Analyzing the Impact of Album Releases on Consumer Engagement in Online Communities of Top Spotify Artists

Group 1: Khoi Gia Pham[a], David Lensen[a], Floris de Lange[a], Lorenzo Belfiore[a]

[a]*RSM, Burgemeester Oudlaan 50, Rotterdam, 3062PA, Netherlands*

Abstract

The advent of music streaming platforms like Spotify has revolutionized the music industry, allowing artists and labels to release new albums directly to fans. In this study, we investigate the impact of album releases from top Spotify artists on the sentiment of user-generated content (UGC) in Reddit communities. By collecting data using the Spotify and Reddit APIs, we analyze the sentiment of UGC before and after new album releases. We also compare the performance of various machine learning models to predict the sentiment of an album based on its sonic characteristics retrieved from the Spotify API. The findings of this research can provide valuable insights for the music industry, artists, and labels, helping them understand the reception of their music among fans and make informed decisions about future album releases. This study ultimately contributes to a deeper comprehension of the relationship between new album releases, sonic characteristics, and the sentiment of UGC on social media platforms.

Keywords: new album releases, online fan communities, sonic characteristics, top artists, sentiment analysis, machine learning

1. Introduction

1.1. Orientation

The music industry has experienced significant changes due to streaming platforms, with Spotify emerging as a leader [1]. As its popularity grows, artists and labels use the platform for album releases [2]. Album releases are anticipated events, generating excitement and discussion on social media platforms [3].

Online fan communities are crucial spaces for fandom expression and social bonds [4, 5]. Album releases by top Spotify artists impact user-generated content (UGC) sentiment within these communities [6]. UGC often includes album discussions, song reactions, and fan-made artwork. Album releases can increase fan engagement and sentiment on social media, with positive sentiment dominating [6]. Social media activity increases during and after an album's release [6].

However, not all album releases positively affect fan communities. Controversial lyrics or negative audio features may result in backlash [7]. Thus, it's essential to examine the impact of album releases on fan communities comprehensively. Studying the impact of new releases from top Spotify artists on UGC sentiment is vital. This research aims to explore the impact of new album releases from highly followed Spotify artists on the sentiment of UGC in corresponding Reddit communities. We will collect data on album releases and analyze UGC sentiment pre and post-release using Spotify and Reddit APIs. Leveraging machine learning (ML) techniques, we will compare models' performance on predicting album sentiment based on sonic characteristics from the Spotify API.

This research has implications for economic, managerial, and societal concerns. It provides insights into how audio features of new releases affect UGC sentiment. This study can help artists and labels understand their music's perception and make informed decisions about future releases. The findings have applications in Consumer Behavior and Decision-Making, Influencer Marketing, and Online Reputation Management. Understanding product characteristics and social media sentiment's influence on consumer choices allows businesses to tailor products and marketing strategies while influencers manage their online reputation effectively.

In the context of Sustainable Development Goals (SDGs), investigating music's influence on emotions supports mental health and well-being (SDG 3). The findings can stimulate educational initiatives promoting creativity and innovation within creative industries (SDG 4), ultimately leading to economic expansion and employment opportunities in these fields (SDG 8).

1.2. Research Objective

The research objective of this study is to explore the impact of new album releases from the most followed artists on Spotify on the sentiment of UGC in corresponding Reddit communities dedicated to those artists' fan bases. The study aims to use the Spotify and Reddit APIs to collect data on new album releases and analyze it impacts on the sentiment of UGC before and after the release. Additionally, the research will leverage machine learning techniques to compare the performance of various models in predicting whether an album release will have a positive or negative sentiment based on the album's sonic characteristics, as retrieved from the Spotify API.

1.3. Research Question

To achieve this objective, the following research question is formulated:

"How can machine learning algorithms be applied to predict the sentiment of user-generated content (UGC) on online fan communities dedicated to an artist in the event of a new album release, using audio features?"

2. Literature Review

2.1. Online Fan Communities

Online fan communities have emerged as important spaces for fans to express their admiration, passion, and loyalty for various forms of media and entertainment, such as music, movies, television series, and sports teams [8]. These communities can be found on social media platforms, forums, and other online sites, where fans come together to share their thoughts, emotions, experiences, and creative works related to their favorite artists, songs, or albums [5].

The rise of the internet and social media platforms has transformed the ways fans interact with each other and their favorite artists, enabling more direct and immediate communication between fans and artists [9]. This unprecedented level of access has fostered a sense of intimacy and connection between fans and artists, potentially strengthening the fan-artist relationship [10]. In the context of music fandom, online fan communities play a crucial role in the promotion and consumption of music, as well as the formation of fan identities [11]. Fans use these online spaces to discuss their favorite artists, songs, and albums, share their interpretations of lyrics, and create fan art, videos, or remixes of songs [12].

The impact of online fan communities on the music industry has also been the subject of recent research, with scholars examining the role of fan communities in influencing music consumption, artist popularity, and the dynamics of artist-fan relationships [10]. Duffet [10] investigated the relationship between online fan communities and music streaming on platforms like Spotify, finding that fan engagement on social media platforms was positively correlated with music streaming numbers. Zhou [6] also explored the impact of album releases on fan communities, noting that album releases often led to increased fan engagement, with fans discussing the album, sharing their reactions to songs, and creating fan art related to the album.

2.2. Album Releases and Sentiment Analysis

The release of a new album by an artist is often an eagerly awaited event for fans, generating significant discussion and engagement within online fan communities [5]. Sentiment analysis, a computational technique for determining the sentiment or emotion expressed in text data, has been increasingly used to study the impact of album releases on the sentiment of UGC in these communities [13].

Past literature found that positive sentiment was the most prevalent emotion expressed by fans, with the album's release resulting in a significant increase in fan engagement on Twitter [14]. However, album releases may not always have a positive impact on fan communities, as some artists may face backlash due to controversial lyrics or audio features of songs in their new albums [7].

Recent studies have also explored the relationship between the audio features of songs in an album and the sentiment of user-generated content in online fan communities. For instance, Zangerle et al. [15] investigated the impact of audio features, such as tempo, energy, and valence, on the sentiment of tweets related to several albums by various artists. The researchers found that audio features could partially explain the sentiment expressed by fans on Twitter, with higher-energy songs being associated with more positive sentiment.

In summary, sentiment analysis is a valuable tool for understanding the impact of album releases on UGC sentiment in online fan communities. Existing literature suggests that album releases often result in increased fan engagement, dominated by positive sentiment. However, backlash due to controversial content can lead to negative sentiment within fan communities. The analysis of the relationship between audio features and sentiment, as well as the temporal aspects of album releases, provide insights into fan engagement and the impact of album releases on fan communities.

2.3. Audio Features and Listeners' Sentiment

The relationship between audio features of music and the sentiment it evokes in listeners has been a subject of interest for researchers in the fields of music psychology, music information retrieval, and emotion recognition [16, 17]. The audio features of a song, such as tempo, mode, energy, and valence, have been found to play a crucial role in shaping listeners' emotional responses and, consequently, the sentiment expressed in UGC within online fan communities [15].

Eerola and Vuoskoski [16] conducted a comprehensive review of studies exploring the relationship between musical features and emotions, highlighting the importance of factors such as tempo, mode, and harmony in evoking emotional responses in listeners. Faster tempos and major modes were generally associated with positive emotions, such as happiness and excitement, whereas slower tempos and minor modes were linked to negative emotions, such as sadness and melancholy [16].

In a study by Koenigstein et al. [17], the researchers used an emotion-based music retrieval system that analyzed various audio features, such as pitch, rhythm, and timbre, to classify songs according to their emotional content. The study found that these audio features could effectively predict the emotions conveyed by a song, demonstrating the potential of using audio features to understand the sentiment expressed in UGC related to music.

Zangerle et al. [15] investigated the impact of audio features, such as tempo, energy, and valence, on the sentiment of tweets related to several albums by various artists. The researchers found that audio features could partially explain the sentiment expressed by fans on Twitter, with higher-energy songs being associated with more positive sentiment. These studies highlight the importance of audio features in shaping listeners' emotional responses to music and the sentiment expressed in UGC within online fan communities. By understanding the relationship between audio features and sentiment, researchers can gain valuable insights into the factors that influence fan engagement and the impact of album releases on fan communities.

3. Methodology

3.1. Data Collection

This section describes the data collection process for this study, which includes information about the audio features of at most the 10 most recent album releases of the top 50 most followed artists on Spotify, along with UGC in corresponding Reddit communities dedicated to these artists' fan bases. The window of analysis covers 90 days before and after the release date of an album. The data collection process consists of three main datasets:

3.1.1. Top 50 Most Followed Artists

The first dataset contains the names of the top 50 most followed artists on Spotify. This list was obtained from a secondary dataset provided on Kaggle [18]. The original dataset includes information on music artists and their popularity, followers, genres, and names obtained from Spotify.

3.1.2. User-Generated Content on Reddit

The second dataset includes user-generated content of fans on Reddit, specifically focusing on threads' titles, texts, and comments in corresponding communities (subreddits) of the top 50 most followed artists. The official subreddits were manually identified based on the number of subreddit members and community descriptions. Other information included in this dataset is authors' names, creation dates of threads/comments, numbers of upvotes, downvotes, scores, total awards received, golds, and URLs to the threads/comments. To extract this data, R (programming language) and the RedditExtractoR package [19] were used.

3.1.3. Audio Features of Tracks

The third dataset comprises the audio features of the tracks including the albums by the top 50 most followed artists on Spotify We chose up to ten of the most recent albums available globally (i.e., available in over 100 markets) and used R, the spotifyR package [20], and the Spotify API to collect the data. We use the function get_artist_audio_features() in the spotifyR package to collect the audio feature information for all or part of an artists' discography. Variables of interest included in this dataset are artist name, artist id, album id, album name, eleven sonic characteristics of each song in the album (danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo), and the number of markets available.

3.2. Data Preprocessing

This section discusses the data preprocessing steps to clean, transform, and structure the collected data for further analysis.

3.2.1. User-Generated Content on Reddit

The data collected from Reddit yielded two distinct data frames: threads and comments. For the threads data frame, we merged the title and text columns into a new column named self-text. We then subsetted the data frame to include only the URL, author, date, self-text, and subreddit columns. In the comments data frame, we renamed the 'comment' column as self-text, and again subsetted it to include only the URL, author, date, and self-text columns. Next, we performed a left join (threads left join to comments) on the URL column, with a many-to-many relationship, to associate each comment with its corresponding subreddit. We then combined the two data frames by row binding and removed any duplicate rows.

We discuss sentiment analysis in this section since the sentiment column from the 'User-Generated Content on Reddit' dataset will be integrated into the second dataset, which has been discussed in section 3.1.3.

To begin with, we preprocess the text data utilizing the tidytext package in R [21]. This process encompasses several steps: removal of all characters except letters and spaces, conversion to lowercase, whitespace trimming, stopword elimination, and word stemming to restore their original forms.

Subsequently, we employ the get_nrc_sentiment() function from the syuzhet package to conduct sentiment analysis on the self-text column [22]. This function is based on the NRC (National Research Council) Word-Emotion Association Lexicon, a pre-compiled dictionary comprising words and their associated emotions. It quantifies the occurrences of words in the input text dataset that correspond to each of the eight fundamental emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) along with two supplementary sentiments (positive and negative). The output generated by the get_nrc_sentiment() function is a data frame containing the frequencies of the ten sentiments for each observation (i.e., each row) in the input text dataset. In the final step, we incorporate the net sentiment column into the corresponding self-text rows in the initial dataset.

3.2.2. Audio Features of the New Album Release by Top 50 artist

In the Audio Features of Tracks dataset (Section 3.1.3), we aggregate the data by artist_id and album_id, and subsequently compute the mean value of each album's sonic characteristics based on the tracks encompassed within that album. It is noteworthy that certain albums, despite possessing distinct album_ids, share identical album names. This phenomenon may be attributed to the presence of multiple versions of an album. Consequently, we exclude duplicate album names, retaining only the earliest version, which can be logically considered the "original" version. Furthermore, for each artist, we select a maximum of 10 recent albums based on their release dates in order to maintain

3.2.3. Final Dataset

In preparation for generating the final dataset, we first associate the pertinent subreddit with each album in the second dataset (Section 3.1.3). Subsequently, for every album, we compute the total sentiment score within a 90-day window preceding the album's release date. Additionally, for each album, we determine the total sentiment score within a 90-day window following the album's release date.

We then introduce a 'release_effects' column that assumes "positive" values if the difference between the post-release sentiment score and the pre-release sentiment score exceeds 0, and "negative" otherwise. Finally, we merge this dataset with the 'Audio Features of the New Album Release by Top 50 Artists' (Section 3.1.3) to generate the comprehensive final dataset.

3.3. Variable Description

Table 1 provides an overview of the variables included in the final dataset used in our analysis. There are 18 variables present in this table, that are unique identifiers for each album and artist, measures of musical characteristics such as danceability and energy, and information about the release date and release effects of each album on the fan sentiment.

Table 1: Variables overview

No.	Variable	Type	Description
1	album_id	chr	Unique identifier for each album
2	album_name	chr	Name of each album
3	artist_name	chr	Name of the artist/band associated with the album
4	artist_id	chr	Unique identifier for each artist
5	album_release_date	chr	Release date of each album (YYYY-MM-DD)
6	danceability	num	Measure of how suitable a track is for dancing based on a combination of musical elements.
7	energy	num	Measure of intensity/activity (higher values = more energetic)
8	key	num	Musical key of a track (encoded as integers)
9	loudness	num	Overall loudness of a track in decibels (dB)
10	mode	num	Modality (major = 1, minor = 0)
11	speechiness	num	Measure of spoken words presence (higher values = more speech-like)
12	acousticness	num	Measure of acousticness (higher values = more acoustic)
13	instrumentalness	num	Measure of vocals amount (higher values = more instrumental)
14	liveness	num	Measure of live audience presence (higher values = more likely live)
15	valence	num	Measure of musical positiveness (higher values = more positive)
16	tempo	num	Estimated tempo of a track in BPM
17	num_markets	int	Number of markets where the album is available
18	release_effects	chr	Effect of album release on artist's popularity

3.3.1. Input Variables

In this research paper, eleven audio features of an album, that were collected through Spotify API, are used to develop the predictive models. Detailed description the input variables are as follows:

1. **Danceability**: A continuous variable ranging from 0.0 to 1.0 that describes how suitable a track is for dancing based on a combination of musical elements. Higher values indicate more danceable tracks.
2. **Energy**: A continuous variable ranging from 0.0 to 1.0 that measures the intensity and activity of a track. Higher values indicate more energetic tracks.
3. **Key**: An integer variable representing the musical key of a track using standard Pitch Class notation (e.g., 0 C̄, 1 C̄#/Db, 2 D̄, etc.). The value -1 is used when no key is detected.
4. **Loudness**: A continuous variable representing the overall loudness of a track in decibels (dB), with values typically ranging between -60 and 0 dB. Higher values indicate louder tracks.
5. **Mode**: A binary variable indicating the modality of a track, with 1 representing major and 0 representing minor.
6. **Speechiness**: A continuous variable ranging from 0.0 to 1.0 that detects the presence of spoken words in a track. Higher values indicate a higher proportion of spoken words in the track.
7. **Acousticness**: A continuous variable ranging from 0.0 to 1.0 that measures the confidence of whether a track is acoustic. Higher values indicate a higher likelihood that the track is acoustic.
8. **Instrumentalness**: A continuous variable ranging from 0.0 to 1.0 that predicts the likelihood a track contains no vocals. Higher values indicate a greater probability of the track being purely instrumental.
9. **Liveness**: A continuous variable ranging from 0.0 to 1.0 that detects the presence of a live audience in the recording. Higher values indicate a higher likelihood that the track was performed live.
10. **Valence**: A continuous variable ranging from 0.0 to 1.0 that measures the musical positiveness conveyed by a track. Higher values indicate more positive-sounding tracks (e.g., happy, cheerful, euphoric), while lower values indicate more negative-sounding tracks (e.g., sad, depressed, angry).
11. **Tempo**: A continuous variable representing the overall estimated tempo of a track in beats per minute (BPM).

3.3.2. Output Variables

In this study, the output variable is release_effects which represents the sentiment of UGC within Reddit communities in response to new album releases by the respective artists. This variable is ascertained through the difference in sentiment scores pre- and post-album release. It assumes a positive value (1) if the difference exceeds 0 and a negative value (0) otherwise. Sentiment scores are computed by employing sentiment analysis with the NRC Word-Emotion Association Lexicon, which is available in the syuzhet package in R. The analysis encompasses a timeframe of 90 days before and after the album's release.

3.4. Machine Learning Modeling

In this study, six machine learning models — Logistic Regression, Support Vector Machine (SVM), Classification Tree, Random Forest, K-Nearest Neighbors (KNN), and Neural Network (NN) were employed to predict the fan sentiment based on the elven sonic characteristics. And the performance of the models was evaluated using five metrics: accuracy, sensitivity, specificity, precision, and area under the receiver operating characteristic curve (AUC).

3.4.1. Logistic Regression, SVM, Classification Tree, Random Forest, and KNN

Initially, we performed a comparative evaluation of the first five algorithms: Logistic Regression, SVM, Classification Tree, Random Forest, and KNN. This evaluation employs k-fold cross-validation on the final dataset, containing only the Input and Output variables specified.

In more detail, we set the seed for reproducibility to 123, guaranteeing consistent random number generation across multiple runs. Next, the dataset is divided into five equal parts for cross-validation. We then initialize accuracy vectors and confusion matrix lists to store each algorithm's confusion matrices across the five folds. The primary loop iterates through these folds, using one fold as the test set and the other four as the training set. We train and evaluate each of the five classification algorithms on these sets, calculating model accuracy by comparing predicted and actual test set values. Additionally, we record each model's confusion matrix.

The functions and hyperparameters used to develop the five models include:

1. **Logistic Regression**: We employ the 'glm' function for model training with a binomial family and the 'predict' function to compute predicted outcome probabilities. A 0.5 threshold is applied to convert probabilities into binary predictions.
2. **SVM**: We use the 'svm' function for model training with C-classification type and the 'predict' function to generate predicted outcomes.
3. **Classification Tree**: We utilize the 'rpart' function for model training with a class method and information split criterion, and the 'predict' function to generate predicted outcomes.
4. **Random Forest**: We apply the 'randomForest' function for model training with 100 trees and the calculated number of variables per split. The "vars_per_split" variable is the square root of the total predictor variables in the model minus one, which is used as the "mtry" parameter in the Random Forest algorithm. The 'predict' function generates predicted outcomes.
5. **KNN**: We use the 'knn' function to generate predicted outcomes with k=5 neighbors.

Upon completing the cross-validation process, the code outputs accuracy values and confusion matrices for each algorithm across the five folds. By analyzing these results, we can compare the algorithms' performance and select the most suitable model for the given dataset.

3.4.2. Neural Network

Next, we conduct an assessment of a NN classification algorithm to determine the ideal number of hidden layers for the model. This assessment is carried out on the Final dataset using a 70/30 train/test split. The initial step involves defining the 'get_nn_accuracies' function. This function accepts the dataset, model formula, and the number of hidden layers as its arguments. Within the function, the dataset is split into training (70%) and testing (30%) sets. The neural network model is subsequently trained with the 'nnet' function, using the specified number of hidden layers and a maximum of 300 iterations. The 'trace' parameter is set to FALSE to prevent output during training. The function computes the accuracy for both the training and testing sets by comparing the predicted values with the actual values. The confusion matrix is also calculated for the predicted outcomes on the test set. Ultimately, the function returns a list containing the training accuracy, test accuracy, and confusion matrix.

The second step requires setting the seed for reproducibility, guaranteeing consistent random number generation across multiple runs. The random seed is set to 123.

The third step entails looping over varying hidden layer sizes. A loop iterates through a range of 1 to 15 layers to assess the performance of the neural network model with different numbers of hidden layers. The 'get_nn_accuracies' function is invoked with the current number of hidden layers (n), and the results are stored in a list.

Subsequently, a data frame is generated to store the number of hidden layers, train accuracy, and test accuracy for each iteration. The index of the highest test accuracy in this data frame is identified using the 'which.max' function. The optimal number of hidden layers is then extracted from this dataframe using the aforementioned index. The neural network model is trained on the entire dataset with the best number of hidden layers using the 'nnet' function. The 'predict' function is employed to compute the probabilities of the predicted outcomes for the entire dataset. A threshold of 0.5 is applied to convert the probabilities into binary predictions.

In the final step, the confusion matrix for the best neural network model is compared with those of the five models mentioned earlier.

3.5. Conclusion

Overall, the methodology used in this study provides a rigorous approach to investigating the impact of new album releases from top Spotify artists on the sentiment of UGC in Reddit communities. The data preparation process ensured that relevant data was collected and processed in a format suitable for analysis. The machine learning modeling process included test and train split, hyperparameter tuning, and model evaluation, providing a robust means of predicting the sentiment of an album based on its sonic characteristics. The findings of this study can provide valuable insights for the music industry, artists, and labels, helping them understand the reception of their music among fans and make informed decisions about future album releases.

4. Research Results

4.1. Descriptive Analysis

Our final dataset consisted of 397 observations, with each observation representing an album and its associated audio features, such as danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. The last column release_effects, represents the sentiment of UGC within Reddit communities in response to the album releases.

4.1.1. Summary of Input Variables

Table 2: Summary Statistics of Input Variables

Statistic	N	Mean	St. Dev.	Min	Max
danceability	397	0.606	0.114	0.188	0.805
energy	397	0.663	0.139	0.128	0.982
key	397	5.382	1.018	2.429	8.500
loudness	397	−6.733	2.385	−20.690	−2.107
mode	397	0.606	0.179	0.100	1.000
speechiness	397	0.109	0.093	0.032	0.923
acousticness	397	0.235	0.172	0.0001	0.833
instrumentalness	397	0.051	0.108	0.000	0.721
liveness	397	0.242	0.166	0.081	0.935
valence	397	0.477	0.136	0.060	0.843
tempo	397	120.960	9.951	82.367	145.812

1. **Danceability**: The average danceability score is 0.606, with a median of 0.624, indicating that the distribution might be slightly left-skewed. The minimum and maximum values are 0.189 and 0.805, respectively, with a standard deviation of 0.114, showing moderate variability in this feature.
2. **Energy**: The mean energy value is 0.663, and the median value is 0.673. This suggests a relatively symmetric distribution. The energy values range from 0.128 to 0.982, with a standard deviation of 0.139, indicating a moderate variation in energy levels across albums.
3. **Key**: The mean and median key values are 5.38 and 5.44, respectively, which suggests a nearly symmetric distribution. The key values range from 2.43 to 8.50, with a standard deviation of 1.018, indicating a moderate level of variation.

4. **Loudness**: The average loudness value is -6.73 dB, with a median of -6.28 dB, indicating a slightly left-skewed distribution. The loudness values range from -20.69 dB to -2.11 dB, with a standard deviation of 2.39 dB, reflecting a considerable variation in loudness levels.
5. **Mode**: The mean and median mode values are 0.606 and 0.615, respectively, suggesting a nearly symmetric distribution. The mode values range from 0.1 to 1, with a standard deviation of 0.179, indicating a moderate level of variation in the modality of the tracks.
6. **Speechiness**: The average speechiness score is 0.109, with a median of 0.083, which implies a right-skewed distribution. The values range from 0.032 to 0.923, with a standard deviation of 0.093, showing a wide variability in the number of spoken words in tracks.
7. **Acousticness**: The mean acousticness value is 0.235, with a median of 0.199, indicating a slight right-skewed distribution. The range of acousticness values is quite wide, from a minimum of 5.86e-05 to a maximum of 0.833, with a standard deviation of 0.172, suggesting a high variability in this feature across albums.
8. **Instrumentalness**: The average instrumentalness score is 0.051, with a median of 0.002, suggesting a highly right-skewed distribution. The values range from 0 to 0.721, with a standard deviation of 0.108, indicating a wide variability in the presence of instrumental tracks.
9. **Liveness**: The mean liveness score is 0.242, while the median score is 0.184, indicating a right-skewed distribution. The values range from 0.081 to 0.935, with a standard deviation of 0.166, showing relatively high variability in liveness across albums.
10. **Valence**: The average valence score is 0.477, with a median of 0.463, suggesting a slightly right-skewed distribution. The valence values range from 0.060 to 0.843, with a standard deviation of 0.136, indicating a moderate level of variability in the perceived positivity or negativity of tracks across albums.
11. **Tempo**: The mean tempo value is 120.96 BPM, and the median value is 121.19 BPM, indicating a roughly symmetric distribution. The tempo values range from 82.37 BPM to 145.81 BPM, with a standard deviation of 9.95 BPM, suggesting a moderate level of variation in tempo across albums.

Overall, the sonic characteristics of the albums in the dataset exhibit varying degrees of variability. Features such as acousticness, instrumentalness, liveness, and speechiness show relatively high variability, while features like danceability, energy, key, mode, tempo, and valence exhibit moderate variability.

Understanding the interdependencies between variables is also important in gaining insights into the research question. To visualize the relationships between the various features in our dataset, we present a correlation matrix in Figure 1. This matrix provides an overview of the correlations between the variables, allowing for the identification of potential trends and patterns in the data. As shown in the figure, the highest correlations are between energy and loudness, as well as between danceability and valence. These observations align with our intuitive expectations. Please note that this plot merely presents correlations between variables, no conclusions can be drawn on causation. Moreover, in our opinion, the relationships are in fact not causal.

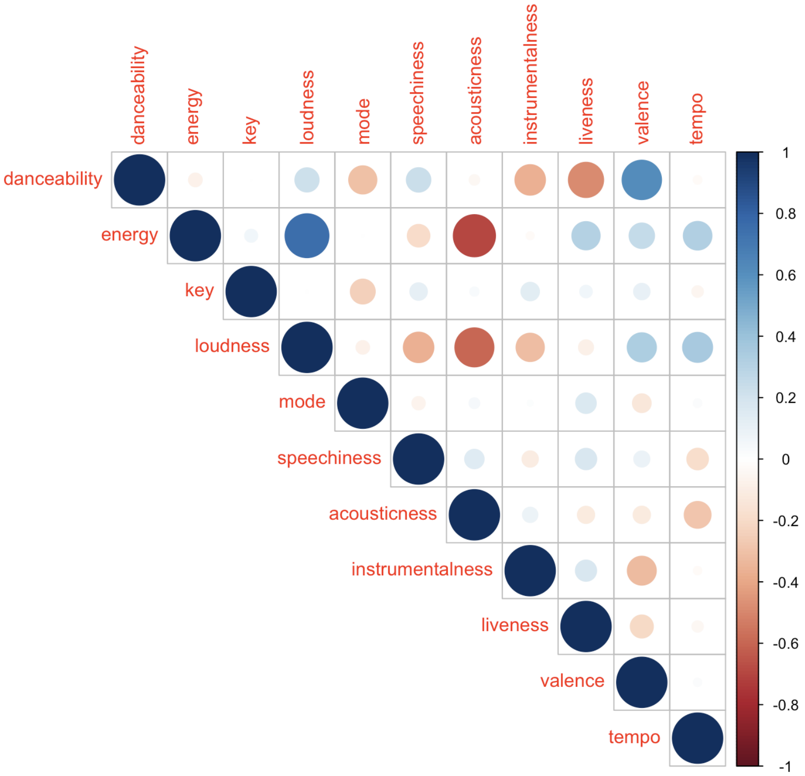

Figure 1: Correlation matrix of input variables

4.1.2. Distribution of the Output Variable

Figure 2 illustrates the distribution of the categorical variable 'release_effects', offering insights into the prevalence of positive and negative sentiment in UGC within Reddit communities following new album releases by top Spotify artists.

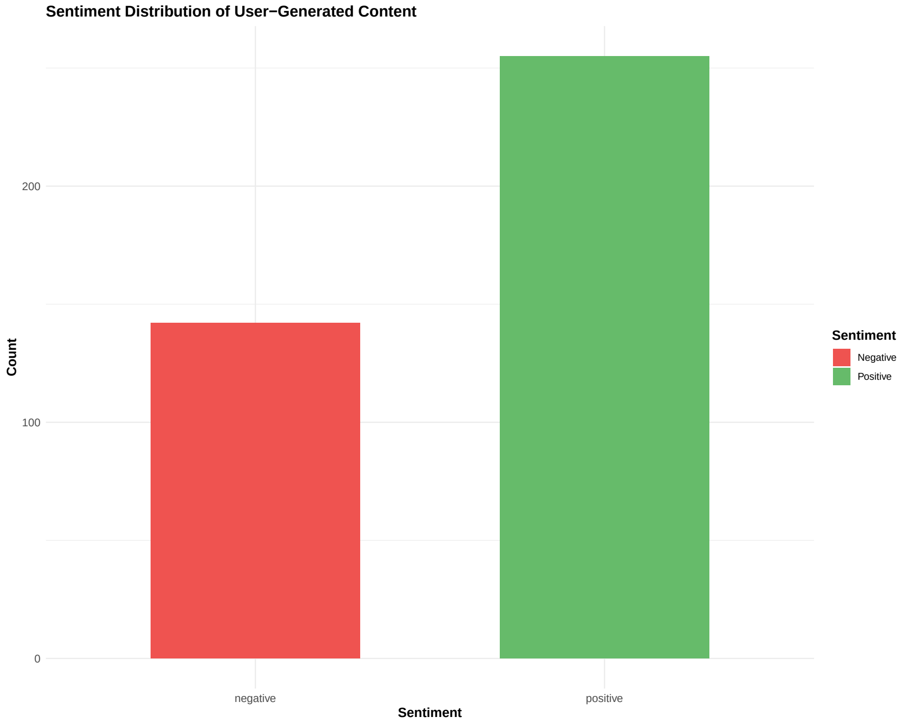

Figure 2: Distribution of the output variable in final dataset

As depicted in the graph, approximately 250 instances (65% of the dataset) correspond to new album releases eliciting positive sentiment in the relevant Reddit communities. In contrast, 142 instances (35% of the dataset) demonstrate new album releases evoking negative sentiment.

These observations imply that new album releases by highly followed artists on Spotify generally garner predominantly positive sentiment from fans on Reddit. Nevertheless, a substantial proportion of album releases (35%) provoke negative sentiment, suggesting that factors such as sonic characteristics, artist popularity, or album context might contribute to the observed variation in sentiment.

4.2. Machine Learning Model Performance

In this section, model performance will be discussed for the six developed models specifically Logistic Regression, SVM, Classification Tree, Random Forest, KNN, and Neural Network. The objective is to predict the sentiment polarity (positive or negative) associated with album releases, taking into account the sonic attributes of the albums. The performance of each model was evaluated using a range of metrics, including accuracy, sensitivity, specificity, precision, and AUC. A comparison of the results can be found in Table 3.

Table 3: Performance metrics of all six models

Metric	LR	SVM	Tree	RF	KNN	NN
Accuracy	0.612	0.637	0.589	0.610	0.567	0.650
Sensitivity	0.918	0.980	0.718	0.835	0.765	1.000
Specificity	0.634	0.021	0.359	0.204	0.211	0.000
Precision	0.638	0.643	0.668	0.653	0.635	0.650
AUC	0.608	0.800	0.822	0.979	0.619	0.629

A detailed analysis of the results is provided below:

- **Accuracy**: This metric indicates the proportion of correct predictions made by the model. The Neural Network model demonstrated the highest accuracy (0.650), followed by the SVM model (0.637). These results suggest that the Neural Network and SVM models outperformed the other models in terms of overall prediction accuracy.

- **Sensitivity**: Sensitivity measures the proportion of true positive instances that are correctly identified by the model. The Neural Network model achieved perfect sensitivity (1.000), while the SVM model also exhibited high sensitivity (0.980). This implies that both the Neural Network and SVM models were particularly effective at identifying albums with positive sentiment.

- **Specificity**: This metric reflects the proportion of true negative instances that are accurately identified by the model. The Logistic Regression model performed best in terms of specificity (0.634), followed by the Classification Tree model (0.359). The Neural Network model, however, had a specificity of 0, indicating that it failed to identify any negative sentiment albums correctly. High specificity is essential for effectively detecting albums with negative sentiment.

- **Precision**: Precision refers to the proportion of true positive instances among the predicted positive instances. The Classification Tree model displayed the highest precision (0.668), followed by the Random Forest model (0.653). This suggests that the Classification Tree and Random Forest models were more reliable in their positive sentiment predictions.

- **AUC**: The area under the receiver operating characteristic curve represents the model's ability to distinguish between positive and negative sentiment albums accurately. The Random Forest model had the highest AUC score (0.979), indicating superior performance in discriminating between the two classes. The Classification Tree model also demonstrated a high AUC score (0.822). Figure 3 visualizes the ROC curves of all six models, and therefore provides insights in the AUC scores.

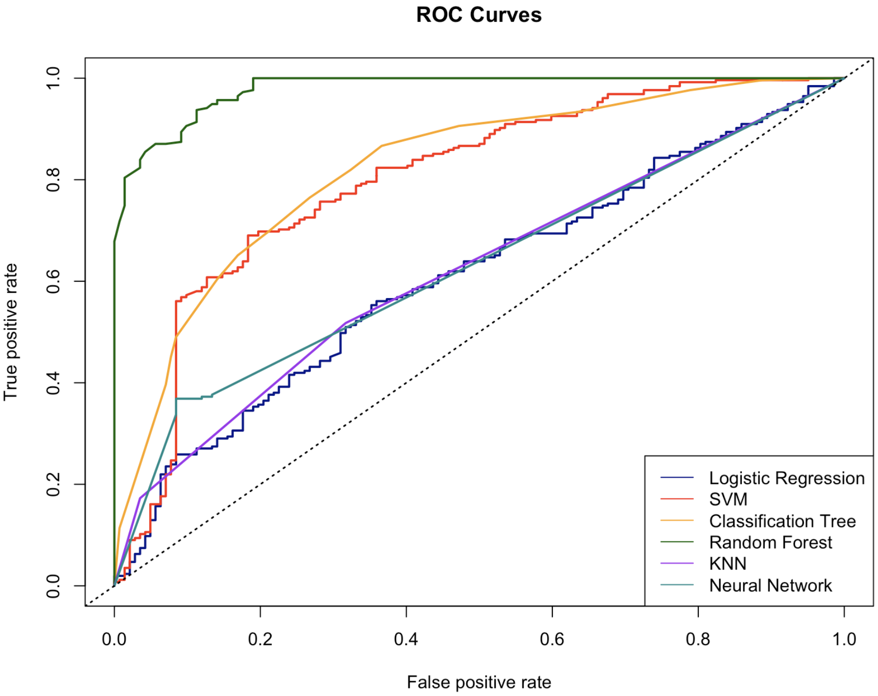

Figure 3: ROC Curves of Machine Learning Models

5. Conclusion

5.1. Recap of research design and methods

In brief, this study aimed to investigate the impact of album releases from the top Spotify artists on the sentiment of UGC in Reddit communities and to compare the performance of various machine learning models in predicting the sentiment of an album based on its sonic characteristics retrieved from the Spotify API.

To achieve this objective, the following research question was formulated:

> *"How can machine learning algorithms be applied to predict the sentiment of user-generated content (UGC) on online fan communities dedicated to an artist in the event of a new album release, using audio features?"*

Primary data was collected from the Spotify and Reddit APIs, and preprocessed to obtain the necessary variables for analysis. Six different machine learning models—Logistic Regression, SVM, Classification Tree, Random Forest, KNN, and Neural Network were employed to predict album sentiment based on sonic characteristics. And the performance of the models was evaluated using five metrics: accuracy, sensitivity, specificity, precision, and AUC.

5.2. Summary of Results

In summary, the results from Table 3 indicated that the Neural Network and SVM models demonstrated the highest accuracy and sensitivity. However, the Neural Network seves as a baseline model that always predict a 'positive value' for the outcome variable. Meanwhile, the Logistic Regression model exhibited the highest specificity. In terms of precision, the Classification Tree model performed best, and the Random Forest model achieved the highest AUC score.

Considering these findings, the SVM and Random Forest models were identified as promising candidates for predicting UGC sentiment in online fan communities. The high sensitivity of the Neural Network and SVM models suggests their effectiveness in identifying positive sentiment, which is likely to be common in fan communities. Additionally, the Random Forest model displayed strong discriminatory capabilities between positive and negative sentiment albums, as evidenced by its high AUC score.

To apply these models to the prediction of UGC sentiment, it is crucial to preprocess and clean the UGC data, extract relevant audio features from the new album, and train and validate the selected models on a dataset containing UGC, audio features, and known sentiment labels. The models' performance should be evaluated using the same metrics as in the initial analysis, and the most suitable model(s) should be selected based on the application requirements and context.

5.3. Contribution to Theory

From a theoretical perspective, the findings of this study contribute to the existing literature by demonstrating the effectiveness of machine learning algorithms in predicting the sentiment of user-generated content in online fan communities. The study also expands on previous research by exploring the use of audio features to predict sentiment, which contributes to the growing body of literature on the impact of social media and online communities on consumer behavior and decision-making. Moreover, the study highlights the strengths and weaknesses of various machine learning models, providing valuable information for future researchers working in this area.

5.4. Contribution to Practice

From a practical perspective, the study has important implications for the music industry. The findings can help artists and music labels to better understand how their music is perceived by their fans and make informed decisions about future album releases. This can lead to better album reception, increased engagement, and ultimately, greater commercial success.

Beyond the music industry, the study has implications for consumer behavior and decision-making, influencer marketing, and online reputation management. By understanding the influence of product characteristics and social media sentiment on consumer choices, businesses can tailor their products and marketing strategies to cater to customer needs. Influencers can use the findings to manage their online reputation more effectively.

Furthermore, the study's findings can be linked to the SDGs. Investigating the influence of music on listeners' emotions supports mental health and well-being (SDG 3). The study can also stimulate educational initiatives that encourage creativity and innovation within creative industries (SDG 4), ultimately leading to economic expansion and employment opportunities in these fields (SDG 8).

5.5. Limitations

5.5.1. Setting

One limitation of this study is the focus on a specific set of Reddit communities dedicated to music discussion. While these communities may represent a substantial portion of online music discussion, they may not reflect the sentiment of all music listeners. Additionally, the study focuses only on the sentiment of written comments on Reddit and does not account for other forms of user-generated content, such as images, videos, or audio recordings. Furthermore, Reddit users may not necessarily be representative of the wider population of music listeners, and the sentiments expressed on the platform may not reflect those expressed in other online or offline settings. As such, the findings of this study may have limited generalizability to other platforms or contexts beyond the Reddit communities analyzed in this study.

5.5.2. Data quality

There are several potential limitations of the dataset which may affect the interpretation and generalizability of the findings. Firstly, the scope is confined to the top 50 artists on Spotify, resulting in a potential bias towards popular artists and their fanbases while neglecting niche or emerging artists' communities. Secondly, the reliance on Reddit user-generated content may not encompass the full range of opinions and sentiments, as other platforms, such as Twitter or music review websites, could offer alternative viewpoints. Thirdly, the 90-day timeframe before and after album releases might be insufficient to capture long-term sentiment trends or delayed reactions.

One of the main limitations of this study is related to the quality of the data used for analysis. Although we collected data from a large number of Reddit communities and used sentiment analysis tools to evaluate the sentiment of the user-generated content, there are still potential limitations in the accuracy and reliability of the data. For instance, the sentiment analysis tools may not be able to capture the complexity of human emotions and can sometimes misinterpret the intended sentiment of the text. Additionally, the data collected from Reddit may not be representative of the entire population of music fans and may be biased towards certain demographics or geographic locations. Furthermore, the data collection process relied on user-generated content, which may not necessarily reflect the opinions of the general population or represent the opinions of professional music critics.

5.5.3. Sentiment Analysis

Furthermore, the sentiment analysis based on the NRC Word-Emotion Association Lexicon may not account for language nuances, slang, or culturally specific expressions unique to certain music genres or artist fanbases. Additionally, the dataset may not capture confounding factors influencing sentiment, such as marketing campaigns, artist controversies, or external events unrelated to the album release. Also, the sentiment analysis method could misinterpret sarcasm, irony, or complex emotional expressions, leading to inaccuracies in sentiment scores. Nevertheless, the method ignores words in other languages and content in picture, sound, or video formats.

5.5.4. Machine Learning Models

Machine learning models have demonstrated efficacy in predicting sentiment polarity based on an album's auditory attributes; however, certain limitations warrant acknowledgment.

Firstly, the models are trained on a relatively limited dataset, potentially constraining their applicability to a broader population. This is particularly pertinent for artists and genres absent from the dataset, which may display divergent patterns of sonic characteristics and audience sentiment. Secondly, the models exclusively consider sonic characteristics, neglecting other factors that might impact fan sentiments, such as album promotion, marketing strategies, and individual preferences.

Thirdly, the models' accuracy is constrained by the precision of sentiment scores derived from the get_nrc_sentiment tool. Although the tool has been extensively employed and validated, it may not consistently capture the intricacies of fan sentiment, particularly when sentiment is more intricate or equivocal (Jockers, 2020).

Lastly, it is crucial to acknowledge that the machine learning models' performance is heavily contingent on hyperparameter selection. While efforts have been made to optimize hyperparameters for each model, it is possible that suboptimal values were chosen, resulting in compromised performance.

5.6. Implications for future research

The findings and limitations of this study suggest several avenues for future research.

Firstly, future studies could expand the analysis beyond the Reddit communities used in this study to include other online platforms, such as Twitter, Instagram, and TikTok, as well as offline settings such as music festivals and concerts. This would provide a more comprehensive understanding of fan sentiment towards music and could reveal additional factors that influence fan sentiment beyond the sonic characteristics of an album.

Secondly, future research could explore the use of alternative data sources and sentiment analysis tools to improve the accuracy and reliability of the sentiment analysis. For instance, research could incorporate data from other music streaming platforms, such as Apple Music, and Tidal, to obtain more granular data on user behavior and preferences.

Finally, future research could build on the machine learning models used in this study to develop more accurate and generalizable models for predicting fan sentiment based on sonic characteristics. This could involve using larger and more diverse datasets, incorporating additional features beyond sonic characteristics, and optimizing hyperparameters more effectively.

References

[1] E. F. Ramos, K. Blind, Data portability effects on data-driven innovation of online platforms: Analyzing spotify, Telecommunications Policy 44 (9) (2020) 102026.

[2] A. M. Beeching, Beyond talent: Creating a successful career in music, Oxford University Press, 2010.

[3] T. Frick, D. Tsekauras, T. Li, The times they are a-changin: Examining the impact of social media on music album sales and piracy, in: Annual Meeting of the Academy of Management, 2014.

[4] L. Geraghty, It's not all about the music: Online fan communities and collecting hard rock café pins, Transformative Works and Cultures 16 (2014).

[5] N. K. Baym, Tune in, log on: Soaps, fandom, and online community, Vol. 3, Sage, 2000.

[6] L. Zhou, H. Lin, W. Liu, Enriching music information retrieval using emotion detection, in: SIGIR 2011 Workshop on Enriching Information Retrieval (ENIR 2011), Beijing, China, 2011.

[7] J. Kegelaers, L. Jessen, E. Van Audenaerde, R. R. Oudejans, Performers of the night: Examining the mental health of electronic music artists, Psychology of Music 50 (1) (2022) 69–85.

[8] H. Jenkins, Fans, bloggers, and gamers: Exploring participatory culture, nyu Press, 2006.

[9] A. Marwick, D. Boyd, To see and be seen: Celebrity practice on twitter, Convergence 17 (2) (2011) 139–158.

[10] M. Duffett, Understanding fandom: An introduction to the study of media fan culture, Bloomsbury Publishing USA, 2013.

[11] M. A. Click, H. Lee, H. W. Holladay, Making monsters: Lady gaga, fan identification, and social media, Popular Music and Society 36 (3) (2013) 360–379.

[12] J. Burgess, M. Foth, H. Klaebe, Everyday creativity as civic engagement: A cultural citizenship view of new media, in: Proceedings 2006 Communications Policy & Research Forum, Network Insight Institute, 2006, pp. 1–16.

[13] B. Pang, L. Lee, Opinion mining and sentiment analysis. found trends inf retr 2 (1–2): 1–135 (2008).

[14] L. Bennett, Patterns of listening through social media: online fan engagement with the live music experience, Social Semiotics 22 (5) (2012) 545–557.

[15] E. Zangerle, M. Pichl, M. Schedl, Culture-aware music recommendation, in: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, 2018, pp. 357–358.

[16] T. Eerola, J. K. Vuoskoski, A review of music and emotion studies: Approaches, emotion models, and stimuli, Music Perception: An Interdisciplinary Journal 30 (3) (2012) 307–340.

[17] N. Koenigstein, Y. Shavitt, E. Weinsberg, U. Weinsberg, On the applicability of peer-to-peer data in music information retrieval research., in: ISMIR, 2010, pp. 273–278.

[18] L. Narnauli, Spotify datasets (2021).
URL `https://www.kaggle.com/datasets/lehaknarnauli/spotify-datasets?select=artists.csv`

[19] I. Rivera, Package ‘redditextractor’ (2023).
URL https://cran.r-project.org/web/packages/RedditExtractoR/RedditExtractoR.pdf

[20] C. Thompson, J. Parry, D. Phipps, T. Wolff, spotifyr (2020).
URL https://www.rcharlie.com/spotifyr/

[21] G. De Queiroz, C. Fay, E. Hvitfeldt, O. Keyes, K. Misra, T. Mastny, J. Erickson, D. Robinson, J. Silge, Package ‘tidytext’ (2023).
URL https://cran.r-project.org/package=tidytext

[22] M. Jockers, Package ‘syuzhet’ (2020).
URL https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html