 347–350.

[89] Boxun Zhang et al. “Understanding user behavior in spotify”. In: *2013 Proceedings IEEE INFOCOM*. IEEE. 2013, pp. 220–224.

[90] Xiaohang Zhang et al. “Exploring the sequential usage patterns of mobile Internet services based on Markov models”. In: *Electronic Commerce Research and Applications* 17 (2016), pp. 1–11.

[91] Dávid Zibriczky et al. “Personalized recommendation of linear content on interactive TV platforms: beating the cold start and noisy implicit user feedback.” In: *UMAP workshops*. 2012.

A

ISMIR Submission

FROM CLICKS TO CUES: EXPLORING USER BEHAVIOUR WHILE WATCHING MUSIC VIDEOS, AS A LANGUAGE

First Author
Affiliation1
author1@ismir.edu

Second Author
Retain these fake authors in submission to preserve the formatting

Third Author
Affiliation3
author3@ismir.edu

ABSTRACT

Prior work has shown that user interactions with multimedia streaming platforms can be affected by the context within which the available content is consumed. Although studies have explored content consumption patterns related to music, film, or TV, little is currently understood about how users consume music videos. In this study, we address this gap in audiovisual content consumption by collecting and analysing a large dataset of music video streaming sessions from a music video streaming company [1]. Within our analysis of 1.8M sessions from around 270K unique users, we tackled the challenges of unequal session lengths through language models and clustered the resulting embeddings. We found that music video streaming sessions exhibit cohesive user interaction patterns, which can be clustered into distinct types. We found that certain user behaviours are associated with specific music video *genres* or the *context*, such as time of the day, in which users participate in a streaming session. Our insights expand the current understanding of user behaviour and their interactions with music videos online and shed light on factors that can affect it.

1. INTRODUCTION

For many decades, music videos have been widely enjoyed by people worldwide, while smart televisions and online streaming platforms have further facilitated access to high-quality content. Music videos have consistently been the most viewed videos on YouTube [2]. Despite their widespread popularity, research into the audio-visual aspects of music consumption has been lacking.

Research has focused on user interactions while listening to music [1, 2] or with videos outside the music domain [3, 4]. Music videos are a unique form of media that combine music, visual cues, and storytelling to create a multimedia experience that can significantly influence how audiences perceive and interact with music. These characteristics are unique compared to the audio-only music modality [5], indicating a need to better understand user behaviour while specifically streaming music videos.

One aspect of this understanding may stem from the ability to see user behaviour as action sequences of various lengths, allowing us to distinguish between patterns that occur at different stages within a streaming session (similarly to [6]). We also acknowledge that user action sequences can be represented in a way that resembles the structure of natural language, thus allowing us to explore user behaviour by employing corresponding techniques, a parallelism already leveraged in other domains [7–9].

With the goal of understanding users interacting with this unique modality of music, we address the following research questions:

- **[RQ1]**: How can we characterize the interaction patterns that emerge during user interactions with music videos through streaming?
- **[RQ2]**: To what extent do temporal and genre-related contextual features shape the interaction patterns?

In our work, we focus on analysing user behaviour data, as gathered on an interactive music video platform [3], to identify and interpret potentially distinct behavioural patterns during music video streaming sessions. We investigate the extent to which contextual aspects such as the time of the day, month, and music genre shape those behavioural patterns. Our results advance the current understanding of the distinct user behaviours that emerge while users interact with music videos. *E.g.,* in our work we were able to distinguish predominantly short/longer passive sessions from the ones which were more exploratory. Similarly, while pursuing the contextual analysis we found that the long passive sessions tend to generally start in the mornings. Our findings have important implications for the design and delivery of music videos on streaming platforms. We publicly share our gathered dataset to promote reproducibility and help future research in the field [4].

2. RELATED WORK

Studies have shown that it is crucial to model user interactions to comprehend user intent and provide more relevant search results [10, 11]. User interactions in terms

[1] Anonymized for the review process.

[2] https://www.statista.com/statistics/249396/top-youtube-videos-views/

[3] The platform name has been anonymised for the review process.

[4] The dataset will be released upon acceptance of this paper.

of clicks (*e.g.,* the number of pauses or forward seeks) have been utilised to model video click behaviours in the e-learning domain to understand student behaviour and cater to their learning needs [3, 4]. In the field of information retrieval, researchers have utilised clustering methods to model user interactions and identify user groups that share similar clickstream activities [12–14]. In the context of online crowd work, prior research has proposed using behavioural traces in worker modelling and pre-selection [15]. In the music domain, others have characterised different behaviours during listening sessions by interpreting the *skips* performed by the user [16]. However, such work has only focused on skips and does not consider other interaction signals such as *playlist change* and *likes*. Therefore, we identify and aim to address the research gap regarding studies focused on user behaviour analysis during music video streaming based on clickstream data.

Natural Language Processing (NLP) techniques have been actively used for unsupervised topic modelling, where linguistic documents are clustered to label similar topics [17–19]. We also found prior works where language models were applied to generate embeddings for non-lingual sequential data successfully. Russac et al. [8], used `word2vec` to create embeddings for sequential data in an unsupervised manner, to tackle credit card fraud detection. Previous works have also employed the `doc2vec` model proposed by Mikolov et al. [20] to model user behaviour in a session. Phi et al. and Ludewig and Jannach [7, 21] used `Doc2vec` to model user behaviour data in e-commerce and hotel rankings, respectively. Specifically, user actions were treated as words, and sequential actions in a session were interpreted as a sentence. Inspired by such prior work, in our study, we treat user behaviour as a sequence of within-session activities.

Users' behaviour on a music video streaming platform is expected to be influenced by their music preferences (*e.g.*, when *liking* or *skipping* videos). Literature has indicated that the music preferences of users can be influenced by their context [22–25]. The studies by Abowd et al. and Razzaque et al. [26, 27], have suggested different categories for such user contexts. However, we find the context classes proposed by Kaminskas and Ricci [28] to be better suited for our case, as their work is based on the music domain. They propose a schema of three main categories: 1) environment-related context, 2) user-related context, and 3) multimedia context. The environment-related context encapsulates user location, time-related information like time of the day, day of the week, and weather-related information. The user-related context contains information about the activity the user is performing, the emotional state of the user, and user demographics. Multimedia context refers to other multimedia sources (other than music) that users might be exposed to. Schedl *et al.* [29] also mention music context as an important factor that influences human music perception which for *e.g.,* encapsulates information on song lyrics, music video clips, and artists' background. Since the platform considered in our work does not retain any data regarding user demographics, and due to the opacity of information in real life regarding user activity, emotional state, and other multimedia platforms the user may be engaged with, we limit the focus of our study to the environmental (*temporal*) and music-related (*genre*) contexts.

3. METHODOLOGY

The focus of our study is multifaceted. We want to understand user behaviour in music video streams, analyse the behavioural patterns that emerge through their interaction with the online platform, and identify contextual factors that may affect their experience.

3.1 Data Gathering

We collected user activity per streaming session, as logged by an interactive music video streaming platform. Table 1 provides an overview of the session attributes we retrieved and processed in our study.

Attribute	Description
`session_id`	Identifier of streaming session on the platform
`video_id`	Identifier of music video played
`playlist_id`	Identifier of playlist curated by experts (optional)
`video_type`	Type of the video (music/ad)
`event`	The action performed by the user
`timestamp`	The local timestamp of the event's occurrence
`genre`	Genre of the specific music video

Table 1. Overview of the data schema used to create user action sequence representations and explore contexts.

Users can stream music videos on this platform, mainly through curated playlists, on their selected digital media players. These playlists tend to have cohesive content, and they auto-start once selected. Hence, a typical user flow is expected to contain content-based interactions with those playlists `select`, `change` and `search` alongside standard video interactions. Similar to other streaming services, ads are shown on this platform during the streaming sessions. We filtered out ads from session data as watching ads is not a user-initiated action through the session attribute `video_type`.

User Interaction Data. To study the behaviour of users during their streaming sessions, we utilised the '`event`' attribute from the session data (see Table 1). This attribute encapsulates all the possible interactions a user can perform. In our case, we made use of the following: 1) `complete video watched` when a music video has been played till its end, 2) `skip`, 3) `like`, 4) `unlike`, 5) playlist `change`, 6) content `search`, 7) `pause`, and 8) `unpause`. Though the pause and unpause features were released on the platform during the period of the study, we decided to still retain them in our data. The lack of information regarding their potential impact on behavioural clusters primarily drives the decision.

Contextual Data. To conduct a contextual analysis in the later parts of this study, `timestamp` has been used. Further, the `genre` attribute has been utilised to identify the genre of each music video. Further context elements are left for immediate future work.

3.2 Data Preparation

During our data preparation, there were several challenges due to the nature of real-time data acquisition. Thus, we put considerable effort into cleaning and standardising our session data.

Sessions as Activity Sequences. As every session contains a series of video plays with associated user actions, we represented user behaviour in every session as a sequence of user actions, grouped by associated videos. Since our current interest is purely in user action sequences characterising single sessions, we decided not to encode which exact video and playlist was played, and for how long. This resulted in each session being represented as a sequence of activity sequences, such as: *[[skip], [like, skip], [complete video watched], [pause, unpause, complete video watched]]*. Our approach aligns with previous practices in literature where the sequentiality in clicks has been described as an important aspect while characterising short-term session behaviour [30, 31].

Upon analysing our collected data, we observed that occasionally duplicates of the same action were logged on the same music video. This could have resulted from a slow internet connection, a false understanding of the interaction (*e.g.,* multiple presses on "like" do not yield different results), or miscommunications between the platform's asynchronous processes. To tackle this semantically uninformative user input, we only retained the first instance of such actions within each video activity sequence.

Session Embeddings. As expected, not all sessions have the same duration. This imposes a challenge when trying to compare their activity sequences, as their duration varies significantly, where a simple padding would severely affect computational performance.

As discussed previously, other studies having similar sequences of user behaviours, have employed language models to represent sequences, as embeddings. We believe that this is an elegant way to capture a user's interaction with the platform, both from an engineering perspective (language models handle sequenced tokens well), but also conceptually.

More specifically, we treated each user action as a letter, each activity sequence within a video as a word, and a session as a sentence. Our approach follows the methodology in [7, 21], who used a similar representation in their studies. Consequently, each value of the `event` attribute has been represented as follows in our generated action sequence: "c" for *playlist change*, "p" for *complete video watched*, "f" for *search (find)*, "s" for *skip*, "l" for *like*, "u" for *unlike*, "h" for *pause* and "k" *unpause*.

To convert session-based action sequences to a lower-level representation, we utilised the `Doc2Vec` model. `Doc2Vec` was chosen as it can encode documents/sentences into embeddings, by capturing the sequential order and context (the surrounding words). As discussed in Section 2, prior works have used the `Doc2Vec` model successfully for encoding similar behavioural sequences, providing evidence that it is a reliable and effective method for this analysis.

We used the Gensim implementation [32] of PV-DM (Distributed Memory version of Paragraph Vector) `Doc2Vec` to encode the sequences. PV-DM was chosen as opposed to Distributed Bag of Words Paragraph Vector as it better understands the semantics of words and also considers word order. The goal of this study is to find different behavioural patterns in a session. Therefore, users who exhibit similar action patterns during their sessions should intuitively be placed in close proximity to one another. This draws parallels with the text similarity task, which suggests that our `Doc2vec` embeddings can be trained using a similarity-based approach. As our study is completely based on unsupervised learning and has no labels, there is no definitive method to evaluate which hyperparameters perform better. This provides us with a good rationale to train our `Doc2vec` model with hyperparameters similar to [33], which have been found to be the most effective for document similarity works.

Consequently, in this study, the `Doc2vec` model was trained exclusively on the unique action sequences, taking into consideration the possibility that using only distinct sequences might produce a more accurate representation of all the sequences, by preventing the language model from overfitting on the most common sequences. After training the `Doc2Vec` model with hyperparameters found by Lau et al. [33] (which includes hyper-parameters such as number of epochs, embedding size, learning rate alpha), for each unique action sequence, a 300-dimensional document vector (doc-vector) was obtained.

Behavioural Clusters. To characterise the behaviours exhibited in each session, we clustered the inferred embeddings for all sessions, using the K-Means++ implementation from Scikit-learn [34].

The "Elbow method" and "Silhouette score" are the two most widely used methods to determine the number of clusters [35]. The Silhouette score is known to be suitable for clearly separable clusters [36]. However, we found that the Silhouette score was not an ideal measure of the number of clusters in our data. Our Silhouette scores ranged from 0.18 to 0.2 for different values of k, which suggested that the behavioural clusters in our data were not clearly separable (as the score was closer to 0 and farther from 1). Thus, we opted to use the Elbow method [37], acknowledging that its interpretation can be subjective since an elbow is not always visible, to determine the optimal number of clusters.

To visualise the clusters of session behaviour, deriving inspiration from Eren et al. [17], we projected the high-dimensional embeddings down to two dimensions using t-Distributed Stochastic Neighbour Embedding (t-SNE) [38].

3.3 Data Analysis

We evaluated our inferred clusters by running a correlation analysis between the different values of the `event` attribute of sessions (see Table 1) and the cluster category as the target variable. We used one hot encoding to retrieve boolean columns per cluster category. To that end, we used Point Biserial correlation, which is mathematically equivalent to Pearson's correlation, to retrieve the correlation between them (as implemented in SciPy [39]). We chose the Point Biserial correlation method because it is appropriate for evaluating the correlation between a continuous variable (number of each interaction per session) and a dichotomous variable (cluster labels) [40]. A t-test with n-1 degrees of freedom was used for statistical significance.

To further understand the user behaviour that each cluster represents, we analysed the user interactions during streaming sessions, within each cluster. Based on the different values for attribute `event` (e.g. play and skips), we identified the most prominent interactions per cluster and qualitatively inferred the cluster characteristics.

Finally, we wanted to understand the effect of the time a session took place and the prominent genre that was enjoyed during the session on user streaming behaviours. As the majority of the data in our study pertained to users from the USA, we have chosen to concentrate our contextual analysis specifically on this region. We selected only sessions longer than a minute to ensure that only meaningful user engagement was analysed. The sessions longer than 16 hours were also removed to emulate the maximum length a realistic streaming session would typically have (assuming the recommended 8 hours of sleep for the user [41]). Additionally, since the streaming service is facilitated via a digital media player, it is possible for the application to continue running in the background even after the user has, in effect, exited the platform. Thus, by eliminating sessions that exceed 16 hours in duration, we were able to maintain a higher data integrity.

We retrieved temporal information about the starting time of each session from the `timestamp` attribute, namely: a) *time of the day*, b) *day of the week*, and c) *month*. Regarding genre, we retrieved the most frequent genre per session, based on the music videos that were played during it. Once we retrieved the most frequent genre per session, we calculated the mode for each cluster to identify the genre that a behavioural cluster is most associated with, for our contextual analysis.

4. RESULTS

In total, we gathered 1.8 million sessions from ≈ 270.000 unique users, with the majority of them located in the United States. The sessions are from a time period of a single year (2022), covering all major holiday seasons (*e.g.,* Christmas, summer) where traffic and behaviour could differ compared to the rest of the year. The sessions are largely evenly distributed across the different times of day, with the exception of lower traffic between the hours of 11 pm and 4 am. Approximately one-fourth of the sessions occurred on weekends. As expected, we witnessed the highest traffic during the vacation/holiday period, specifically August and December, where $\approx$ 20% of the total sessions took place. Demographic information about users was not available to take into account for this research.

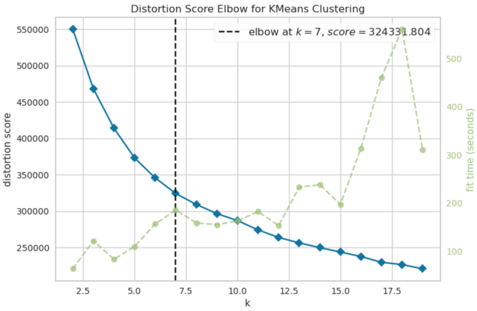

Figure 1. Optimal number of clusters, as determined using the Elbow method.

4.1 Clusters of Embedded Behavioural Sequences

Via the elbow method, we found the optimal number of clusters to be seven (see Figure 1). Within each cluster, we found that sessions have similar behavioural characteristics based on the actions user performed during the session.

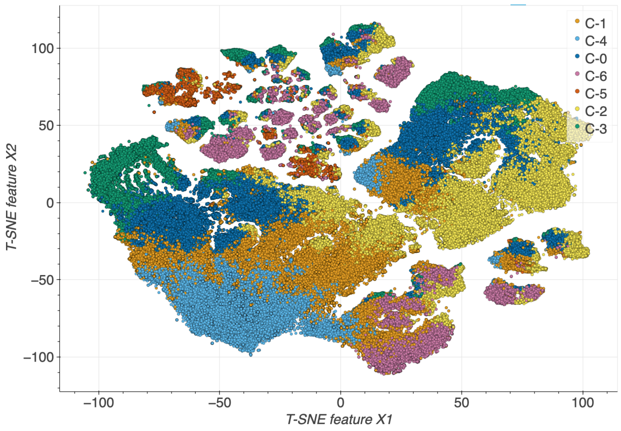

Figure 2. 'Cluster map' visualisation of embedded action sequences using t-SNE, where same-coloured dots represent a single behavioural cluster.

We analysed each cluster by examining the point biserial correlation between the user actions within the `event` attribute, e.g., *plays* and *likes*, and the cluster they were assigned. The p-value's corresponding to almost all of the point biserial correlation coefficients, associated with the different `event` attributes were found to be significant at the $p < 0.05$ level (cf. Table 2). It is worth noting that each cluster was distinctly correlated to different user actions, implying that our clustering method successfully revealed patterns within the user behaviour embeddings.

The retrieved clusters are distinctly correlated to different user interactions within a session. Our qualitative analysis of their properties expanded our understanding of the type of within-session behaviours they represent.

Additionally, the resulting two dimensions from t-SNE, although not interpretable by themselves, when used to visualise our clusters, clearly illustrate the relations of the clusters to each other. As shown in Figure 2, clusters,

Cluster	Plays	Likes	Skips	Playlist changes	Searches	Pauses	Unpauses
C0	**-0.12***	-0.01*	-0.05*	0.01*	-0.03*	0.00*	0.00*
C1	**-0.15***	-0.03*	-0.14*	-0.11*	-0.06*	-0.01*	-0.01*
C2	-0.1*	0.00	**0.44***	0.02*	-0.02*	0.00	0.00
C3	-0.16*	-0.01*	-0.07*	**0.24***	-0.03*	0.00*	0.00*
C4	**0.54***	-0.02*	-0.10*	-0.06*	-0.04*	0.00*	0.00*
C5	-0.05*	0.01*	0.02*	-0.03*	**-0.25***	0.05*	0.04*
C6	-0.03*	**0.13***	0.06*	0.00*	-0.02*	0.00	0.00

Table 2. Point Biserial correlation and p-values for each cluster, where statistical significance ($p < 0.05$) is marked using an asterisk (*). Bold values indicate action types with the highest correlations.

represented by coloured dots, are spread across the graph (due to projecting the high-dimensional vectors to a plane). Still, we find that each cluster neighbours only a specific set of other clusters. Qualitatively, this indicates that the generated embeddings are informative and capture different behaviours.

In Table 3, we see that sessions within C0, C1, C2, and C5 have similar averages for videos played to completion, with C3 having far fewer and C4 many more. In the case of C0, C1, C2, and C3, these clusters seem to neighbour each other closely, as seen in Figure 2. Sessions within these clusters also exhibit a high average number of playlist changes (2.9, 1.3, 3.3, and 14, respectively), albeit with different patterns on the rest of the interactions.

Through our clustering approach, we managed to identify a group of behavioural patterns that centre around more skips than others, as indicated by C2. This characteristic could be why those sessions neighbour others from C6, which yields the second-highest average number of skips. While both seem to have "skips" as a common interaction, their average number of video completions largely differ (sessions in C6 have almost double the number of video completions in C2). Also, users during sessions like those in C6 seem to most actively "like" videos compared to all other sessions, showing a more selective experience.

We also found a group of sessions characterised by seemingly passive user behaviour (see C4). Figure 2 shows these sessions almost exclusively neighbour those of C1. On average, both C4 and C1 exhibit low user activity, indicating a preference for "sitting back and watching" content rather than interacting with the platform.

The highest average count of searches in C5, compared to other clusters, is potentially indicative of a greater degree of purposeful and targeted listening. Although few sessions exhibit such behaviour, they seem to neighbour most other clusters, but primarily those from C2, C3, C6, and to a much lesser degree, those from C1 and C4. This can be attributed to C2, C3 and C6 showcasing selective behaviour (highest average on "skip", "playlist change" and "likes"), which can be associated with the targeted experience that searching content can provide.

4.2 Session Contexts and Users' Behaviour

In our contextual analysis of the behavioural clusters, we wanted to focus on the sessions from USA-based users whose sessions comply with our duration conditions. After filtering the data, we retained 1.58M sessions in total.

Upon examining the clusters with respect to environmental context, we indeed observe that long passive listening sessions (C4) typically start in the morning around 10 am, as illustrated in Figure 3. The clear difference between the start of those sessions compare to those of the other clusters, indicates that those users tend to have more interactive sessions as the day progresses. This is true even for the shorter passive sessions (C1), which although peak around 17:00, they still have the second highest occurrence until 16:00.

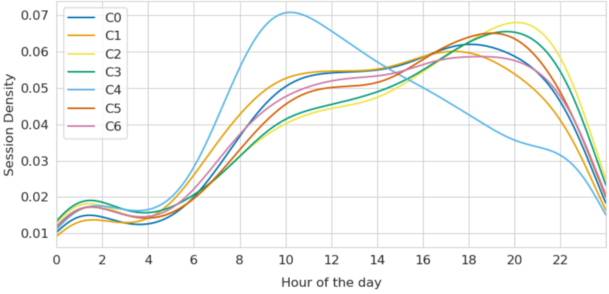

Figure 3. Session start density in relation to time of the day

As we see in Table 4, clusters associated with more active behaviours (C0, C2, C3, C6) peak on August, while the more passive behavioural clusters (C1 and C4) peak on December. One reasonable explanation regarding the peak on December, would be that the month is mainly associated with the holiday season in the USA; users might be more inclined to start a session and consume the music videos more passively, compared to other periods.

August though seems harder to explain with just the gathered data. Seeking further insights from the streaming company, we identified a marketing campaign during that month, which promoted the interactive features of the platform. We believe that this could explain the peak of the more "interactive" sessions that period. Although further inspection on a wider span of time could help identify if other factors could affect such phenomenon.

Upon observing the relation between behaviour and music video genres, we found that "Pop", "Country" and "Rap/Hip-Hop" are the most popular genres across the behavioural clusters. "Pop" and "Rap/Hip-Hop" are associated more with active and explorative sessions (clusters C0, C2, C3, C5, C6). "Pop" appears to be more associated with relative short sessions, while "Rap/Hip-Hop" to those that exhibit more selective behaviour (more "skips", "likes" and "searches").

The top genre for sessions in C1 and C4 on the other hand was "Country". Both clusters are identified as more "passive" as they exhibit considerably less interactive behaviour compared to the other clusters, differing on the length of the sessions.

Cluster	Video Completion	Likes	Skips	Playlist change	Search	Pause	Unpause	Session Characteristics
C0	27	0.16	0.9	2.9	0.15	0	0	Short sessions, plays, few channel changes
C1	36	0.06	0.4	1.3	0	0	0	Shorter, passive listening sessions
C2	22	0.3	**9.6**	3.3	0.01	0	0	Exploration sessions
C3	9.9	0.2	0.8	**14**	0.01	0	0	Playlist juggling sessions
C4	**175**	0.6	0.26	1.4	0	0	0	Long, passive listening sessions
C5	28	1.4	2.2	1.4	**3**	0.1	0.1	Targeted listening sessions
C6	43.7	**7.8**	3.5	2.5	0.1	0	0	Selective listening sessions

Table 3. Mean values of user actions per behaviour cluster; the highest value per action is highlighted in bold.

Cluster	Month	Genre
C0	Aug	Pop
C1	Dec	Country
C2	Aug	Rap/Hip-Hop
C3	Aug	Pop
C4	Dec	Country
C5	Jan	Rap/Hip-Hop
C6	Aug	Rap/Hip-Hop

Table 4. Most active months and most popular genres per behavioural cluster.

5. DISCUSSION

With this work, we sought to identify distinct patterns of user interaction with streaming music videos and to gain a deeper understanding of what these patterns mean. We found distinguishable behaviour patterns across user sessions through a structured analysis of user interactions, represented as embeddings. While some of these patterns overlap, as visualised in Figure 2, there are also clear differences. Clusters C2 and C3 share the attribute of involving many playlist changes, potentially indicating more exploratory user behaviour. On the other hand, sessions belonging to C4 indicate more passive listening periods that peak around 10 am. This could indicate a need to explore whether longer playlists should be created to support such user behaviour.

Further, the characterized behaviours also tie with the prior works, it has been acknowledged that people often play music and music videos as a background [42, 43] which might explain behavioural clusters C0, C1 and C4. It has also been found that people often listen to music for mood management [42, 44], this might trigger more targeted and selective listening sessions i.e behavioural clusters C5 and C6. Lonsdale *et al.* [42] also established that people listen to music for the purpose of exploration and surveillance. This might be a plausible reason behind the behavioural clusters C2 and C3.

Having found identifiable patterns of interaction, we delved further deeper into whether such behaviour can be associated with more contextual aspects, such as the time of day or the genre of music listened to. Interestingly, certain patterns occur more frequently at specific times of the year, *e.g.,* C1 and C4 happening most frequently in December. Insights like these can have lasting effects on music video streaming platforms. Such platforms can utilise similar insights from their session data to inform design decisions and exploration regarding user interfaces, user services, and elements of user experience. Our analysis, therefore, reveals potentials for adaptation and personalisation based on user behaviour patterns. For instance, behaviour patterns similar to C5, *i.e.,* targeted listening sessions, could indicate a need for well-organised and curated catalogues or libraries of music videos. Each of the individual behaviour patterns can be, in part, a set of features to power a recommender system tuned to the domain of music videos. Finally, the user behaviours revealed in this work can provide a starting point for modelling users that emerging music platforms should be prepared to support.

Limitations. At the current stage our the research, we have only carried out a visual, qualitative analysis of where certain action patterns tend to occur within sessions, which has not led to conclusive results yet. Also, the content and curated playlists of our gathered data might inspire interaction behaviours specific to the platform and, to an extent, to the demographics of our sample. Considering the lack of similar studies, our insights cannot yet generalise to streaming sessions beyond the selected platform. Therefore, we encourage further research in the domain of music video streaming so that a more generalised understanding on the user interaction behaviours can be achieved.

6. CONCLUSIONS AND FUTURE WORK

In this work we identified the patterns in user behaviour by characterising the sequences of user interactions during music video streaming sessions. Additionally, we examined these behavioural patterns in relation to contextual features to gain a better understanding of the exhibited behaviours. We believe our findings provide valuable insights into the user interaction behaviours which can be used by music streaming platforms to personalise and enhance the overall user experience.

In the imminent future, we will delve deeper into the analysis of within-session action sequence patterns, and investigate user behaviour across sessions over time, and explore the relationship between exhibited behaviour and content types further (*e.g.,* understanding which behavioural patterns correspond to specific kinds of playlists). We will also further utilise the encoded behavioural patterns to predict user behaviour within sessions as well as upcoming ones.

7. REFERENCES

[1] A. Abdul, J. Chen, H.-Y. Liao, and S.-H. Chang, "An emotion-aware personalized music recommendation system using a convolutional neural networks approach," *Applied Sciences*, vol. 8, no. 7, p. 1103, 2018.

[2] L. Aguiar and B. Martens, "Digital music consumption on the internet: Evidence from clickstream data," *Information Economics and Policy*, vol. 34, pp. 27–43, 2016.

[3] N. Li, Ł. Kidziński, P. Jermann, and P. Dillenbourg, "Mooc video interaction patterns: What do they tell us?" in *European Conference on Technology Enhanced Learning*. Springer, 2015, pp. 197–210.

[4] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller, "Understanding in-video dropouts and interaction peaks in online lecture videos," in *Proceedings of the first ACM conference on Learning@ scale conference*, 2014, pp. 31–40.

[5] Z. Duan, S. Essid, C. C. Liem, G. Richard, and G. Sharma, "Audiovisual analysis of music performances: Overview of an emerging field," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 63–73, 2018.

[6] 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*, ser. CHIIR '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 95–106. [Online]. Available: https://doi.org/10.1145/3576840.3578312

[7] M. Ludewig and D. Jannach, "Learning to rank hotels for search and recommendation from session-based interaction logs and meta data," in *Proceedings of the Workshop on ACM Recommender Systems Challenge*, 2019, pp. 1–5.

[8] Y. Russac, O. Caelen, and L. He-Guelton, "Embeddings of categorical variables for sequential data in fraud context," in *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)*. Springer, 2018, pp. 542–552.

[9] G. Reddy, L. Desban, H. Tanaka, J. Roussel, O. Mirat, and C. Wyart, "A lexical approach for identifying behavioural action sequences," *PLOS Computational Biology*, vol. 18, no. 1, p. e1009672, 2022.

[10] G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi, and W. Fan, "Optimizing web search using web click-through data," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 2004, pp. 118–126.

[11] A. Kathuria, B. J. Jansen, C. Hafernik, and A. Spink, "Classifying the user intent of web queries using k-means clustering," *Internet Research*, 2010.

[12] Ş. Gündüz and M. T. Özsu, "A web page prediction model based on click-stream tree representation of user behavior," in *Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining*, 2003, pp. 535–540.

[13] Q. Su and L. Chen, "A method for discovering clusters of e-commerce interest patterns using click-stream data," *electronic commerce research and applications*, vol. 14, no. 1, pp. 1–13, 2015.

[14] I.-H. Ting, C. Kimble, and D. Kudenko, "Ubb mining: finding unexpected browsing behaviour in clickstream data to improve a web site's design," in *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. IEEE, 2005, pp. 179–185.

[15] U. Gadiraju, G. Demartini, R. Kawase, and S. Dietze, "Crowd anatomy beyond the good and bad: Behavioral traces for crowd worker modeling and pre-selection," *Computer Supported Cooperative Work (CSCW)*, vol. 28, pp. 815–841, 2019.

[16] 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*, 2021, pp. 3333–3337.

[17] M. E. Eren, N. Solovyev, E. Raff, C. Nicholas, and B. Johnson, "Covid-19 kaggle literature organization," in *Proceedings of the ACM Symposium on Document Engineering 2020*, 2020, pp. 1–4.

[18] A. Budiarto, R. Rahutomo, H. N. Putra, T. W. Cenggoro, M. F. Kacamarga, and B. Pardamean, "Unsupervised news topic modelling with doc2vec and spherical clustering," *Procedia Computer Science*, vol. 179, pp. 40–46, 2021.

[19] J. Uys, N. Du Preez, and E. Uys, "Leveraging unstructured information using topic modelling," in *PICMET'08-2008 Portland International Conference on Management of Engineering & Technology*. IEEE, 2008, pp. 955–961.

[20] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.

[21] V.-T. Phi, L. Chen, and Y. Hirate, "Distributed representation based recommender systems in e-commerce," in *DEIM Forum*, 2016.

[22] E. Horvitz, C. Kadie, T. Paek, and D. Hovel, "Models of attention in computing and communication: from principles to applications," *Communications of the ACM*, vol. 46, no. 3, pp. 52–59, 2003.

[23] P. Korpipää, M. Koskinen, J. Peltola, S.-M. Mäkelä, and T. Seppänen, "Bayesian approach to sensor-based context awareness," *Personal and Ubiquitous Computing*, vol. 7, pp. 113–124, 2003.

[24] M. Kaminskas and F. Ricci, "Location-adapted music recommendation using tags," in *International conference on user modeling, adaptation, and personalization*. Springer, 2011, pp. 183–194.

[25] H.-S. Park, J.-O. Yoo, and S.-B. Cho, "A context-aware music recommendation system using fuzzy bayesian networks with utility theory," in *International conference on Fuzzy systems and knowledge discovery*. Springer, 2006, pp. 970–979.

[26] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles, "Towards a better understanding of context and context-awareness," in *International symposium on handheld and ubiquitous computing*. Springer, 1999, pp. 304–307.

[27] M. A. Razzaque, S. Dobson, and P. Nixon, "Categorisation and modelling of quality in context information," in *Workshop on AI and Autonomic Communications, held at International Joint Conference on Artificial Intelligence (IJCAI)*, 2005, p. EJ.

[28] M. Kaminskas and F. Ricci, "Contextual music information retrieval and recommendation: State of the art and challenges," *Computer Science Review*, vol. 6, no. 2-3, pp. 89–119, 2012.

[29] M. Schedl, A. Flexer, and J. Urbano, "The neglected user in music information retrieval research," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 523–539, 2013.

[30] M. Quadrana, P. Cremonesi, and D. Jannach, "Sequence-aware recommender systems," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–36, 2018.

[31] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," *arXiv preprint arXiv:1511.06939*, 2015.

[32] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[33] J. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," 07 2016, pp. 78–86.

[34] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[35] D. M. Saputra, D. Saputra, and L. D. Oswari, "Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method," in *Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*. Atlantis Press, 2020, pp. 341–346.

[36] T. Thinsungnoena, N. Kaoungkub, P. Durongdumronchaib, K. Kerdprasopb, N. Kerdprasopb *et al.*, "The clustering validity with silhouette and sum of squared errors," *learning*, vol. 3, no. 7, 2015.

[37] B. Bengfort, R. Bilbro, N. Danielsen, L. Gray, K. McIntyre, P. Roman, Z. Poh *et al.*, "Yellowbrick," 2018. [Online]. Available: http://www.scikit-yb.org/en/latest/

[38] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[39] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[40] R. F. Tate, "Correlation between a discrete and a continuous variable. point-biserial correlation," *The Annals of mathematical statistics*, vol. 25, no. 3, pp. 603–607, 1954.

[41] Y. Liu, A. G. Wheaton, D. P. Chapman, T. J. Cunningham, H. Lu, and J. B. Croft, "Prevalence of healthy sleep duration among adults—united states, 2014," *Morbidity and Mortality Weekly Report*, vol. 65, no. 6, pp. 137–141, 2016.

[42] A. J. Lonsdale and A. C. North, "Why do we listen to music? a uses and gratifications analysis," *British journal of psychology*, vol. 102, no. 1, pp. 108–134, 2011.

[43] C. Lagger, M. Lux, and O. Marques, "What makes people watch online videos: An exploratory study," *Computers in Entertainment (CIE)*, vol. 15, no. 2, pp. 1–31, 2017.

[44] S. Knobloch and D. Zillmann, "Mood management via the digital jukebox," *Journal of communication*, vol. 52, no. 2, pp. 351–366, 2002.