Alavijeh et al. *EPJ Data Science* (2025) 14:29
https://doi.org/10.1140/epjds/s13688-025-00549-7

EPJ Data Science
a SpringerOpen Journal

RESEARCH Open Access

Twitter-MusicPD: melody of minds - navigating user-level data on multiple mental health disorders and music preferences

Soroush Zamani Alavijeh[1], Xingwei Yang[2*], Zeinab Noorian[2], Amira Ghenai[2] and Fattane Zarrinkalam[3]

*Correspondence:
nancy.yang@torontomu.ca
[2]Ted Rogers School of Information Management, Toronto Metropolitan University, Toronto, Canada
Full list of author information is available at the end of the article

Abstract

Social media platforms have become integral spaces for individuals to express emotions, seek advice, and disclose mental health conditions. While existing research primarily focuses on analyzing textual content for predicting mental disorders, music listening, as a fundamental aspect of human experience, has gained attention for its potential to influence psychological well-being. This paper introduces the Twitter-Music-Psychological Disorder (Twitter-MusicPD) dataset, which includes data from 5767 music-listening Twitter users, covering both individuals with six self-reported psychological disorders and non-disordered users, along with a matched control group of 38,086 non-music-listening Twitter users across six disordered and non-disordered groups. The dataset spans from August 2007 to May 2022, comprising 8,976,628 English tweets reported as embeddings and the content of 78,413 music tracks shared by users. Detailed information on music tracks, including sources, titles, artists and associated lyrics, is provided, along with sentiments and emotions related to the music. Twitter-MusicPD serves as a comprehensive resource for investigating the relationships between Twitter engagement, music choices, and psychological well-being, offering insights into how tweeting behaviors and music preferences evolve over time. Our data is available at: https://github.com/szamani20/Twitter-MusicPD_Melody-of-Minds.

Keywords: Psychological Disorders; Music Preference; Social Media; Twitter

1 Introduction

Social networks have become a prominent platform where people express emotions, seek advice, share personal experiences, and even disclose their mental health conditions in a virtual setting [1]. Given the wealth of information users share, extensive research has explored how language patterns in social media posts, such as vocabulary selection, emotional expressions, and psycholinguistic attributes, can be linked to various mental health conditions. These studies often focus on predicting disorders such as depression [2], suicidal thoughts [3], post-traumatic stress disorder (PTSD) [4], and bipolar disorder [5] based on textual content.

In parallel with the analysis of language patterns, music is also recognized as a powerful medium that can reflect and influence human emotions [6]. Music has been shown to impact cognitive, emotional, and physiological parameters, including mood regulation, pain perception, and relaxation [7, 8]. In recent decades, a growing body of research has examined music as a cognitive phenomenon. Studies have explored key elements such as pitch, tempo, rhythm, and melodic contour, as well as music's impact on psycho-physiological parameters, influencing pain perception, relaxation, blood pressure, respiratory patterns, and heart rate dynamics [7]. With its ability to evoke deep emotions, music is frequently employed for emotion regulation [9, 10], making it a promising intervention for addressing cognitive, emotional, and social challenges associated with psychiatric conditions. Research on music therapy has shown improvements in mood and psychopathology among individuals with depression and schizophrenia-like disorders [11, 12], with high acceptance and participation of patients.

The integration of online music streaming platforms into social media has provided an additional channel through which individuals express their moods and emotions by sharing music [13]. This has led to research on the analysis of music sentiment, the relationship between music preferences and personality traits, and the intersection of mental health conditions with music listening behaviors [14–17]. However, most existing studies are limited to small-scale clinical trials or traditional cohort studies using self-reported data [9, 18–20]. There is a lack of large-scale datasets that examine the combination of music preferences, tweeting activities, and social behaviors, highlighting the need for broader research in this area.

To address these gaps, we present the Twitter-Music-Psychological Disorder (Twitter-MusicPD) dataset, a large-scale, comprehensive resource for investigating the relationships between Twitter engagement, music preferences, and psychological well-being. The dataset spans from August 2007 to May 2022, containing the profiles and temporal tweet behaviors of *5767* music-listening Twitter users, including those with six self-reported psychological disorders (i.e., depression, bipolar disorder, anxiety, panic disorder, PTSD, and borderline personality disorder) and non-disordered users, along with a group of *38,086* non-music-listening Twitter users (we refer to this group as the control group). The control group includes over *8.9 million* English tweets, *78,413* music tracks, and the corresponding lyrics of these tracks. It also captures over *35.8 million* English tweets, providing a potential for a robust comparison to explore differences in user behavior. To comply with Twitter's API terms, we share tweet embeddings rather than raw tweet text, allowing for richer linguistic and behavioral analysis while maintaining privacy.

This dataset offers several key features:

- Detailed information on music tracks shared by users, including the source, title, artist, and lyrics. Additionally, sentiments and emotions related to the music tracks are extracted, allowing an in-depth analysis of how music preferences intersect with online behaviors.
- Temporal data spanning over a decade, enabling researchers to study the evolution of user behavior, tweet content, and music preferences over time.
- A comprehensive set of user embeddings, facilitating advanced analyses of social media engagement and the potential influence of music on the psychological well-being.

In addition to releasing the dataset, we performed exploratory data analysis (EDA), revealing key insights into the differences between music-listening and non-music-listening groups. The EDA highlights social media engagement metrics, linguistic patterns, and emotional expression, showing how music influences user interactions and mental health expressions online. We anticipate that the Twitter-MusicPD dataset will be a valuable resource for researchers in mental health, computational social science, and natural language processing. It will provide a foundation for exploring the complex relationships between social media behavior, music engagement, and psychological health.

2 Literature review

Recently, there has been a growing interest in analyzing mental health issues using social media platforms. Researchers have tried to deduce users' mental states by examining their online behaviors and linguistic patterns in their online posts. Diverse techniques have been used, including various machine learning algorithms, the examination of different features, the consideration of various online interactions, and the sourcing of data from specific social media platforms.

Specifically, researchers in computer science explore features derived from the posting and behavioral history of social media platforms such as Twitter, Reddit, and Facebook to identify indicators of mental disorders and develop automated detection systems for various mental disorders such as depression [21, 22], anxiety and Obsessive-Compulsive Disorder (OCD) [23], bipolar disorder [24], and PTSD [25].

As social media becomes a prevalent medium to reflect the mood and behavior of users, researchers are more incentivized to construct large-scale, well-labeled datasets with different mental disorders to enable thorough analysis of users' chronological activities, social interactions, and online behaviors. The initial effort by Coppersmith et al. [26] led to the creation of a dataset featuring four distinct mental disorders incorporating a limited selection of users' recent tweets without considering the progression of their mental health conditions. Subsequent research by Shen et al. [27] resulted in the formation of a more extensive dataset of over 3000 users who self-reported depression, including details such as profile statistics, tweet history, and patterns of social engagement and activities over time. Cohen et al. [28] were inspired by this approach. They created a large-scale dataset called SMHD (Self-reported Mental Health Diagnoses), including nine mental health conditions and their respective matching groups from Reddit. This dataset is a valuable resource for research to identify users with various psychological disorders by analyzing their language use. Suhavi et al. [29] created Twitter-STMHD (Self-Reported Temporally-Contextual Mental Health Diagnosis Dataset). This user-level dataset categorizes eight types of mental disorders along with a control group consisting of a profile of 54,006 users. This dataset includes the temporal context of user activities and timelines, selected to reflect the onset and progression of mental health conditions. As another labeled dataset with multiple mental disorders, Ji et al. [30] published a large-scale dataset from Reddit which contains 54,412 data instances collected from several subreddits related to suicide, depression, anxiety, and bipolar disorder.

Recently, Villa-Pérez et al. [31] published a dataset for mental health detection from Twitter which contains data from approximately 1500 Twitter users diagnosed with nine different mental disorders (ADHD, Autism, Anxiety, Bipolar, Depression, Eating disorders, OCD, PTSD, and Schizophrenia) and 1700 matched control users. The dataset supports both binary and multi-class classification tasks. Additionally, Raihan et al. [32] have published MentalHelp, a large-scale semi-supervised dataset containing over 14 million instances collected from Reddit. It was labeled using an ensemble of three models: flan-T5, Disor-BERT, and Mental-BERT. This dataset aims to facilitate training systems capable of modeling different mental disorders.

All the aforementioned datasets, alongside a substantial body of research on mental health analysis, affirm the richness and potential of social media data for investigating user-level behaviors to detect and predict psychological disorders and their symptoms early from social media. However, most datasets only provide post text, leading to a heavy reliance on textual features of user-generated content. As a result, existing approaches primarily focus on extracting linguistic cues such as cognitive, emotional, and sentiment markers to infer users' psychological states [33].

Music can potentially provoke deep emotions and is commonly utilized for regulating feelings in listeners [9]. The relationship between mental health and how people interact with music, including their listening habits, choices, and requirements, is increasingly becoming a focus of research [34, 35]. Previous studies explored how individuals with different mental health conditions employ music to manage their emotions [10]. It has been shown that emotional dependency on music increases during episodes of depression and psychological distress [36]. Research further indicates that the emotional experiences derived from listening to music do not always result in positive health outcomes and, instead, may intensify the pathological symptoms of the listeners with mental health problems. In [37], authors confirmed the dual-edged power of music by discovering that people who listen to music while having high levels of distress would feel more intense and experience negative moods afterward. Additional studies have assessed how choices in music listening affect the mood and overall well-being of individuals with depression [38, 39], suggesting that depressed individuals might engage with music as a form of unhealthy coping mechanism, leading to rumination and social withdrawal [18]. Garrido and Schubert [40] discovered that people with depression tend to choose sad music, adversely affecting their emotional state. This aligns with findings [41] that those suffering from depression struggle to select music that could improve their mood. Meanwhile, another research avenue has explored music therapy's efficacy in treating various mental disorders [6, 42]. In [19], authors found significant improvement in depression symptoms, sleep quality, quality of life, and anhedonia among participants of music-based intervention. Music therapy has also been shown to effectively lessen anxiety and depression levels in individuals diagnosed with General Anxiety Disorder (GAD) [43]. Nevertheless, the majority of existing studies have been centered around patients in clinical environments [18, 19] or carried out via conventional observational methods that utilize questionnaires and self-reported surveys [9, 20]. Although these investigations have provided valuable insights, they often focus on particular mental health conditions, predominantly depression and its related symptoms, and include only a small group of participants.

In a recent study, [16] explored the connection between users' self-reported mental health disorders on social media and their musical preferences through a comprehensive analysis of data collected from Twitter. They examined the linguistic characteristics of the music listened to by individuals with various mental health conditions, comparing

Table 1 Prominent benchmark datasets in mental health without music research

Dataset	Reference	Category	Source	Instances	Control Group?	Music info
Coppersmith et al.	Coppersmith et al. (2015) [26]	4 mental disorders	Twitter	28,832	yes	N/A
Shen et al.	Shen et al. (2018) [27]	Depression	Twitter	3000+	yes	N/A
SMHD	Cohen et al. (2019) [28]	9 mental disorders	Reddit	36,948+	yes	N/A
STMHD	Suhavi et al. (2021) [29]	8 mental disorders	Twitter	54,006	yes	N/A
Villa-Perez et al.	Villa-Perez (2022) [31]	9 mental disorders	Reddit	3200	yes	N/A
SWMH	Ji et al. (2022) [30]	5 mental disorders	Twitter	54,412	no	N/A
MentalHelp	Raihan et al. (2023) [32]	7+ mental disorders	Reddit	14 million	no	N/A
Music-MentalHealth	Zamani et al. (2023) [16]	6 mental disorders	Twitter	3999	no	59,468
Twitter-MusicPD	This work (2025)	6 mental disorders	Twitter	43,853	yes	78,413

these patterns to a non-disordered group without any psychological conditions. While their findings highlight significant differences in language and sentiment between the disorder and control groups, the dataset has certain limitations. Notably, their dataset did not include a control group of users who did not listen to music. This lack of a non-music-listening control group prevents comparison between music listeners and non-listeners within both disorder and non-disordered populations. Additionally, their work does not support in-depth longitudinal or temporal analysis, which would allow researchers to track shifts in music-listening habits and psychological states over time. As a result, the dataset cannot capture the evolution of users' music interactions and mental health conditions over a prolonged period.

To address the limitations of previous studies, our Twitter-MusicPD dataset offers a more comprehensive examination of user behaviors, including both music-listening and non-music-listening groups for users with and without psychological disorders. This enables a more detailed analysis by allowing us to apply matching techniques to evaluate the causal impact of music on tweet behaviors, emotions, and psychological states. Our dataset supports multifaceted feature extraction, incorporating tweet content and music-listening data while providing an extensive timeline for longitudinal and temporal analysis. This allows for effectively tracking trends in music preferences and social media activity, offering deeper insights into how music influences mental health over time.

Table 1 summarizes key benchmark datasets in mental health research. The *Instances* column shows the total number of users included in each dataset. The *control group?* column indicates whether the dataset includes a control group that can be used for comparative analysis. The *Music info* column specifies whether the dataset includes information about music listening habits.

3 Data collection

This section describes the construction and characteristics of the Twitter-Music-Psychological Disorder (Twitter-MusicPD) dataset. The dataset includes six mental disordered groups corresponding to branches in the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition) [44]: Depression, Anxiety, PTSD, Bipolar, Borderline, and Panic. Additionally, we include users without any of the studied mental health conditions. A unique aspect of our dataset is that all users, both disordered and non-disordered, have a record of music listening in their tweets, allowing us to explore the relationship between music and mental health.

Table 2 The statements used for collecting tweets. X: mental disorder name, Y: date specifier

Statements of Diagnosis Tweets
Y i was diagnosed with X
Y i was clinically diagnosed with X
i'm diagnosed with X Y
i'm clinically diagnosed with X Y
i am diagnosed with X Y
i am clinically diagnosed with X Y
diagnosed me with X Y
Y i have been diagnosed with X
Y i have been clinically diagnosed with X
i've been diagnosed with X Y
i've been clinically diagnosed with X Y

To provide a robust comparison, we also include a user group composed of users—both disordered and non-disordered—who have no music-related content in their timelines. Including such a user group enables studies that distinguish the influence of music listening on mental health from other factors, as reflected in social media activity.

3.1 User group construction

Music-Listening Group This group consists of Twitter users who either self-reported a diagnosis (i.e., one of the six psychological disorders) or are non-disordered but have engaged in music listening, as indicated by their tweets. To identify the disordered group, we processed tweets that explicitly mention a diagnosis using high-precision patterns, as shown in Table 2. Regular expressions were employed to capture tweets containing these patterns, which were then used as queries to the Twitter API. For example, we considered tweets such as "Y I was diagnosed with X," "I was diagnosed with X Y," and "diagnosed me with X Y," where X represents a psychological disorder and Y represents a date specifier (e.g., today, yesterday, this week).This method aligns with the dataset construction practices used in previous studies [28, 29], which are well-established methods that serve as a validated proxy for identifying mental health conditions on social media.

In this study, we considered tweets in which users explicitly self-report a specific mental health diagnosis (e.g., "I was diagnosed with depression") as anchor tweets. These tweets were collected over a two-year period, from January 2020 to February 2022. Two students independently annotated 17,370 anchor tweets, achieving a Cohen's Kappa score of 0.99, indicating near-perfect agreement. In cases where disagreements arose, a third annotator was introduced to resolve conflicts. The inter-annotator agreement among all three annotators was measured using Fleiss' Kappa, which resulted in a score of 0.90, reflecting high reliability. When disagreement persisted, the final label was determined by majority vote. We then extracted the most recent tweets of the remaining users (up to 3200 tweets per user which dated back to August 2007.[1]), resulting in a total of 13,154 anchor tweets from 11,652 users in the disordered group, with 16,339,519 timeline tweets extracted.

To form a non-disordered group, we collected users who are unlikely to have any of the studied mental conditions. We ensured temporal consistency by sampling tweets randomly from the same period as the disordered group (January 2020 to February 2022). We further collected 3200 of their most recent tweets from their timelines. We carefully

[1] This cap is not strictly enforced by Twitter API and has a less than 3% slippage.

Table 3 Summary of the music-listening group, categorized by disorder Type

Disorder Type/Group	Users	Anchor Tweets	Total Timeline Tweets	Music Count
Depression	829	900	1,468,693	8217
PTSD	413	453	799,503	5263
Anxiety	343	454	618,661	3311
Bipolar	250	270	414,713	3645
Borderline	81	90	143,917	720
Panic	28	30	52,012	417
Non-Disordered	3823	N/A	5,479,129	56,840
Total	5767	2197	8,976,628	78,413

Table 4 Summary of the Control Group (non-music-listening), categorized by disorder type

Disorder Type/Group	Users	Anchor Tweets	Total Timeline Tweets
Depression	3847	4253	4,816,597
PTSD	2576	3094	3,681,979
Anxiety	1803	2003	2,301,932
Bipolar	991	1088	1,377,646
Borderline	344	360	450,253
Panic	147	159	213,613
Non-Disordered	28,378	N/A	23,020,659
Total	38,086	10,957	35,862,679

examined these timelines to ensure no mention of mental health diagnoses, using a predetermined list of mental disorder lexicons. This process resulted in collecting the timelines of 32,201 non-disordered users, with a total of 28,499,788 tweets.

After identifying the disorder and non-disordered groups, we ensured that both user groups had records of music listening in their timelines by scanning for links to popular music platforms such as Spotify, SoundCloud, and Apple Music. Thus, we filtered out 38,086 users from disordered and non-disordered groups, identifying a total of 5767 users in the music-listening group, with 8,976,628 tweets in their timelines. Table 3 shows the distribution of users and tweets in the music-listening group.

Control Group This group consists of users with no records of music listening in their timelines and couldn't be included in a music-listening group. The control group included 9708 disordered users (those diagnosed with one of the six psychological disorders) and 28,378 non-disordered users with a total of 35,862,679 tweets in both groups. (see Table 4 for a detailed distribution of tweets and users in the control group.)

3.2 Music preference collection

As mentioned in Sect. 3.1, for gathering records of music listening activity for different user groups, we scanned their timelines to identify tweets containing links to popular music platforms (i.e., Spotify, SoundCloud, and Apple Music). Once identified, we select the corresponding user accounts for further analysis. We then use the Genius API[2] to retrieve the lyrics for each music track, filtering out non-English content to ensure consistency in our analysis. To ensure that our music dataset captures only the music tracks that are listened by individual users, we filter URLs associated with albums and playlists. The column "Music count" in Table 3 shows the number of music tracks across different user groups.

[2] https://genius.com/developers.

3.3 User data collection

We gathered three distinct types of data for every user: user profile data, timeline-related data, and music data. Each of these data categories possesses its own unique set of attributes. During the attribute collection process, any information containing personally identifiable details, such as profile names, is deliberately excluded to safeguard user anonymity and privacy.

User profile The user profile information contains the following attributes for each user: (1) *author_Id*: an anynomized unique user identifier; (2) *account_creation date*: time and date of creation of user account; (3) *description*: the short bio of users displayed on their user profile; (4) *user_location*: self-expressed location by users; (5) *following_count*: users' number of contacts; (6) *followers_count*: users' number of followers; (7) *verified_check*: a flag to note whether a user is verified; (8) *tweet_count*: number of tweets posted by users; and (9) *listed_count*: the number of accounts listed by a user.

User timeline Each user's timeline contains the contextual information of the last 3200 tweets posted by users. The attributes that are collected for each tweet are: (1) *embeddings* of tweet: 384-dimensional representation of tweet learned by all-MiniLM-L6-v2;[3] (2) *tweet_id*: an anynomized unique identifier of a tweet; (3) *tweet_timestamp*: tweet's creation time; (4) *tweet_source*: the device where the tweet is generated; (5) *anchor_flag*: a flag to specify whether the tweet is an anchor tweet in disorder class; (6) *referenced_tweet_type*: whether a tweet is a reply or a quote or neither of them (7) *like_count*: number of likes on the tweet; (8) *quote_counts*: number of times tweets are quoted; (9) *retweet_counts*: number of times tweets are retweeted; (10) *reply_counts*: number of replies on the tweets; (11) *hashtag*: the list of hashtags used in each tweet; (12) *URL*: the list of URLs embedded in tweets. The summary on the tweet sources can be found in Table 10 and 11 in Appendix C.

Music data The music information of each user contains the textual and contextual information of the music tracks listened by users and mentioned in their tweets in chronological order. The attributes collected for every music track are: (1) *music_lyrics*: the content of music; (2) *music_name*: name of the music; (3) *artist*: name of the artist who sang the music, (4) *music_source*: where the music is listened from; and (5) *music_timestamp*: the time that the music is listened to.

For each music, in addition to the attributes obtained directly from Twitter and Genius APIs, we enrich our dataset with some auxiliary music information. Furthermore, we applied j-hartmann/emotion-english-distilroberta-base [45] on music lyrics to extract the *7-dimension of emotions* (anger, disgust, joy, sadness, fear, neutral, and surprise) associated with each music lyric; We also applied siebert/sentiment-roberta-large-english [46] and calculated a *sentiment_score* associated with each music lyric. Table 5 summarizes the variables available in the Twitter-MusicPD.

4 Exploratory data analysis

Our exploratory data analysis focuses on two main user groups: (1) the music-listening group, consisting of Twitter users who have actively shared music tracks along with

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

Table 5 Variables and descriptions in the Twitter-MusicPD dataset

Variable	Description
User Profile	
Author ID	Anynomized unique user identifier
Account Creation Date	The time and date when the user's account was created
Description	The bio of the user as displayed on their profile
User Location	The self-expressed location of the user
Following Count	The number of users the individual follows
Followers Count	The number of users following the individual
Verified Check	A flag indicating whether the account is verified
Tweet Count	The total number of tweets the user has posted
Listed Count	The number of lists in which the user is included
User Timeline	
Tweet Embedding	Pre-trained 384-dimensional embedding of the tweet's text
Tweet ID	Anonymized unique identifier for each tweet
Tweet Timestamp	The time when the tweet was posted
Tweet Source	The device or platform from which the tweet was posted
Anchor Flag	A flag indicating whether the tweet is an anchor tweet in the disorder class
Referenced Tweet Type	Whether the tweet is a reply, a quote, or neither
Like Count	The number of likes the tweet has received
Quote Count	The number of times the tweet has been quoted
Retweet Count	The number of times the tweet has been retweeted
Reply Count	The number of replies to the tweet
Hashtag	The list of hashtags used in the tweet
URL	The list of URLs embedded in the tweet
Music Data	
Music Lyrics	The lyrics of the music tracks mentioned in the tweets
Music Name	The name of the music track
Artist	The name of the artist who performed the track
Music Source	The platform from which the music was listened (e.g., Spotify, Apple Music)
Music Timestamp	The time when the music was shared or mentioned in the tweet
Music Emotion	7-dimension of emotion expressed in the music lyrics (e.g., sadness, joy)
Music Sentiment	The sentiment score of the music lyrics, indicating positive, negative, or neutral

their tweets, and (2) the non-music-listening group, composed of users without recorded music-sharing activity. Both groups are further divided into users with and without psychological disorders to analyze the impact of music engagement on mental health expression in social media behavior.

The following sections will extensively analyze users' social engagement, tweet activity, music preferences, and topic modeling to compare behaviors across the different groups.

4.1 Social media engagement analysis

As shown in Fig. 1, we measure the social media engagement of users across four distinct groups using metrics such as the average number of followers, followings, and tweets per user. To maintain consistency with prior studies that report statistics of social media data [29, 47], we adopt the average measure as our primary metric. Figure 1 reveals that while music-listening groups generally maintain a similar number of followings and post a comparable number of average tweets as the non-music-listening groups, they exhibit fewer average followers. When comparing disorder and non-disordered groups, disordered groups tend to have more followers than non-disordered groups, despite having a slightly lower average number of tweets. This suggests that users with disorders tend to tweet less frequently but attract a broader audience who follows them. To capture variations beyond the mean, we provide percentile-based information (50th, 75th, 95th, and

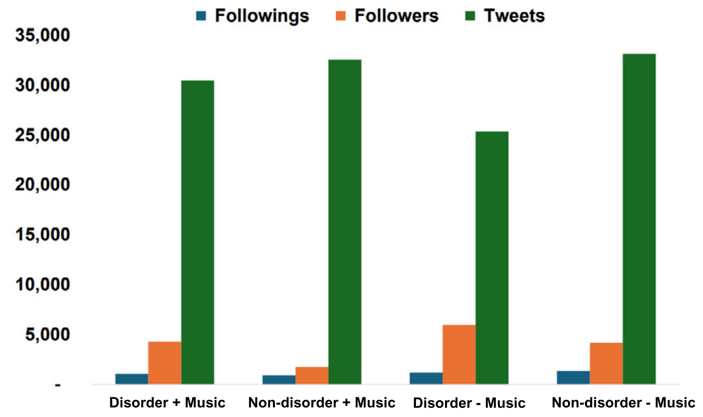


Figure 1 Social media engagement of disorder and non-disorder users with (+) and without (–) music, measured by the average number of followers, followings, and tweets per user

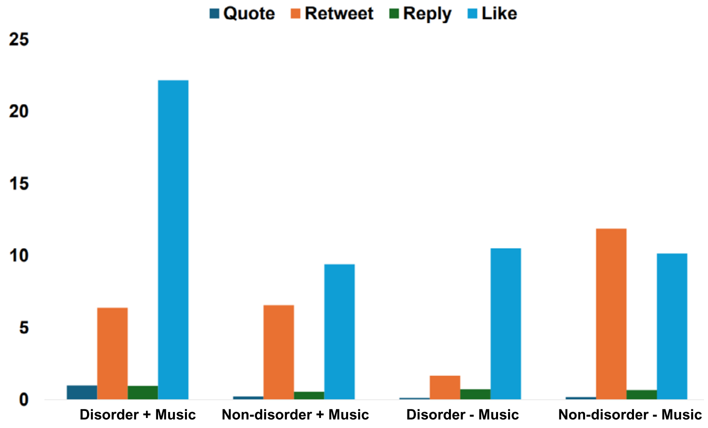


Figure 2 Overall tweeting activity of disordered and non-disordered users with (+) and without (–) music, including the average number of likes, replies, retweets, and quotes per user

99th percentiles) on engagement metrics, included in Appendix A, Table 7. The overall trends in Table 7 and Fig. 1 align well, but percentile-based data offer additional granularity that enhances our understanding of engagement behaviors. For example, users with disorders who do not listen to music have the highest tweet count at the 50th percentile (median) and 75th percentile but are surpassed at the 99th percentile, where users without disorders who do listen to music post the most tweets.

Figure 2 illustrates the overall tweeting activities within our dataset, such as average likes, replies, retweets, and quotes per user. Overall, music-listening behavior is associated with higher direct engagement (likes, replies, quotes), particularly among users with disorders. This indicates that music-listening users tend to generate more interaction and personal responses from their audience. In contrast, non-music-listening behavior, especially among non-disordered users, is more aligned with content amplification (retweets), suggesting that their posts are shared more widely, even if they receive less direct engagement. Similarly, Appendix A, Table 8 provides percentile-based engagement metrics, offering additional insights beyond the mean. The lower numbers observed in quotes,

Table 6 Music information for each user group

Group	Avg # of Music per User	Avg # of Words in Lyric	Spotify Songs	Apple Music Songs	Soundcloud Songs
Depression	9.91	391.23	6735	1334	148
PTSD	12.74	548.62	3710	1382	171
Anxiety	9.65	411.50	2537	722	52
Bipolar	14.58	379.78	2729	874	42
Borderline	8.88	342.63	635	73	12
Panic	14.89	222.42	256	136	25
Non-disordered	14.86	315.06	37,510	18,224	1106
All Users	13.59	343.76	54,112	22,745	1556

retweets, and replies align with existing research [29], suggesting that these forms of engagement occur less frequently than likes. Additionally, the Table 8 confirms that likes have the highest engagement levels across all groups, suggesting that liking is the dominant form of interaction on social media, regardless of disorder or music-listening behavior.

Additionally, we examined the music-listening behavior of users across different psychological disorders. Table 6 shows music-level statistics, including the average number of music listened to by each user group, the popular music platforms, as well as the average length of music lyrics in each user group. The observation indicates that panic and bipolar disordered users listen to the most songs on average (14.89 and 14.58, respectively), while borderline and depression users listen to fewer (8.88 and 9.91). PTSD users engage with the longest song lyrics (548.62 words), whereas panic disordered users prefer shorter lyrics (222.42 words). Spotify is the dominant platform across all groups, particularly for non-disordered users (37.510 songs), with Apple Music and SoundCloud being less frequently used. Depression and PTSD users show a relatively higher engagement with Apple Music. A list of the top three music and popular artists listened to by different user groups is shown in Table 9 in Appendix B.

4.2 Tweet analysis

In this section, we examine the linguistic differences between the music-listening and non-music-listening (control) groups to understand how music may influence language use on social media. Given the extensive amount of data available, we focus our analysis on the year 2021, which provides the most recent and comprehensive dataset for both tweet content and music-listening behaviors.

To investigate how music influences language usage in the two groups, we calculated the LIWC (Linguistic Inquiry and Word Count) [48] scores for different categories of words in tweets across both groups. LIWC provides a psycholinguistic analysis by categorizing words into various cognitive, emotional, and social domains, enabling us to assess how the presence or absence of music shapes users' language.

To facilitate the comparison, we calculate the percentage difference for each LIWC category g between the music-listening and non-music-listening groups as follows:

$$\text{Percentage Difference} = \frac{\text{LIWC}_{\text{non-music},g} - \text{LIWC}_{\text{music},g}}{\text{LIWC}_{\text{music},g}} \times 100 \qquad (1)$$

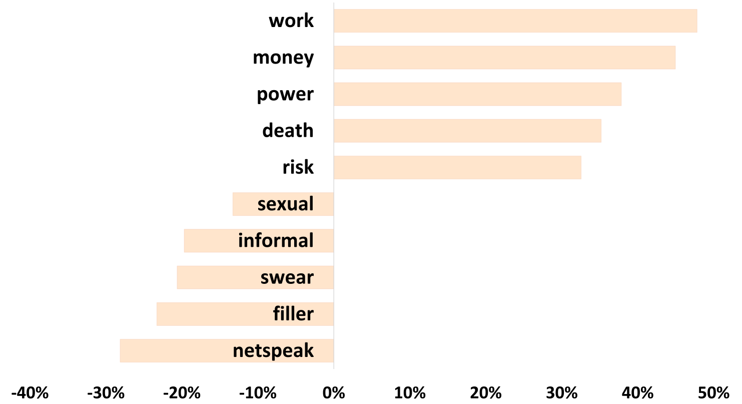


Figure 3 Top five LIWC categories with the largest percentage differences between music- listening and non-music-listening groups (2021)

Where $\mathrm{LIWC}_{\text{non-music},g}$ and $\mathrm{LIWC}_{\text{music},g}$ represent the LIWC scores for category g in the non-music-listening and music-listening groups, respectively. A positive value indicates higher scores for the non-music-listening group, while a negative value suggests higher scores for the music-listening group.

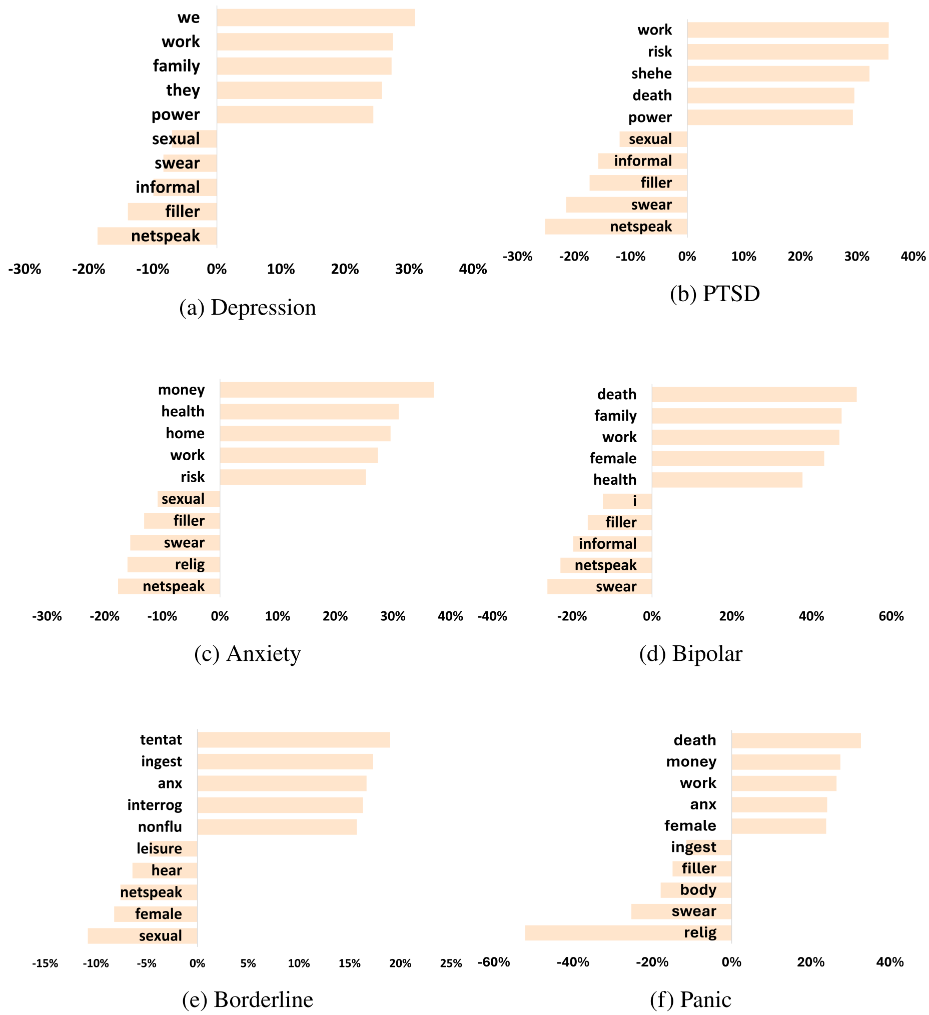
Figure 3 illustrates the top five LIWC categories with the most significant percentage differences between the music-listening and non-music-listening groups in 2021. The non-music-listening group scored higher in categories such as "work","money,"power", "death", and "risk", indicating a stronger focus on professional, financial, and existential topics. Common words in these categories include "work", "class", "cash", "dead", and "protect", suggesting that in the absence of music, users are more likely to engage in serious reflections on life and personal security.

Conversely, the music-listening group exhibited higher scores in categories related to "fillers", "swear words", "sexual terms", and "netspeak". This group frequently used words such as "wow", "you know", "shxt", "fxck", "sex", and "lol", indicating a more casual, informal, and sometimes explicit communication style. The influence of music may contribute to this more relaxed and spontaneous expression, consistent with prior research on the role of music in mood regulation and informal language use [49].

We also analyzed the differences in LIWC scores for users with mental disorders, comparing their language in tweets when they listened to music versus when they did not. Figure 4 highlights the top LIWC categories with the most significant percentage differences across six mental disorders.

For users with depression, tweets without music showed a higher frequency of social words (e.g., "we"), indicating a focus on social bonds and external validation, which is consistent with research showing that individuals with depression seek self-enhancing feedback from their social network [50].

For users with anxiety, tweets without music expressed concerns related to "money", "health", "home", and "risk", reflecting common areas of worry [51] and highlighting how stress from work can generate home-based anxiety [52]. Users with Bipolar disorder showed higher use of family-related words, pointing to the significant impact of family dynamics on their condition, aligning with previous research on family influence in bipolar disorder [53].



(a) Depression

(b) PTSD

(c) Anxiety

(d) Bipolar

(e) Borderline

(f) Panic

Figure 4 Top five LIWC categories with the largest percentage differences between music-listening and non-music-listening groups (2021), presented separately for individual disorders

For borderline personality disorder, users exhibited more tentative languages (i.e., “if”, “or”, “any”, “something”) and nonfluent language (i.e., “oh”, “um”, “uh”, “i”) in the absence of music, reflecting emotional instability and uncertainty, alongside concerns about ‘ingestion’ which are common due to the prevalence of eating disorders in this population [54]. Users with Panic disorder had increased focus on anxiety-related language when not listening to music, consistent with studies showing that a panic attack involves intense anxiety with sudden onset and brief duration [55].

PTSD users in the non-listening group are shown to use more personal pronouns (e.g., “she”, “he”), suggesting a narrative focus on trauma-related experiences, which aligns with the literature on the relational aspects of trauma in PTSD [56].

On the other hand, when users with mental disorders listened to music, we observed a notable increase in the use of informal language, including “swear words”, “fillers”, and “netspeak”, across all disorders. Additionally, users with anxiety and panic disorders discussed religion more often, possibly as a coping mechanism or due to the influence of themes in the music they listened to [57]. Borderline users showed higher usage of leisure-related and auditory terms (i.e., “hear”, “music”), reflecting relaxation or positive experiences associated with music, which may suggest that music serves as a form of emotional regulation for these individuals [58].

For users with panic disorder, music listening is associated with an increased focus on “ingest” (e.g., “food”) and “body” (e.g., “ache”, “heart”,“cough”) terms, indicating heightened awareness of physical sensations. This suggests that music may amplify individuals’ attention to their bodily experiences, possibly reflecting music’s emotional and physical impact on users with panic disorder [59].

4.3 Music analysis

In this section, we conduct two key analyses on the music lyrics listened to by users with and without psychological disorders. First, we examine the dominant emotions expressed in the lyrics to understand the emotional landscape of the music preferred by different user groups. Second, we perform a linguistic analysis to explore the linguistic features in the lyrics, comparing the prevalence of various linguistic categories between disordered and non-disordered groups.

4.3.1 Emotion analysis

We analyzed the emotions expressed in the music lyrics listened to by different user groups over the entire timeline. For each song, lyrics were extracted, and an emotion score was calculated for all emotions as described in Sect. 3.3. The emotion with the highest score was selected as the dominant emotion for each song. For each disordered group, we calculated the percentage distribution of songs in which each emotion was the top emotion. This approach allowed us to assess the emotional content of the music preferred by individuals with various disorders.

Figure 5 presents the distribution of dominant emotions in music across different disorders. Sadness emerged as the most prevalent emotion across all groups, with anxiety showing the highest association with sadness. Anger was consistently high, particularly for individuals with panic disorder. Fear was also prominent in anxiety and depression, while positive emotions like joy and surprise were less frequent compared to the dominant negative emotions (sadness, anger, and fear). Disgust was the least represented emotion across all disordered groups.

4.3.2 Linguistic analysis

We computed LIWC scores for music lyrics to analyze the linguistic differences between disordered and non-disordered groups in relation to the music they listen to. We began by separating unique music tracks for each user group to prevent overlap and ensure a clearer understanding of linguistic patterns specific to the disordered (comprising all six disordered groups) and non-disordered groups. LIWC scores were calculated based on the lyrics of these unique tracks, and percentage differences between disordered and non-disordered groups were derived to quantify how much more or less prominent each linguistic category was in the disordered group.

The percentage difference for each LIWC category was calculated using the approach shown in Equation (1) in the previous section. This formula allows us to assess how certain linguistic features in the lyrics are comparatively more or less dominant across groups.

Figure 6 shows the top music lyrics LIWC analysis result. Positive values indicate that the LIWC category is more prevalent among users with disorders, while negative values

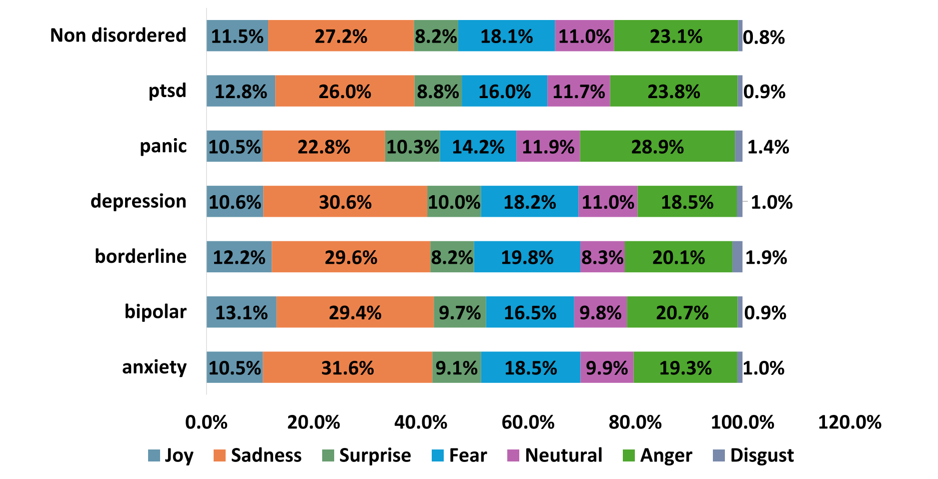


Figure 5 Distribution of top music emotions across different disorders

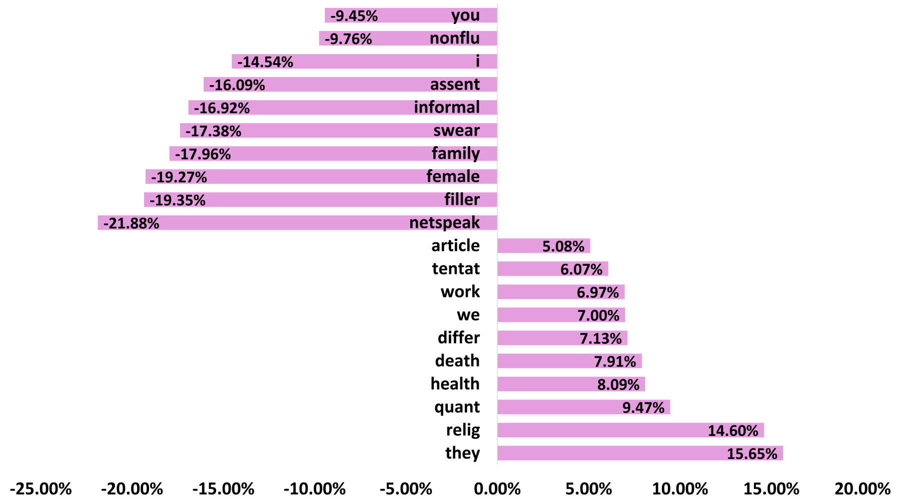
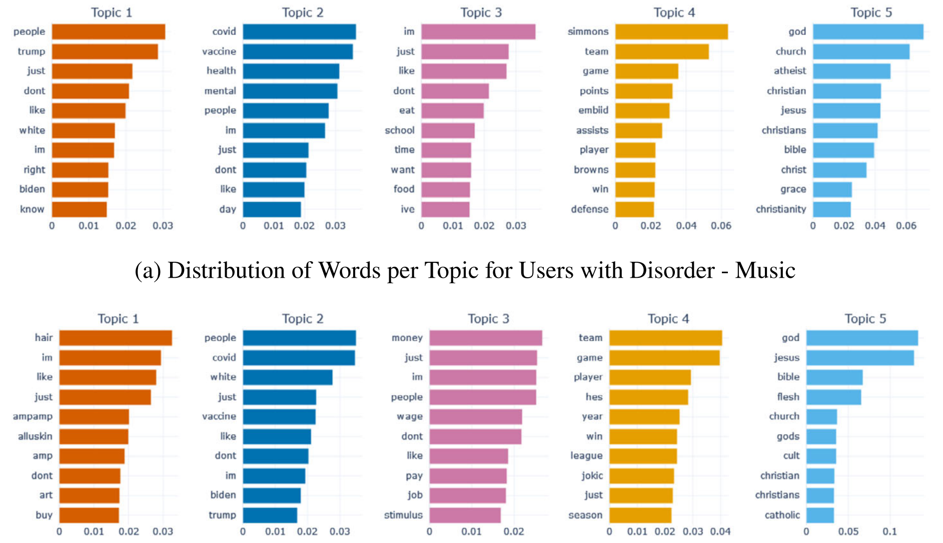


Figure 6 Percentage differences in LIWC usage for top music lyrics between disordered and non-disordered groups

suggest that it is less prevalent in the disordered groups compared to the non-disordered group. This comparison provides insights into the linguistic patterns in music lyrics and how they vary between users with and without disorders.

The analysis reveals that users with disorders tend to use collective pronouns like "they" and "we" more frequently, suggesting a focus on external or group dynamics, potentially reflecting feelings of detachment or isolation. Their music also includes more tentative language ("tentat"), such as "maybe" or "perhaps" which suggests a connection with uncertainty and hesitation. Additionally, the lyrics in their preferred music often address serious themes like "work," "death," "health," and "religion" pointing to a preoccupation with life's significant challenges and existential concerns. Their language is also more structured, frequently using quantifiers ('quant') and articles.



(a) Distribution of Words per Topic for Users with Disorder - Music

(b) Distribution of Words per Topic for Users with Disorder + Music

Figure 7 Topic analysis for disordered users with (+) and without (–) music

In contrast, the non-disordered group's music features more personal pronouns like "I" and "you", highlighting individual relationships and self-expression. Their lyrics tend to be more informal, with higher occurrences of non-fluencies ("nonflu"), swearing ("swear"), and conversational language ("netspeak" and "filler"), reflecting a more casual and relaxed engagement with music. Additionally, the non-disordered group has a stronger connection to personal and social relationships, as shown by more frequent references to family and female-related terms in their music. The use of assent (e.g., "yes" "okay") suggests a more affirming or positive tone, possibly indicating a more optimistic or agreeable outlook.

4.4 Topical analysis

In this section, we aim to explore the key topics discussed by users with psychological disorders, comparing those who listen to music with those who do not. We have leveraged collective word embeddings derived from the all-MiniLM-L6-v2 [60] model in the Twitter-MusicPD dataset and applied hierarchical clustering to facilitate the extraction of topics from the tweet data. We specifically focus on January 2021, as the start of a new year may reflect user sentiments and priorities shifts. For efficient topic modeling of the tweet embeddings, we employ BERTopic [61, 62]. Following the recommended preprocessing steps from [61], we lowercase the tweets and remove URLs, mentions, punctuation, and special characters. After inspecting the topics manually, we limit the number of topics to 10.

Figure 7 shows the word distribution of the top 5 topics among users with disorders, comparing those who listen to music and those who do not (a detailed list of topics with corresponding manually generated labels can be found in Appendix D).

The analysis shows distinct themes between those who listen to music (Disorder + music) and those who do not (Disorder - music). Music listeners focus more on lifestyle topics, including product promotions (Topic #1), engage deeply in political and racial discussions around COVID-19 (Topic #2), and are concerned with economic struggles (Topic #3). They also participate in sports conversations (Topic #4) and religious discussions (Topic #5), suggesting music helps them process emotions and engage with societal issues.

In contrast, non-music listeners emphasize U.S. political discourse (Topic #1), focus more on mental health and the pandemic's impact (Topic #2), and engage in practical conversations about daily life and work (Topic #3). Their discussions on sports (Topic #4) and religion (Topic #5) highlight a more introspective and personal approach, with a strong focus on mental health and daily challenges.

Both groups show concern about financial issues, but the focus differs: music listeners tend to link these concerns to broader economic systems and stock trading, while non-music listeners discuss wages and work-life balance more directly.

5 Discussion and conclusion

In this paper, we present the Twitter-MusicPD dataset, providing a comprehensive resource for examining the relationship between Twitter engagement, music preferences, and psychological well-being. To the best of our knowledge, this dataset is the first for investigating the intersection of music and tweet behaviors. We anticipate that this dataset will serve as a valuable asset for researchers investigating mental health topics in the context of online activity, providing a detailed understanding of users' evolving preferences and behaviors.

The exploratory data analysis revealed significant differences between the music-listening group and the non-music-listening group (i.e., the control group), particularly in linguistic pattern and word usage, emotional expression, and personal concerns. These differences highlight how individuals with and without music engagement express themselves on social media.

First, the dataset can be utilized to uncover correlations between users' tweet behaviors, music consumption habits, and self-reported psychological disorders. By examining patterns within tweet embeddings, music-sharing activities, and reported emotions, researchers can identify valuable information on how various mental health conditions manifest in online behaviors. For example, tweet embeddings can reveal distinct linguistic styles or topics of discussion prevalent among individuals with certain mental health disorders. These insights contribute to a deeper understanding of how mental health conditions are expressed and reflected in online interactions, offering new avenues for research in the field of computational social science and mental health.

Second, the Twitter-MusicPD dataset offers an extensive resource for diverse feature extraction experiments by incorporating both tweet and music data. Specifically, the dataset provides an opportunity to study potential predictive factors associated with specific psychological disorders. Researchers can explore patterns/indicators within tweet embeddings, sentiments or lyrical themes that may correlate with certain mental health conditions.

Third, the analysis uncovered insights into users' music listening behaviors, indicating a universal inclination towards music engagement across all user groups. This finding suggests the potential for leveraging music as a tool for tailored treatments or interventions for individuals with mental disorders. We hope the dataset helps researchers investigate the role of music as a coping mechanism for managing mental health. By examining the types of music shared by users with different psychological disorders and analyzing associated sentiments, researchers can gain insights into how music is utilized for emotional regulation and stress management.

Fourth, through exploration of the Twitter-MusicPD dataset, researchers have the opportunity to uncover emotional patterns and temporal variations in users' music preferences over an extensive period. This dataset, covering a substantial time frame, allows for rich temporal analysis, enabling researchers to gain deeper insights into users' psychological states over time. With access to this valuable resource, researchers can develop music-based interventions tailored to individuals' emotional needs, leveraging the longitudinal examination of music preferences provided by the Twitter-MusicPD dataset.

5.1 Limitations

Social media activity as a proxy for behavior has inherent challenges. One limitation of the Twitter-MusicPD dataset comes from relying on keyword searches to identify users within the disordered group, which may not capture all relevant users accurately and comprehensively. To address this limitation, future plans involve implementing a machine learning-based classifier that can leverage contextual cues, semantic relationships, and linguistic patterns to improve the identification of relevant content. Another limitation is that some users in the non-music listener group may engage with music but do not post about it or use less common platforms. To mitigate this issue, future work will explore alternative strategies and incorporate machine learning techniques to analyze engagement signals such as likes, retweets, and replies to music-related content, ultimately refining user classification. Furthermore, we acknowledge that the dataset only contains tweet-level embeddings without explicit mappings to assigned topics, making it impossible to link specific tweets to their corresponding topic clusters directly. To address this limitation, we have made the fitted BERTopic model available, enabling researchers to reproduce topic assignments and analyze topic distributions in a consistent manner. Moreover, social media data in general often follow a power-law distribution, highlighting the disproportionate influence of high-impact users (e.g., those with many followers, quotes, and replies). This suggests a direction for future research—examining how these users shape online discourse and amplify themes such as emotional expression through music and mental health awareness. Understanding their role in driving engagement could potentially offer valuable insights into digital influence and public conversations on well-being. Additionally, the dataset is constrained by the moderation policies enforced by the Twitter platform. This constraint means that content removed or deleted due to violations of platform rules is not included in the dataset, potentially leading to an incomplete representation of content related to the disordered group.

5.2 FAIR consideration

The Twitter-MusicPD dataset adheres to the FAIR principles of Findability, Accessibility, Interoperability, and Reusability and is accessible online.[4] Further, Twitter-MusicPD holds significant promise for researchers interested in exploring user-level data on multi-

[4] https://github.com/szamani20/Twitter-MusicPD_Melody-of-Minds.

ple mental health disorders and music preferences on Twitter. By providing a comprehensive array of metadata, Melody of Minds serves as a valuable resource for analyzing the intersection of mental health disorders and music tastes on social media. As a result, the dataset ensures both reusability and interoperability, aligning with the best data sharing practices.

5.3 Ethical considerations

During the careful collection of the Twitter-MusicPD dataset, we focus solely on publicly available information on Twitter via the Twitter API. To protect user privacy and align with ethical standards, we anonymized user IDs in both the music listening and non-music listening (control group) groups featured in our dataset. This process involves substituting identifiable user details with anonymized identifiers. This measure is of high importance, especially considering the inclusion of individuals with mental health disorders in the data, highlighting the critical need to maintain their anonymity in accordance with ethical protocols. Further, we maintain strict adherence to the Twitter Terms of Service[5] by not sharing the textual content of the tweets. Instead, we provide embeddings, which can be utilized for a diverse range of applications. These embeddings offer a flexible resource for various analyses, including sentiment analysis, topic modeling, user profiling, and similarity measures. Moreover, they can facilitate research in computational social sciences, Natural Language Processing, and machine learning. By prioritizing embeddings over textual content, we guarantee compliance with Twitter's policies while enabling diverse research opportunities in mental health disorders and music preferences.

Furthermore, our dataset collection methodology received an exemption from ethics review by the Toronto Metropolitan University's Research Ethics Board (REB).[6] This exemption was granted based on specific criteria: (1) research activities that don't involve direct interaction with individuals and don't raise privacy concerns; (2) utilization of datasets for analysis that are either publicly accessible or protected by law.

It's important to acknowledge that, similar to other datasets in this domain, the Twitter-MusicPD dataset carries the potential for misuse and adverse societal consequences. Specifically, the dataset may contain content related to specific mental disorders such as suicide, depression, anxiety, and others. If mishandled, inadequately analyzed, or subject to biased interpretation, there's a risk of aggravating stigmatization, enforcing harmful stereotypes, and triggering emotional distress among vulnerable individuals. Additionally, using such data for research purposes may raise ethical concerns regarding privacy and consent, as well as the potential for unintentional harm to individuals or communities represented in the dataset. It is crucial to approach the analysis and interpretation of this data with sensitivity and caution, ensuring that research methodologies prioritize the well-being and dignity of those affected by mental health issues. Transparent disclosure of limitations and ethical considerations is essential in mitigating these risks and promoting responsible use of the dataset.

[5] https://developer.twitter.com/en/developer-terms/agreement-and-policy.

[6] https://www.torontomu.ca/research/resources/ethics/course-based-research-ethics/#!accordion-1636735697402-what-does-not-require-review.

Appendix A: Additional statistics on social media engagement and overall tweeting activity

Table 7 Social media engagement of Disordered and Non-disordered users with (+) and without (–) music, measured using the percentile distribution of followers, followings, and tweets per user

User Group	50th Percentile	75th Percentile	95th Percentile	99th Percentile
Following				
Non-disorder+Music	482.0	946.0	2897.2	5689.8
Disorder+Music	560.5	1080.0	3417.5	6246.5
Non-disorder-Music	422.0	1034.0	4760.4	12,914.9
Disorder-Music	470.5	1080.0	4422.5	10,002.9
Followers				
Non-disorder+Music	519.0	1291.0	5358.6	26,357.6
Disorder+Music	597.5	1544.3	8134.5	30,314.1
Non-disorder-Music	312.0	1043.0	6855.0	41,058.3
Disorder-Music	358.0	1225.8	8879.8	39,165.8
Tweet				
Non-disorder+Music	13,180.0	35,439.0	128,251.8	282,267.7
Disorder+Music	16,487.5	37,722.5	116,844.3	242,566.1
Non-disorder-Music	9034.0	30,429.5	138,525.8	344,183.9
Disorder-Music	8581.0	25,708.3	94,558.0	219,058.6

Table 8 Overall tweeting activity of Disorder and Non-disorder users with (+) and without (–) music, measured using the percentile distribution of likes, replies, retweets, and quotes per user

User Group	50th Percentile	75th Percentile	95th Percentile	99th Percentile
Quote				
Non-disorder + Music	0.0	0.0	0.0	2.0
Disorder + Music	0.0	0.0	0.0	1.0
Non-disorder – Music	0.0	0.0	0.0	2.0
Disorder – Music	0.0	0.0	0.0	1.0
Retweet				
Non-disorder + Music	0.0	0.0	1.0	11.0
Disorder + Music	0.0	0.0	1.0	7.0
Non-disorder – Music	0.0	0.0	2.0	13.0
Disorder – Music	0.0	0.0	1.0	9.0
Reply				
Non-disorder + Music	0.0	1.0	2.0	5.0
Disorder + Music	0.0	1.0	2.0	5.0
Non-disorder – Music	0.0	1.0	2.0	6.0
Disorder – Music	0.0	1.0	2.0	6.0
Like				
Non-disorder + Music	1.0	2.0	12.0	82.0
Disorder + Music	1.0	2.0	13.0	74.0
Non-disorder – Music	0.0	1.0	12.0	90.0
Disorder – Music	1.0	2.0	14.0	87.0

Appendix B: Top artists and songs

Table 9 Top artists and songs for each user group

	Top1 Artist	Top2 Artist	Top3 Artist	Top1 Song	Top2 Song	Top3 Song
Depression	Taylor Swift	Nicki Minaj	Mitski	All Too Well, Taylor Swift	You Don't Own Me, Tasha Bloom	Vibez, Zayn
PTSD	The Superlatives	Venus Leone	Big KRIT	Sonder, The Superlatives	Higher Place, Skip Marley	I Want You, Common
Anxiety	Taylor Swift	Mac Miller	The Weeknd	Gave Em Hope, $hy Boog	Fallingforyou, The 1975	All Too Well, Taylor Swift
Bipolar	Taylor Swift	Ariana Grande	Demi Lovato	Sober, Demi Lovato	Happier, Marshmello	1-800-273-8255, Logic
Borderline	Taylor Swift	Lucas Savant	Halsey	Wish List, Lucas Savant	Black Bathing Suit, Lana Del Rey	Everything Is Embarrassing, Sky Ferreira
Panic	MC Cone	Elton John	BONES	Keep Grinding, MC Cone	After All, Elton John	Donald Trump's Neck P***y, MC Cone
Non-Disorder	Richard C. Rocha	Taylor Swift	Drake	Winds of Freedom, Richard C. Rocha	Just A Feeling, Joy Xande	Never Forget, Tam Dogg
All Users	Taylor Swift	Richard C. Rocha	Drake	Winds of Freedom, Richard C. Rocha	Just A Feeling, Joy Xande	Never Forget, Tam Dogg

Appendix C: Tweet source

Table 10 Tweet counts from different sources (Music Listening Group)

	iPhone	Android	Web App	Others	All Sources
Depression	876,070	368,816	173,123	50,684	1,468,693
PTSD	490,606	172,726	82,333	53,838	799,503
Anxiety	381,584	135,558	79,172	22,347	618,661
Bipolar	287,657	70,058	38,708	18,290	414,713
Borderline	75,994	30,472	23,216	14,235	143,917
Panic	41,809	4777	3901	1525	52,012
Non-Disorder	3,527,272	1,134,690	536,942	280,225	5,479,129
All Users	5,680,992	1,916,221	938,271	441,144	8,976,628

Table 11 Tweet counts from different sources (Control Group)

	iPhone	Android	Web App	Others	All Sources
Depression	2,163,652	1,422,721	870,233	359,991	4,816,597
PTSD	1,696,272	999,854	695,933	289,920	3,681,979
Anxiety	1,101,324	641,936	419,911	138,761	2,301,932
Bipolar	578,994	300,742	189,866	308,044	1,377,646
Borderline	204,473	134,464	74,488	36,828	450,253
Panic	129,104	37,086	32,395	15,028	213,613
Non-Disorder	9,144,823	5,792,578	3,901,505	4,181,753	23,020,659
All Users	15,018,642	9,329,381	6,184,331	5,330,325	35,862,679

Appendix D: Topic analysis

Table 12 Topic Themes and Top 10 Words for Each Topic

#	Topic theme	Top 10 Words for Each Topic
Disorder Class + music		
1	Product Promotion	(‘hair’, ‘im’, ‘like’, ‘just’, ‘ampamp’, ‘alluskin’, ‘amp’, ‘dont’, ‘art’, ‘buy’)
2	COVID-19, Politics	(‘people’, ‘covid’, ‘white’, ‘just’, ‘vaccine’, ‘like’, ‘dont’, ‘im’, ‘biden’, ‘trump’)
3	Economy, Labor Concerns	(‘money’, ‘just’, ‘im’, ‘people’, ‘wage’, ‘dont’, ‘like’, ‘pay’, ‘job’, ‘stimulus’)
4	Sports	(‘team’, ‘game’, ‘player’, ‘hes’, ‘year’, ‘win’, ‘league’, ‘jokic’, ‘just’, ‘season’)
5	Religion	(‘god’, ‘jesus’, ‘bible’, ‘flesh’, ‘church’, ‘gods’, ‘cult’, ‘christian’, ‘christians’, ‘catholic’)
Disorder Class - music		
1	U.S. Politics	(‘people’, ‘trump’, ‘just’, ‘dont’, ‘like’, ‘white’, ‘im’, ‘right’, ‘biden’, ‘know’)
2	COVID-19, Mental Health	(‘covid’, ‘vaccine’, ‘health’, ‘mental’, ‘people’, ‘im’, ‘just’, ‘dont’, ‘like’, ‘day’)
3	Daily Life	(‘im’, ‘just’, ‘like’, ‘dont’, ‘eat’, ‘school’, ‘time’, ‘want’, ‘food’, ‘ive’)
4	Sports	(‘simmons’, ‘team’, ‘game’, ‘points’, ‘embiid’, ‘assists’, ‘player’, ‘browns’, ‘win’, ‘defense’)
5	Religion	(‘god’, ‘church’, ‘atheist’, ‘christian’, ‘jesus’, ‘christians’, ‘bible’, ‘christ’, ‘grace’, ‘faith’)

Abbreviations

ADHD, Attention Deficit Hyperactivity Disorder; API, Application Programming Interface; DSM-5, Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition; EDA, Exploratory Data Analysis; FAIR, Findability, Accessibility, Interoperability, and Reusability; GAD, General Anxiety Disorder; LIWC, Linguistic Inquiry and Word Count; OCD, Obsessive-Compulsive Disorder; PTSD, Post-Traumatic Stress Disorder; REB, Research Ethics Board; SMHD, Self-reported Mental Health Diagnoses; Twitter-MusicPD, Twitter-Music-Psychological Disorder; Twitter-STMHD, Self-Reported Temporally-Contextual Mental Health Diagnosis Dataset; URL, Uniform Resource Locator.

Acknowledgements

Not applicable

Author contributions

S.Z.A: Data collection, analysis. X.Y.: Conceptualization, visualization. Z.N: Analysis, methodology development A.G.: Visualization, writing of the methods section F.Z. Literature review, methodology development Everyone contribute to overall writing to the paper.

Funding information

This research does not have funding.

Data Availability

Our data is available at: https://github.com/szamani20/Twitter-MusicPD Melody-of-Minds.

Declarations

Ethics approval and consent to participate

This research received an exemption from ethics reviewing from the Toronto Metropolitan University’s Research Ethics Board (REB).

Consent for publication

All authors consent to the publication of this manuscript.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Computing and Software, McMaster University, Hamilton, Canada. [2]Ted Rogers School of Information Management, Toronto Metropolitan University, Toronto, Canada. [3]School of Engineering, University of Guelph, Guelph, Canada.

Received: 27 October 2024 Accepted: 2 April 2025 Published online: 07 April 2025

References

1. Bathina KC, Ten Thij M, Lorenzo-Luaces L, Rutter LA, Bollen J (2021) Individuals with depression express more distorted thinking on social media. Nat Hum Behav 5(4):458–466
2. Yang X, Joukova A, Ayanso A, Zihayat M (2022) Social influence-based contrast language analysis framework for clinical decision support systems. Decis Support Syst 159:113813

3. Ophir Y, Tikochinski R, Asterhan CS, Sisso I, Reichart R (2020) Deep neural networks detect suicide risk from textual Facebook posts. Sci Rep 10(1):1–10
4. Reece AG, Reagan AJ, Lix KL, Dodds PS, Danforth CM, Langer EJ (2017) Forecasting the onset and course of mental illness with Twitter data. Sci Rep 7(1):1–11
5. Gkotsis G, Oellrich A, Hubbard T, Dobson R, Liakata M, Velupillai S, Dutta R (2016) The language of mental health problems in social media. In: Proceedings of the third workshop on computational linguistics and clinical psychology, pp 63–73
6. Hegde S (2017) Music therapy for mental disorder and mental health: the untapped potential of Indian classical music. BJPsych International 14(2):31–33
7. Gupta U, Gupta B (2016) Gender differences in psychophysiological responses to music listening. Music and Medicine 8(1):53–64
8. Zatorre R (2005) Music, the food of neuroscience? Nature 434(7031):312–315
9. Sakka LS, Juslin PN (2018) Emotion regulation with music in depressed and non-depressed individuals: goals, strategies, and mechanisms. Music Sci 1:2059204318755023
10. Gustavson DE, Coleman PL, Iversen JR, Maes HH, Gordon RL, Lense MD (2021) Mental health and music engagement: review, framework, and guidelines for future studies. Transl Psychiatry 11(1):1–13
11. Maratos A, Gold C, Wang X, Crawford M (2008) Music therapy for depression. Cochrane Database Syst Rev (1)
12. Mössler K, Chen X, Heldal TO, Gold C (2011) Music therapy for people with schizophrenia and schizophrenia-like disorders. Cochrane Database Syst Rev 12
13. Johnson BK, Ranzini G (2018) Click here to look clever: self-presentation via selective sharing of music and film on social media. Comput Hum Behav 82:148–158
14. Shukla S, Khanna P, Agrawal KK (2017) Review on sentiment analysis on music. In: 2017 international conference on infocom technologies and unmanned systems (trends and future directions) (ICTUS). IEEE, pp 777–780
15. Çano E, Morisio M (2017) Moodylyrics: a sentiment annotated lyrics dataset. In: Proceedings of the 2017 international conference on intelligent systems, metaheuristics & swarm intelligence, pp 118–124
16. Alavijeh SZ, Zarrinkalam F, Noorian Z, Mehrpour A, Etminani K (2023) What users' musical preference on Twitter reveals about psychological disorders. Inf Process Manag 60(3):103269
17. Yang Y-H, Liu J-Y (2013) Quantitative study of music listening behavior in a social and affective context. IEEE Trans Multimed 15(6):1304–1315
18. Stewart J, Garrido S, Hense C, McFerran K (2019) Music use for mood regulation: self-awareness and conscious listening choices in young people with tendencies to depression. Front Psychol 10:1199
19. Braun Janzen T, Al Shirawi MI, Rotzinger S, Kennedy SH, Bartel L (2019) A pilot study investigating the effect of music-based intervention on depression and anhedonia. Front Psychol 10:1038
20. McFerran KS, Hense C, Koike A, Rickwood D (2018) Intentional music use to reduce psychological distress in adolescents accessing primary mental health care. Clin Child Psychol Psychiatry 23(4):567–581
21. Nakamura T, Kubo K, Usuda Y, Aramaki E (2014) Defining patients with depressive disorder by using textual information. In: 2014 AAAI. Spring symposium series
22. Reece AG, Danforth CM (2017) Instagram photos reveal predictive markers of depression. EPJ Data Sci 6(1):15
23. Wheaton MG, Prikhidko A, Messner GR (2021) Is fear of covid-19 contagious? The effects of emotion contagion and social media use on anxiety in response to the coronavirus pandemic. Front Psychol 11:567379
24. Kadkhoda E, Khorasani M, Pourgholamali F, Kahani M, Ardani AR (2022) Bipolar disorder detection over social media. Inf Med Unlocked 32:101042
25. Murarka A, Radhakrishnan B, Ravichandran S (2021) Classification of mental illnesses on social media using roberta. In: Proceedings of the 12th international workshop on health text mining and information analysis, pp 59–68
26. Coppersmith G, Dredze M, Harman C (2014) Quantifying mental health signals in Twitter. In: Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, pp 51–60
27. Shen G, Jia J, Nie L, Feng F, Zhang C, Hu T, Chua T-S, Zhu W (2017) Depression detection via harvesting social media: a multimodal dictionary learning solution. In: IJCAI, pp 3838–3844
28. Cohan A, Desmet B, Yates A, Soldaini L, MacAvaney S, Goharian N (2018) Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. In: 27th international conference on computational linguistics, pp 1485–1497. ACL
29. Singh AK, Arora U, Shrivastava S, Singh A, Shah RR, Kumaraguru P, et al (2022) Twitter-stmhd: an extensive user-level database of multiple mental health disorders. In: Proceedings of the international AAAI conference on web and social media, vol 16, pp 1182–1191
30. Ji S, Li X, Huang Z, Cambria E (2022) Suicidal ideation and mental disorder detection with attentive relation networks. Neural Comput Appl 34(13):10309–10319
31. Villa-Pérez ME, Trejo LA, Moin MB, Stroulia E (2023) Extracting mental health indicators from English and Spanish social media: a machine learning approach. IEEE Access 11:128135–128152
32. Raihan MN, Puspo SSC, Farabi S, Bucur A, Ranasinghe T, Zampieri M (2024) Mentalhelp: a multi-task dataset for mental health in social media. In: Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation, LREC/COLING 2024, Torino, Italy, 20–25 May, 2024, pp 11196–11203
33. Chen X, Sykora MD, Jackson TW, Elayan S (2018) What about mood swings: identifying depression on Twitter with temporal measures of emotions. In: Companion proceedings of the web conference 2018, pp 1653–1660
34. Blasco-Magraner JS, Bernabé-Valero G, Marín-Liébana P, Botella-Nicolás AM (2023) Changing positive and negative affects through music experiences: a study with university students. BMC Psychol 11(1):76
35. Golden TL, Springs S, Kimmel HJ, Gupta S, Tiedemann A, Sandu CC, Magsamen S (2021) The use of music in the treatment and management of serious mental illness: a global scoping review of the literature. Front Psychol 12:880
36. Garrido S, Schubert E (2015) Moody melodies: do they cheer us up? A study of the effect of sad music on mood. Psychol Music 43(2):244–261
37. Saarikallio S, Gold C, McFerran K (2015) Development and validation of the h ealthy-u nhealthy m usic s cale. Child Adolesc Ment Health 20(4):210–217
38. Kanagala SC, Schäfer T, Greenberg DM, Gabińska A (2021) Depression symptoms relationship with music use: investigating the role of trait affect, musical ability, music preferences. Music Sci 4:20592043211057217

39. Tang Q, Huang Z, Zhou H, Ye P (2020) Effects of music therapy on depression: a meta-analysis of randomized controlled trials. PLoS ONE 15(11):0240862
40. Garrido S, Schubert E (2015) Music and people with tendencies to depression. Music Percept 32(4):313–321
41. Wilhelm K, Gillis I, Schubert E, Whittle EL (2013) On a blue note: depressed peoples' reasons for listening to music. Music and Medicine
42. Silverman MJ (2022) Music therapy in mental health for illness management and recovery. Oxford University Press, London
43. Gutiérrez EOF, Camarena VAT (2015) Music therapy in generalized anxiety disorder. Arts Psychother 44:19–24
44. American Psychiatric Association (2013) Diagnostic and statistical manual of mental disorders: DSM-5, vol 5. American psychiatric association, Washington
45. Hartmann J (2022) Emotion English DistilRoBERTa-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/
46. Hartmann J, Heitmann M, Siebert C, Schamp C (2023) More than a feeling: accuracy and application of sentiment analysis. Int J Res Mark 40(1):75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005
47. Jang Y, Park C-H, Seo Y-S (2019) Fake news analysis modeling using quote retweet. Electronics 8(12):1377
48. Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) The development and psychometric properties of liwc2015
49. Ludke KM (2019) Songs and music. In: The handbook of informal language learning, pp 203–213
50. Hames JL, Hagan CR, Joiner TE (2013) Interpersonal processes in depression. Annu Rev Clin Psychol 9(1):355–377
51. Fan L-B, Blumenthal J, Watkins L, Sherwood A (2015) Work and home stress: associations with anxiety and depression symptoms. Occup Med 65(2):110–116
52. Doby VJ, Caplan RD (1995) Organizational stress as threat to reputation: effects on anxiety at work and at home. Acad Manag J 38(4):1105–1123
53. Reinares M, Bonnín C, Hidalgo-Mazzei D, Sánchez-Moreno J, Colom F, Vieta E (2016) The role of family interventions in bipolar disorder: a systematic review. Clin Psychol Rev 43:47–57
54. Khosravi M (2020) Eating disorders among patients with borderline personality disorder: understanding the prevalence and psychopathology. J Eat Disord 8(1):38
55. Association AP (2013) Diagnostic and statistical manual of mental disorders, Fifth edn. American Psychiatric Publishing, Arlington
56. Dorahy MJ, Corry M, Shannon M, MacSherry A, Hamilton G, McRobert G, Elder R, Hanna D (2009) Complex ptsd, interpersonal trauma and relational consequences: findings from a treatment-receiving northern Irish sample. J Affect Disord 112(1–3):71–80
57. Williams ML (2022) I will sing to the lord as long as I live: christian music as an effective coping mechanism for catholic undergraduate students. The Institute for the Psychological Sciences
58. Baltazar M, Saarikallio S (2019) Strategies and mechanisms in musical affect self-regulation: a new model. Music Sci 23(2):177–195
59. Akhshabi M, Rahimi M (2021) The impact of music on sports activities: a scoping review. J New Stud Sport Manag 2(4):274–285
60. Transformers S (2021) all-MiniLM-L6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
61. Grootendorst M (2022) Bertopic: Neural topic modeling with a class-based tf-idf procedure. Preprint. Available at arXiv:2203.05794
62. Egger R, Yu J (2022) A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify Twitter posts. Front Sociol 7:886498

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.