1 Introduction and Motivation

Our project uses Google Store reviews to understand users' satisfaction levels, their perceptions of app functionality, and broader review patterns toward Spotify. By comparing linguistic and topical patterns across users with different ratings (e.g., 1-star vs. 5-star reviews), we can see how people express frustration, expectations, and appreciation through different sentiment and topic patterns. To explore this, we break our analysis into three parts: overall sentiment trends, common words and topics, and whether review ratings can be predicted from text features. Together, these approaches help us identify what users value most in a music app and which parts of user feedback Spotify should pay closest attention to. These insights can guide Spotify in improving user-experience design and refining features to better meet users' needs. From a broader social science perspective, this research also helps us understand how people communicate emotions in online environments, and how these expressions reflect larger patterns in consumer behavior, platform perception, and technology trust.

2 Data

This study uses the Spotify Google Play Store Reviews dataset[1] sourced from Kaggle, consisting of 3,377,423 reviews. The dataset includes review text, user star ratings, and basic metadata such as timestamps and other app-related information.

We refined the raw dataset through several filtering steps to ensure quality and relevance. First, we restricted the data to reviews posted between November 15, 2019, and November 15, 2023, capturing the most recent 5 years of user feedback. Next, we removed reviews containing words longer than 15 characters to reduce noise from corrupted or anomalous text entries. These filtering steps produced a final working dataset of 1,669,701 reviews ready for further analysis.

We applied a review-length segmentation strategy using raw word counts to support more granular NLP analysis, since very short reviews often lack sufficient linguistic content for advanced analysis. Reviews were categorised as Very Short (1–3 words), Short (4–6 words), Medium (7–10 words), and Long (10+ words). The final distribution was: Very Short (614,051), Short (269,962), Medium (197,652), and Long (587,948). To streamline downstream analysis, we then consolidated these categories into two groups: Short (1–6 words) and Long (7+ words). This broader classification efficiently separates text suitable for targeted analysis and supports tailored model training for each length category later.

Text preprocessing was conducted using the BERT tokenizer, which effectively handles informal language features common in user-generated reviews, including slang and misspellings.

[1] Source: Kaggle, *3.4 Million Spotify Google Store Reviews*, 2024. Available at: https://www.kaggle.com/datasets/bwandowando/3-4-million-spotify-google-store-reviews

A key limitation of the data is the inability to fully filter out non-English reviews. Due to the scale and linguistic diversity of the dataset, removing multilingual content was technically challenging. As a result, some non-English reviews remain, which might affect interpretability in topic modelling and may reduce the accuracy of classification tasks by introducing linguistic noise into the models.

3 Methods

3.1 Sentiment Analysis

To ensure sufficient textual content for reliable sentiment analysis, we retained only "Medium"(7–10 words) and "Long" (>10 words) reviews, excluding potentially noisy short entries. We selected VADER to perform sentiment analysis, for its proven effectiveness with short, informal text such as user reviews. VADER's compound score, which ranges from –1 (most negative) to +1 (most positive), served as a unified measure of overall sentiment for each review.

Finally, we categorized the continuous compound scores into three distinct labels: "Negative" (score ≤ –0.1), "Neutral" (–0.1 < score < 0.1), and "Positive" (score ≥ 0.1). This threshold-based approach reduces ambiguity around neutral values while maintaining clear and interpretable sentiment distinctions.

3.2 Topic Modeling

For topic modeling, we first conducted a general comparative token analysis before examining topic-cluster patterns across different subsets of the data. For the comparative token analysis, we extracted 1-star and 5-star reviews as two subsets and applied bag-of-words and TF-IDF methods to identify the most common and distinctive words in highly negative versus highly positive user feedback.

Building on this, we then applied BERTopic to three review subsets, including good reviews (4–5 stars), mid reviews (3 stars), and bad reviews (1–2 stars), and set a limit of 50 topics for each model to extract topic clusters within each group. This approach allows us to clearly identify what users praise, what they complain about, and which aspects of the app require the most improvement.

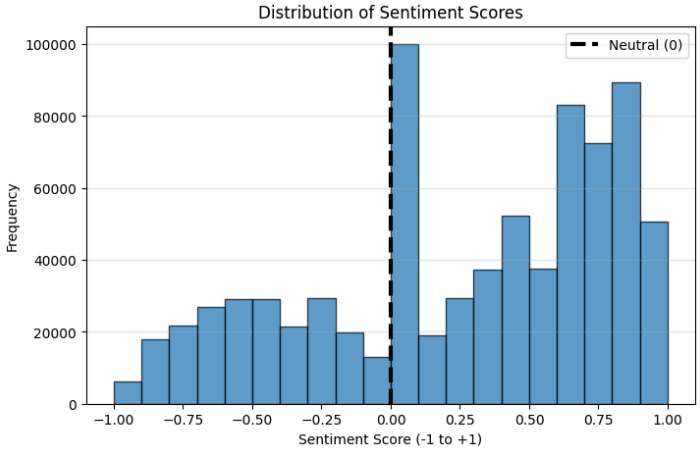
3.3 Classification

To predict review ratings from text, we trained logistic regression models based on two feature representations. The first was TF-IDF, which represents text based on word frequency patterns. The second used was Word2Vec, which represents each review using the average of its word embeddings. To understand how review length influences classification accuracy, we trained models on the full dataset and separately on categorized long and short reviews. Model performance was evaluated using standard evaluation metrics, including overall accuracy, precision, recall, F-1 score for each rating level, and weighted average F-1 Score. Using those metrics, we can compare how different feature representations and review lengths affect a model's ability to classify star ratings from text.

4 Results

4.1 Sentiment Analysis

Figure 1. Distribution of Review Sentiment Score

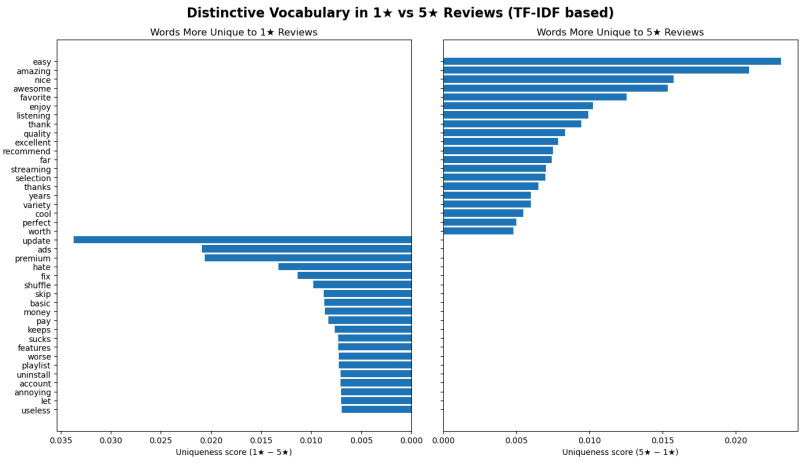


Overall, the compound score of the reviews shows a bimodal distribution, which is strongly skewed towards the positive end. The mean sentiment score is 0.251, and the median is 0.3753, both significantly positive, presenting a generally favorable sentiment towards the Spotify app.

The histogram exhibits two primary peaks: a very sharp, dominant peak clustered around the Neutral marker, and a second broader mode clustered in the highly positive range (between 0.7 and 1.0). This pattern suggests that the majority of the review texts are categorized as positive or neutral. The 75th percentile at 0.726 and the 25th percentile at -0.123 also confirm the high frequency of positive reviews.

4.2 Topic Modeling

Figure 2. Distinctive Vocabulary in 1-star and 5-star Reviews (TF-IDF based)



For the comparative analysis, using bag-of-words, we found that the top five words in 1-star reviews were premium, ads, update, playlist, and fix, whereas the top five words in 5-star reviews were premium, easy, amazing, listening, and ads. Using TF-IDF, the distinctive vocabulary in 1-star comments included terms such as update, fix, shuffle, skip, money, uninstall, account, and annoying, while 5-star reviews featured more positive words like easy, amazing, favorite, and enjoy, as shown in Figure 2.

For the topic modeling, the top two topic clusters remained consistent across all subsets and centered on general terms such as app, music, song, and Spotify. However, beyond these broad themes, the specific topic clusters differed substantially depending on the review type.

For 1–2 star reviews, prominent topics involved account login and password issues, offline or internet connection problems, new updates, and the lyrics display function. Users frequently mentioned concerns such as limited skip counts without Premium, unexpected account logouts, connectivity failures, and lyrics not appearing for songs.

For 3-star reviews, topics focused on internet/update connectivity, Premium features and ad frequency, podcast episodes, and the Instagram Stories sharing function. Many users expressed that they like the app overall but find the ads overwhelming without Premium, that podcast features feel less smooth even when shows are exclusive to Spotify, and that they are unable to share music to Instagram Stories.

For 4–5 star reviews, the main topics included ease of use, Premium value, and playlist/shuffle/queue functionality. These reviews featured more positive comments such as “this is the best app ever,” praise for the ability to create personalized playlists, and satisfaction with Spotify’s music and podcast selection.

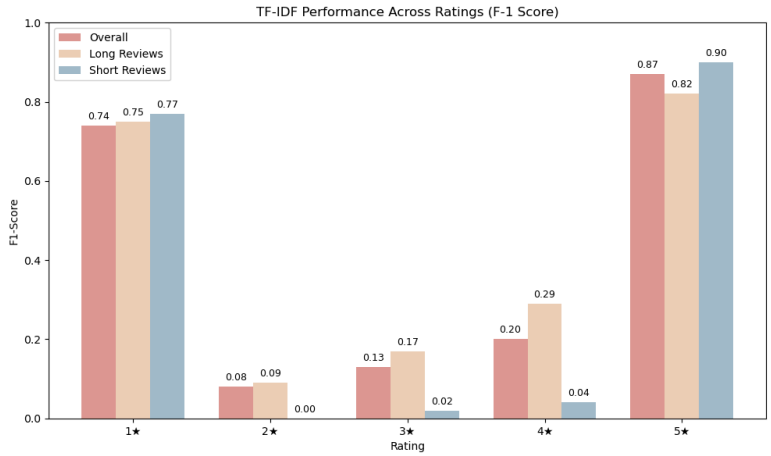
4.3 Classification

Table 1. Comparison of Accuracy and F-1 Score Between TF-IDF and Word2Vec

Category	Accuracy (TF-IDF)	F1-Score (TF-IDF)	Accuracy (Word2Vec)	F1-Score (Word2Vec)
Overall	0.75	0.70	0.71	0.64
Long text	0.67	0.62	0.63	0.56
Short text	0.83	0.77	0.80	0.75

Table 1 summarizes the accuracy and weighted average F-1 score for both feature representations. Across all categories, TF-IDF consistently produced higher accuracy and F-1 score than Word2Vec. Also, for both models, they demonstrated overall better performance in classifying short reviews than long reviews.

Figure 3. Distribution of F-1 Score Across Ratings for TF-IDF

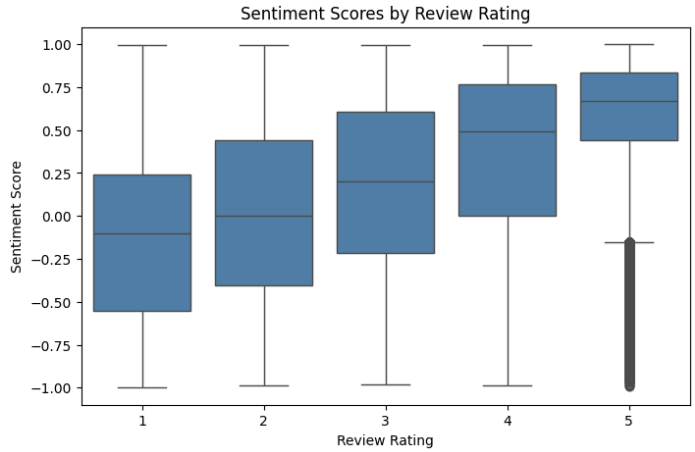


According to Figure 3, models with TF-IDF as feature representation performed best on 1-star and 5-star reviews, with F-1 scores consistently high across all categories, regardless of whether reviews were long or short. However, performance dropped significantly in the middle ratings, from 2-star to 4-star, where the model struggled to classify them accurately.

5 Interpretation

5.1 Sentiment Analysis

Figure 4. Boxplot of VADER Sentiment Scores By User Review Ratings



The comparison of VADER compound sentiment scores against user star ratings reveals a strong positive correlation, confirming VADER's ability to capture sentiment from text reviews effectively. As star ratings increase from 1 to 5, median sentiment scores rise consistently. Low ratings (1-2 stars) produce appropriately negative or neutral sentiment scores, while 3-star reviews already shift positive.

This positive trend is amplified for high ratings, where 4 and 5-star reviews exhibit median scores of 0.5 and 0.7, respectively, with their interquartile ranges firmly in the positive domain.

The alignment between ratings and text sentiment suggests that most users write reviews that genuinely reflect their experiences. Satisfied customers use positive language, while dissatisfied ones express clear negativity. This consistency indicates that star ratings are not arbitrary; they are backed by emotional responses captured in the text.

However, a critical finding emerges: negative sentiment outliers exist even within 5-star reviews. These cases represent a mismatch where the users' highest explicit rating contradicts the negative polarity detected in the written text. This mismatch shows the complexity of real user sentiment. Some customers give top ratings despite expressing frustration in their reviews, often praising the product overall while criticizing peripheral aspects like ads.

5.2 Topic Modeling

The topic modeling analysis highlights both the aspects of the app that users complain about and the features they consistently praise, providing valuable insight into customer pain points and areas of satisfaction. Across review patterns, both the comparative token analysis and the topic modeling show clear distinctions by rating group. In 1–2 star reviews, users describe broadly negative experiences and focus heavily on unresolved pain points. In 3-star reviews, users acknowledge the app's strengths but also provide concrete suggestions for improvement. In 4–5 star reviews, users increasingly use positive adjectives and expressive praise, indicating an overall shift toward emotional appreciation rather than evaluation of specific features.

In terms of specific pain points and user favorites, one theme that emerges across all reviewer groups is the importance of Premium features and ad frequency. A common sentiment among low-rated reviews is frustration that ads cannot be skipped without subscribing to Premium, whereas many 5-star reviewers note that Premium is "worth it" precisely because it removes ads and allows unlimited skips. This contrast highlights differing user values and perceptions of whether the benefits of the paid service justify the cost.

Beyond ads and Premium, several functionality-related issues appear frequently. Negative sentiments often mention internet connectivity problems, new app updates, and password or login issues, all of which relate to the app's compatibility with different devices or system environments. Users also raise experience-oriented concerns, such as lyrics not displaying (especially for non-English songs), lack of smoothness in podcast playback, difficulty sharing songs to social media, and shuffle mechanics that limit users' control over their listening choices. These insights point directly to users' core pain points and can guide Spotify in refining specific features to deliver a more seamless and satisfying user experience.

5.3 Classification

Across all review lengths, TF-IDF performed better than Word2Vec, suggesting that rating prediction relies heavily on specific words and phrase patterns that strongly signal star ratings, such as “love”, “terrible”, rather than average word embeddings from Word2Vec, in which it lost details and therefore performed worse. Then, both models achieve higher accuracy on short reviews than on long reviews because short reviews always contain direct sentiment words, such as “perfect” and “hate”, which enable more accurate classification. However, for long reviews, it contained a mix of both positive and negative phrases and nuanced language, making classification much more difficult.

When examining TF-IDF performance across rating patterns using F-1 scores, it performs well on predicting 1-star and 5-star reviews but poorly on middle-range ratings. The rationale for these findings is that extreme ratings, such as 1-star and 5-star, have stronger and clearer sentiment signals. For example, 1-star reviews always contain direct negative words or phrases, such as “worst” or “disappointed”, while 5-star reviews contain consistent positive language, such as “love” and “fantastic”, which makes TF-IDF easier to capture patterns. However, for middle range star reviews, from 2-star to 4-star, they always contain mixed and nuanced sentiments, such as “good product overall but the ads are annoying”, which made the model difficult to distinguish between them and therefore lower their F-1 scores. What’s more, the weighted average F-1 score appears strong with a score of 0.70, but the score was inflated by the model’s strong performance at predicting the extremes. High F1-scores for 1-star and 5-star reviews elevate the overall average, even though the model performs much worse on mid-range ratings.

6 Implications

Across all analytical messages, our findings suggest a consistent theme: while star ratings can provide a general understanding of users’ experience on Spotify, it is crucial to delve into the specific review text for richer, more comprehensive feedback. Sentiment Analysis shows that while most reviews align with star ratings, some 5-star reviews still contain negative language about the app. This indicates that users will always reward Spotify’s overall value while simultaneously expressing frustration with specific parts of the app. This hidden disconnection suggests that Spotify cannot just look at general ratings; it also needs specific textual information for improvement.

Then, the findings from Topic Modeling tend to provide us with rich information about the topics and themes that users are most satisfied and disappointed with, which enables the platform to deliver very targeted interventions. Findings show that 3-star reviews provide some of the most abundant material for actionable insights. Those users always express both appreciation and disappointment. Applying strategies aligned with their pain points is the main way to convert these groups into highly satisfied and

loyal users. Therefore, it is really important for Spotify to identify the main pain points through topic modeling analysis and implement targeted interventions to enhance the user experience.

The classification results tend to reinforce the findings from the previous two analytics approaches. Models show good predictive accuracy for extreme star reviews but struggle with the middle range. This pattern mirrors what we find in Sentiment Analysis and Topic Modeling: extreme positive and negative experiences and feedback are expressed directly, while moderate experiences are always more complex, suggesting that there should be greater effort to understand middle-range text reviews.

Taken together, we proposed two actionable business strategic opportunities. First, the product team should prioritize issues that repeatedly surface across sentiment, topics, and classification errors, and track app feature improvements in specific areas, such as technical reliability, lyrics support, ad reduction, and podcast usability. Spotify could address these concerns by adding clearer FAQ sections focused on common technical problems, offering step-by-step guides to resolve bugs, and expanding customer service options, such as chat support, so users can resolve issues more easily. By incorporating those supports, users could have a better experience and therefore avoid the churn caused by the problems they encounter. Second, beyond looking only for the most positive and negative feedback, pay attention to middle-range reviews, which always offer the highest return on investment for product improvement. Mid-range feedback often includes early warnings about bugs, confusing UI workflows, or minor frustrations. These issues may not affect overall ratings yet, but ignoring them risks long-term dissatisfaction. Prioritizing fixes for mid-range feedback can reduce negative sentiment before it spreads.