Successful Album or Forgotten: Predicting Album Ratings Using Sentiment Analysis

By Alex Sapinoso, Alison Wilbur, and Sofia Villalpando

Statistics 133 Final Project

Introduction

Music has always been known to display emotions through lyrics, writing, and ratings. Pitchfork is an online music publication that publishes album reviews for artists across all genres of music, publishing five reviews per business day. The website was founded in 1996 and originally covered alternative and independent music, eventually expanding to cover more genres and reach a larger audience. Pitchfork is widely considered to be the "Most Trusted Voice in Music", and many music fans turn to the platform to determine whether or not new albums are worth a listen (Cabral, 2021). As such, Pitchfork is extremely influential in the music scene and has the potential to make or break artists' careers.

The digital world also relies on these reviews for individuals to discover new music, streaming platforms to personalize their consumers' playlists, and create algorithm-based recommendations to users across several platforms. Music is something that is shared around the world and reviews have become an essential part of consumerism, customer satisfaction, and personal enjoyment.

We aim to understand the underlying trends across the many album reviews on Pitchfork's website. Using the Kaggle data set "Pitchfork Reviews: Music Critiques Over the Years" with over 25,000 album reviews, we aim to use text mining methodologies to perform sentiment analysis and uncover the factors that influence whether an album's review is positive or negative.

Methodology

I. Hypothesis

Ratings have now become an influential factor in music taste and are prevalent on every streaming platform worldwide. These reviews come from humans who have varying tastes, emotions, and opinions on music. Based on the literature review and our personal love for music and their reviews, we suspect that there will be a surplus of positively rated albums based on sentiment analysis and data cleaning. We aim to uncover any trends in these reviews and use sentiment analysis to see how many albums are rated highly.

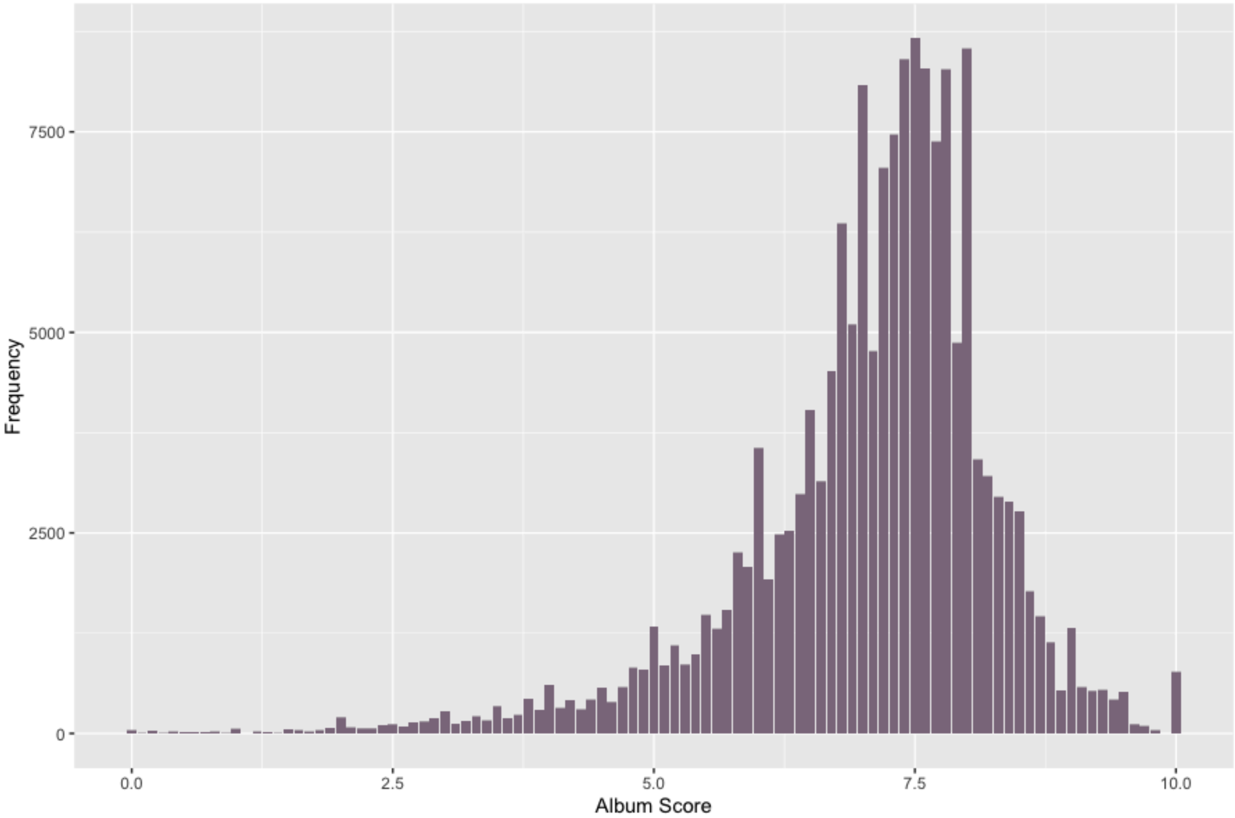
II. Data Retrieval

The Pitchfork Reviews were taken from the Kaggle dataset "Pitchfork Reviews: Music Critiques Over the Years", collected by user "timstafford" on Kaggle, which has a collection of various albums from different genres and ranging release dates from 1952 to 2023. Each album has been archived by artist name(s), album name, album score (rated on a scale from 0 to 10), the album's release year, the author of the review, the album's genre(s), record label, the date the review was published, the summary of the album review, the full album review, and whether the album was awarded "Best New Reissue" or "Best New Music". Our main variables of interest were the album scores, the album genres, and the album reviews. Examining the relationships between these variables allows for a thorough understanding of the various factors that contribute to an album's overall rating on Pitchfork.

III. Exploratory Data Analysis

The first step toward understanding the underlying trends across Pitchfork reviews is to examine the overall distribution of ratings given to all albums throughout the years on the website. In doing so, it becomes clear whether the distribution of album ratings is skewed in a particular direction, which could indicate bias in Pitchfork's criteria when rating albums.

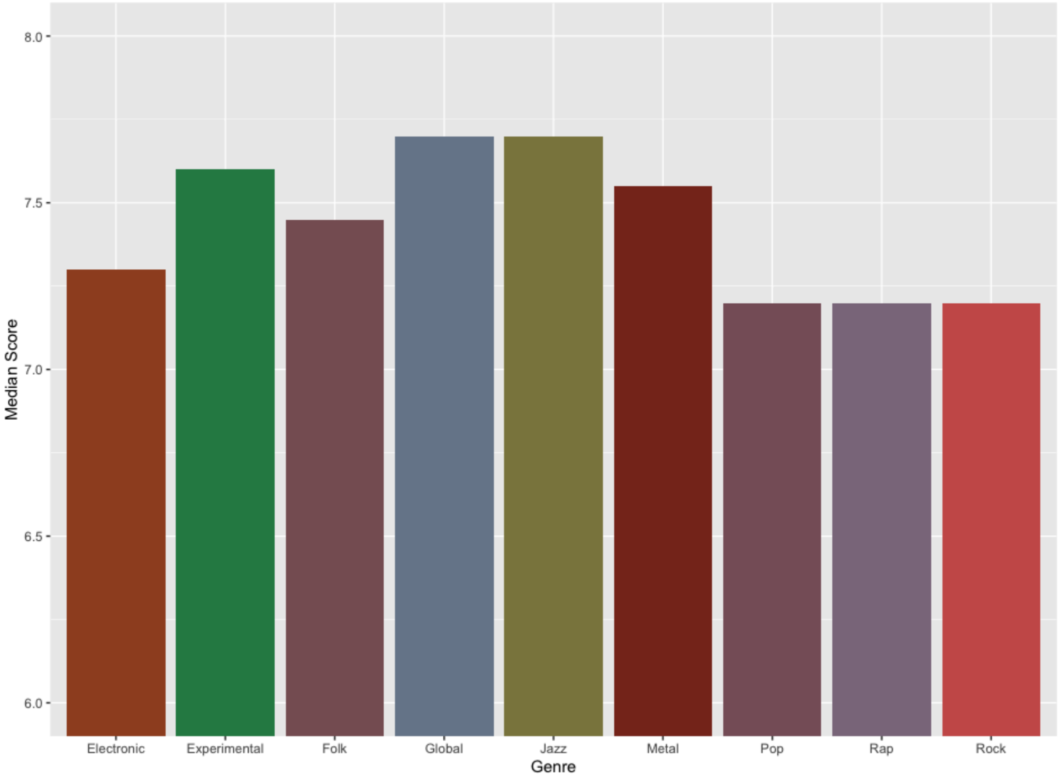
Figure 1. Distribution of All Album Scores



Based on the distribution of the album scores, we see that the median score is 7.3, with most reviews scoring between 6 and 8. This left-skewed distribution of scores indicates that Pitchfork is selective when it comes to the albums they choose to publish reviews on and that they generally are not too harsh in their ratings of albums. Given that Pitchfork has been around for several decades and that the landscape of the music industry has significantly changed since the beginning of the platform, authors on Pitchfork may select the albums they review based on whether an album or artist is generating buzz among the music community or whether an artist has already established a strong presence within the community.

Another important step toward understanding the underlying trends across Pitchfork reviews is to examine how ratings vary across different genres. Identifying differences in ratings based on album genres could indicate whether Pitchfork reviews tend to be biased toward any particular genre.

Figure 2. Distribution of Median Score per Genre

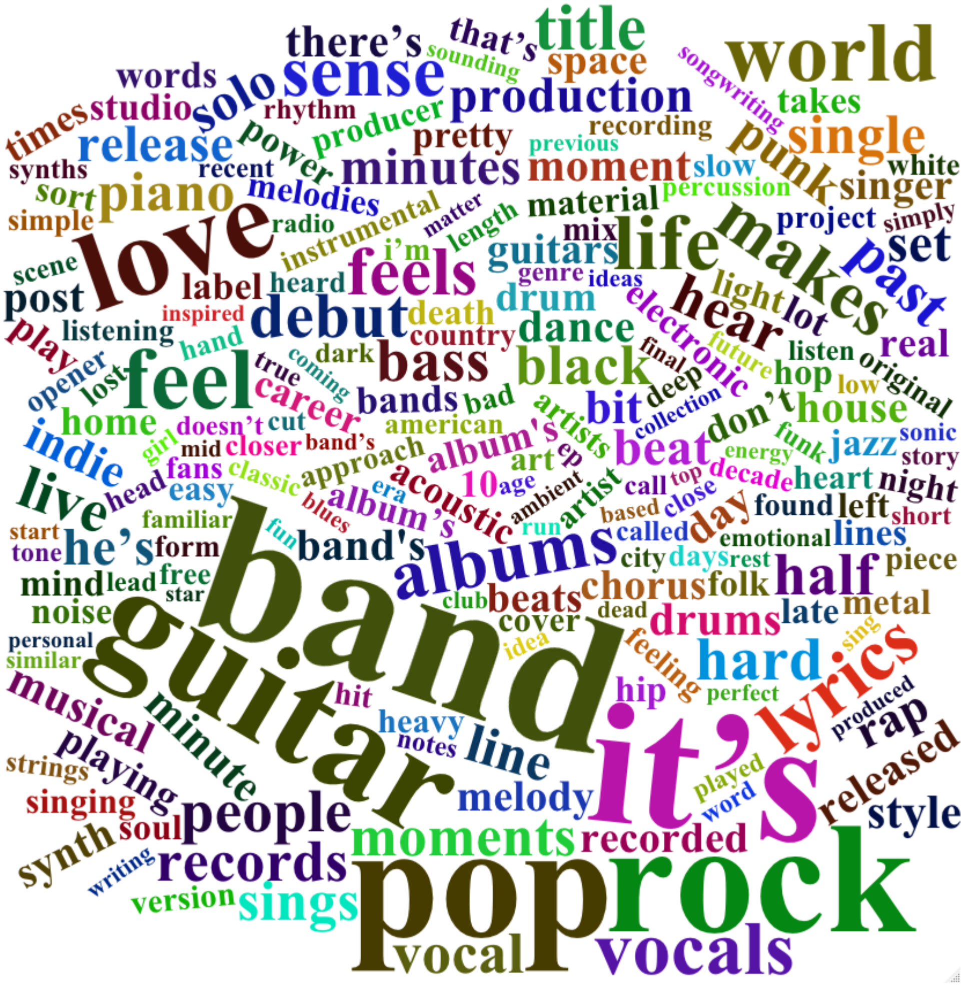


From the distribution above, we see that the median score for each genre falls between 7 and 8. Examining the medians of each group ensures that any outlier album ratings do not affect our findings. We see that the genres with the highest median album ratings were Global, Jazz, and Experimental. Since the median scores are similar across all genres, this indicates that Pitchfork reviews favor no particular genre. It should be noted that, for the top 3 genres with the highest median scores, these genres tended to have fewer reviews than those with lower median scores, such as Pop, Rap, and Rock. Since the frequency of reviews are more heavily skewed toward these genres, there is a larger range of albums to review and therefore a wider range of scores than for Global, Jazz, and Experimental.

Results

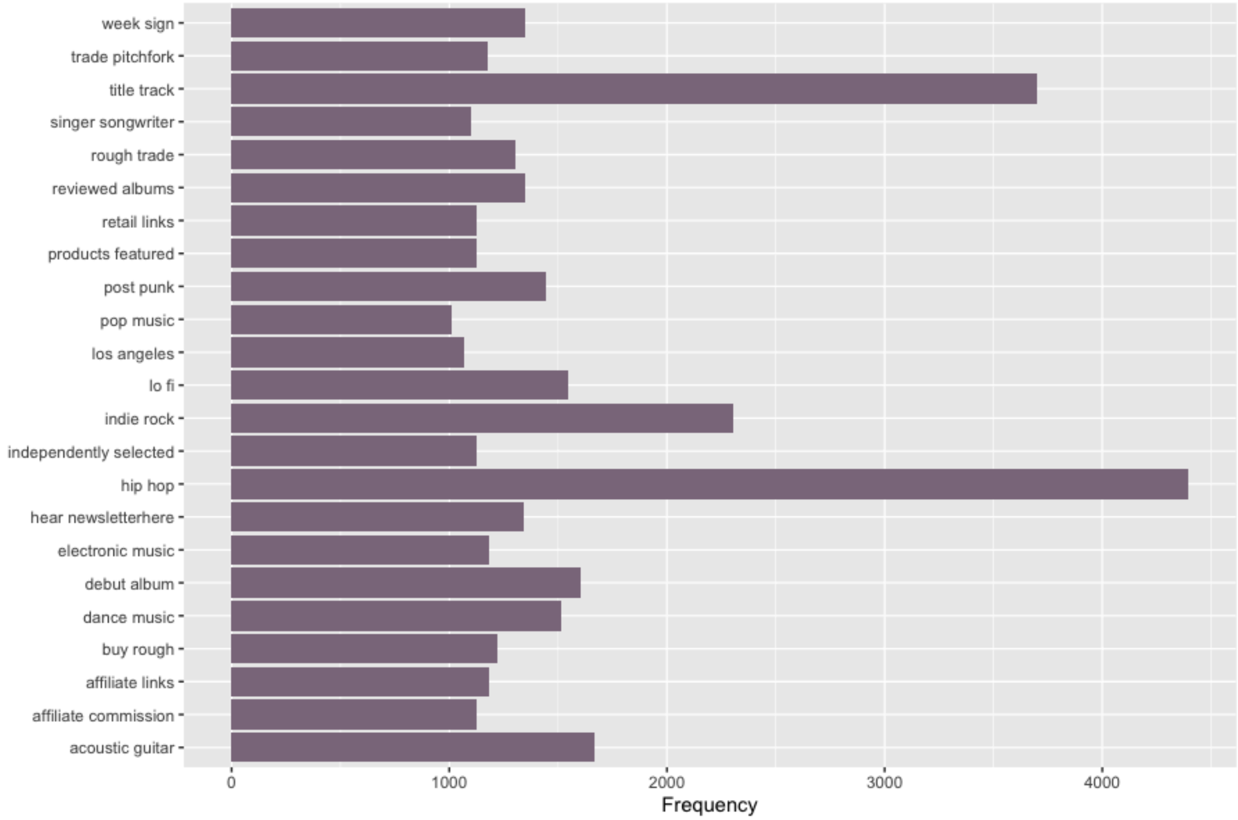
To get a better understanding of the text data we were working with, we began by converting the data from the reviews column into a tidy text format, tokenizing the terms, and removing stop words from the text. We began with the following word cloud, where the most frequently used terms are larger than terms that are used less frequently.

Figure 3. Word Cloud



Based on the largest terms in the word cloud, we see that the most frequently used terms are musical, particularly relating to genre. In addition to musical terms such as “guitar”, “production”, and “beats”, there are many words that are used to describe the emotions an album can evoke in the listener, with terms such as “feels”, “moments”, and “sense.”

Figure 4. Most Frequent Bigrams

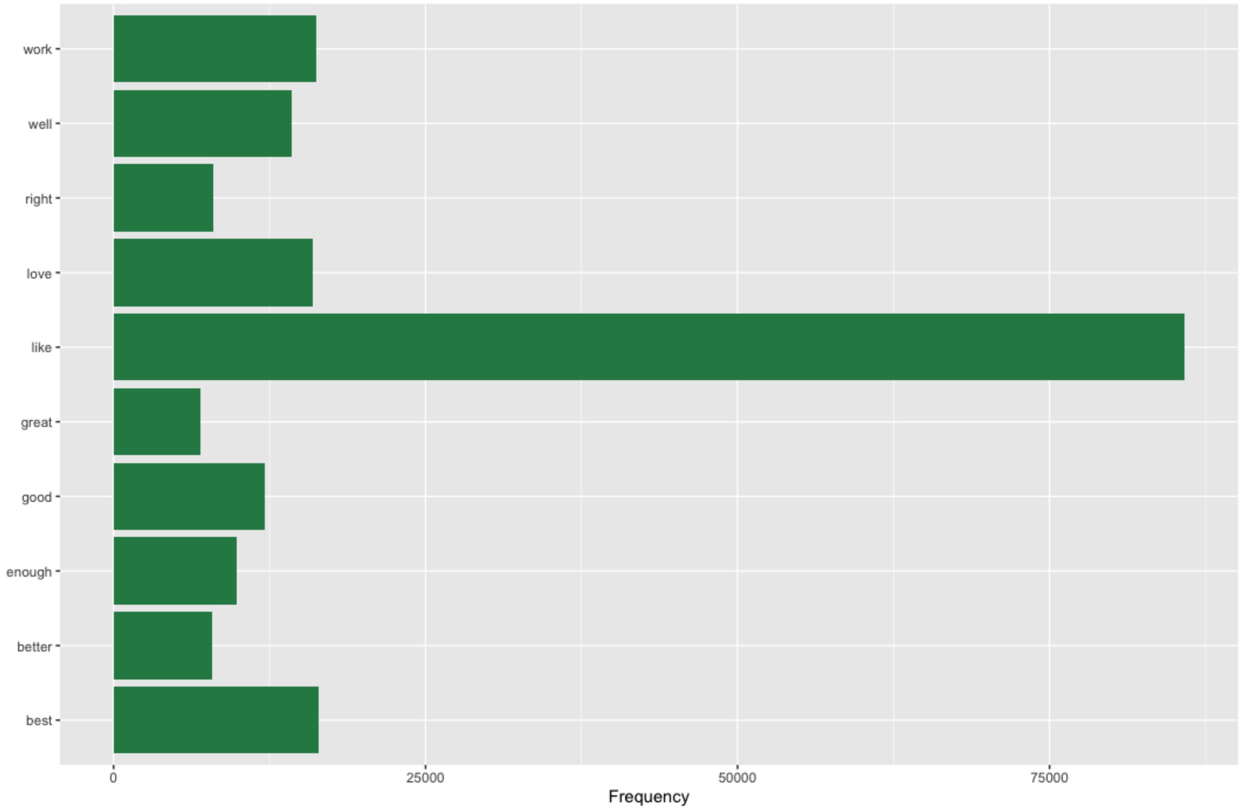


Looking closer at the most characteristic terms, the most frequently used bigrams are shown in Figure 4. We see that the two terms that appear together the most, after cleaning the text data and removing stop words, are “hip hop”, which makes sense given that many reviews are given to albums of the Hip Hop genre. The second most frequent bigrams are “title track”, which is often used to kick off an album’s review when describing the various songs on an album. A majority of the most frequent bigrams are subgenre descriptors such as “singer-songwriter”, “post-punk”, and “indie rock”, that are often used to describe the album in more detail than their main categorized genre.

I. Sentiment Analysis

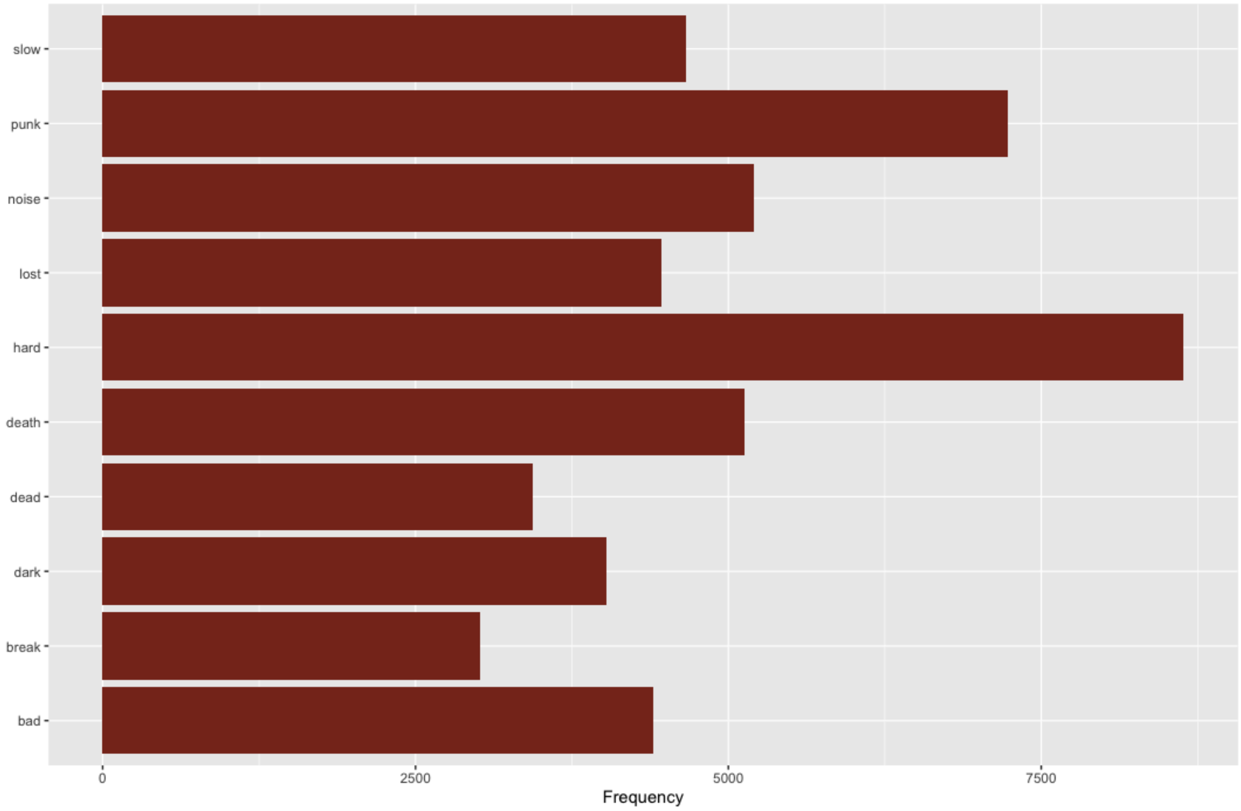
To perform sentiment analysis on our data set, we converted the text data into a tidy format using the tibble() function, extracted sentiments from the Bing lexicon, and followed the procedure of tokenizing the tidy text data into words and performing an inner join with the Bing sentiments. We began by analyzing the most frequent terms associated with positive reviews.

Figure 5. Most Frequent Positive Terms



From the distribution of the top ten most frequent positive terms, we see that the word “like” is used far more frequently than any of the others. In the context of the album reviews, the word “like” was used to describe what the author liked about each album, and was additionally used to compare elements of the albums to other musical elements that are often characteristic of specific genres. Since Pitchfork writes specifically for the online community of music fans, it is very common for album reviews to contain sections that are very comparative in nature. The word “like” is often used for the purpose of allowing the reader to understand what a new album sounds like based on a description of a song or album they are most likely familiar with. Other frequently used positive terms include “best”, “work”, “well”, and “love”, which are often used to describe what the author of the review argued were the best elements of well-rated albums, as well as what made the album good from their perspective.

Figure 6. Most Frequent Negative Terms

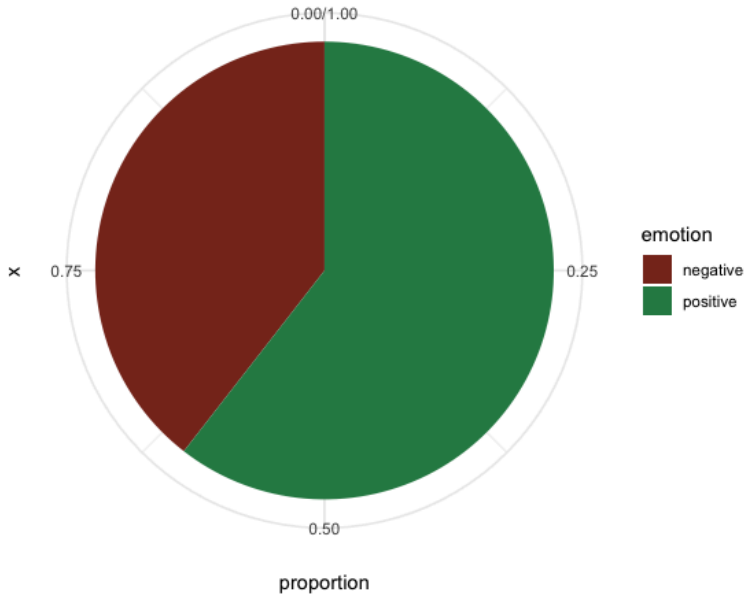


From the distribution of the top ten most frequently used negative terms, we see that the most frequently used term is “hard”, which was used in phrases such as “hard to know” and generally used by reviewers to express confusion over the album. Negative reviews on Pitchfork tend to criticize the content of the album, whether it be confusion over the creative direction the artist was going for on their album or a lack of interesting changes in sound as the album progresses. The latter is often reflected with the words “slow” and “lost”, which are also the most frequently used negative words. The rest of the negative terms depict the more negative themes that albums can be based around, with words such as “dark” and “death.”

II. Distribution of Sentiments

After examining the most frequent terms that were categorized as positive and negative, comparing the overall proportions of positive and negative reviews in the data set can give us a better understanding of how a review’s sentiment affects the distribution of album ratings.

Figure 7. Distribution of Positive and Negative Reviews

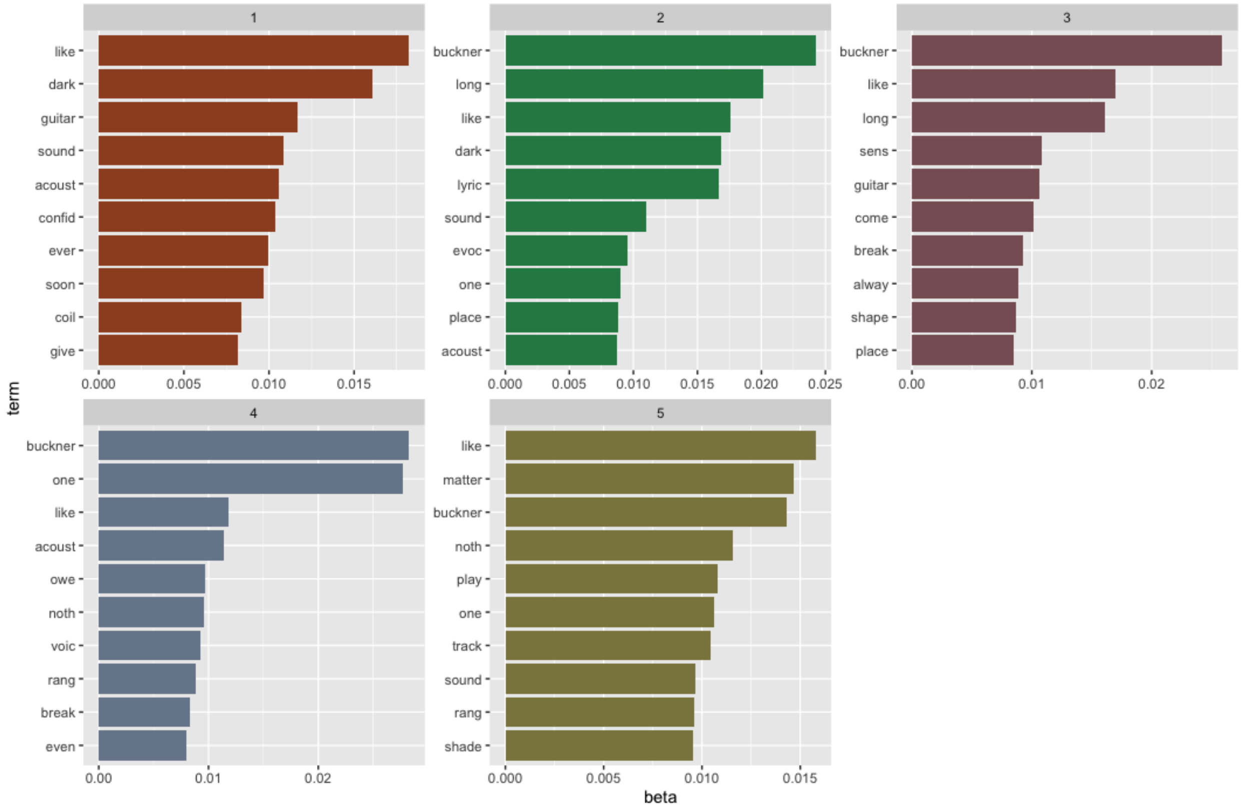


From the distribution in Figure 7, it is clear that the majority of Pitchfork reviews have an overall positive sentiment. When comparing this distribution of sentiments with the distribution of album scores across the data set, it is clear that our overall findings properly reflect the relationship between a review's sentiment and the album's score.

III. Topic Modeling

Topic modeling was used on our large dataset of album reviews as one approach to determine the natural groups that the reviews could be clustered into. Using Latent Dirichlet Allocation (LDA), the topic model for the Pitchfork reviews used the text context to differentiate between various groups and split them into different topics. We created a five topic model (k = 5), with the top ten most frequent terms of each topic depicted in the figure below.

Figure 8. Latent Dirichlet Allocation Topics



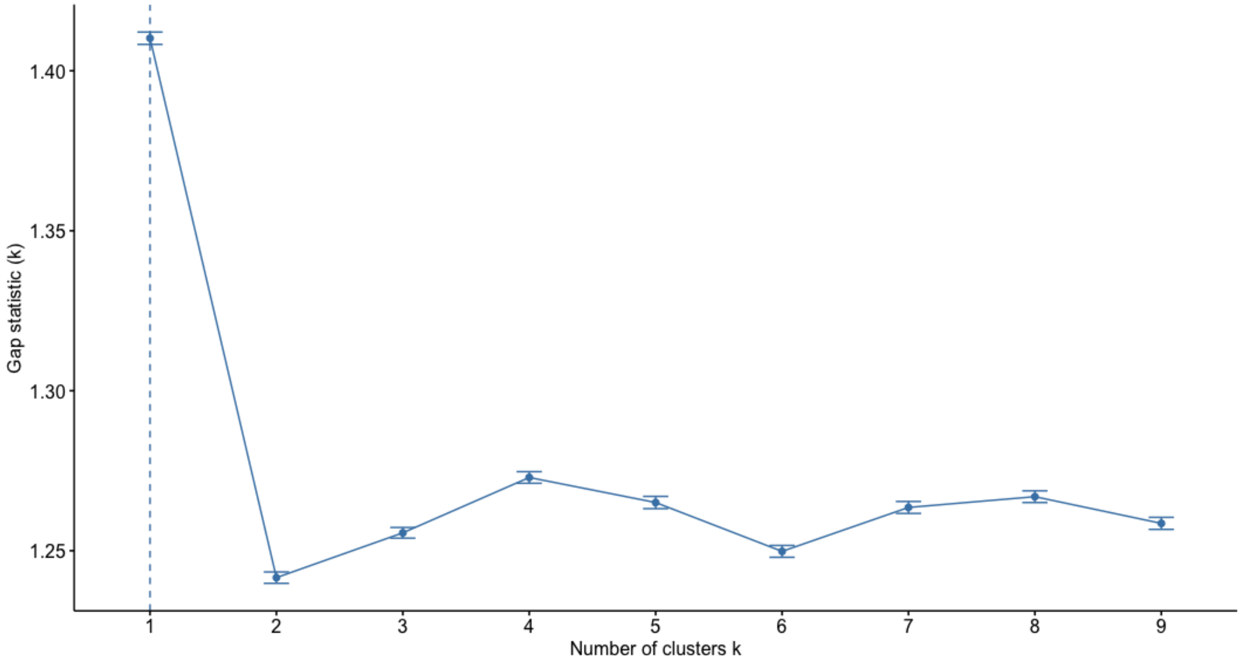
Based on the most frequent words from each topic, we can infer which topics refer to the various groups of reviews. Since Topic 1's most frequent words include "dark", "guitar", and "acoustic", we can infer that these reviews fall under the "Rock/Metal" genre category. Topic 2's most frequent words include "long", "lyric", and "acoustic", which indicates that these reviews fall under the "Folk/Country" genre category. The most frequent words in Topic 3 include "sense", "break", and "always" which could indicate that these reviews fall under the "Jazz" category given that reviews for this genre use more abstract terms to describe the music. Topic 4's most frequent words include "one", "voice", and "range", which indicates that these reviews fall under the "Pop/R&B" genre category since pop albums most notably discuss the artist's use of their voice. Finally, since Topic 5 includes the most frequent terms "matter", "track", and "shade", this could indicate that these reviews fall under the "Rap" category since one of the characteristic themes of rap artists' lyrics is often described as "throwing shade."

Another notable aspect of the LDA output is that the term "Buckner" is seen throughout most of these topics as one of the most frequent terms. This likely refers to the artist Richard Buckner, whose reviewed albums all fall under the Rock genre category. One reason Buckner would be mentioned across reviews for different genres could be that some Pitchfork reviews include information about newly released music announcements in the subheadings and that there are many Pitchfork articles called "News in Brief" which discuss new music releases, combining all genres into one review article.

IV. Cluster Analysis

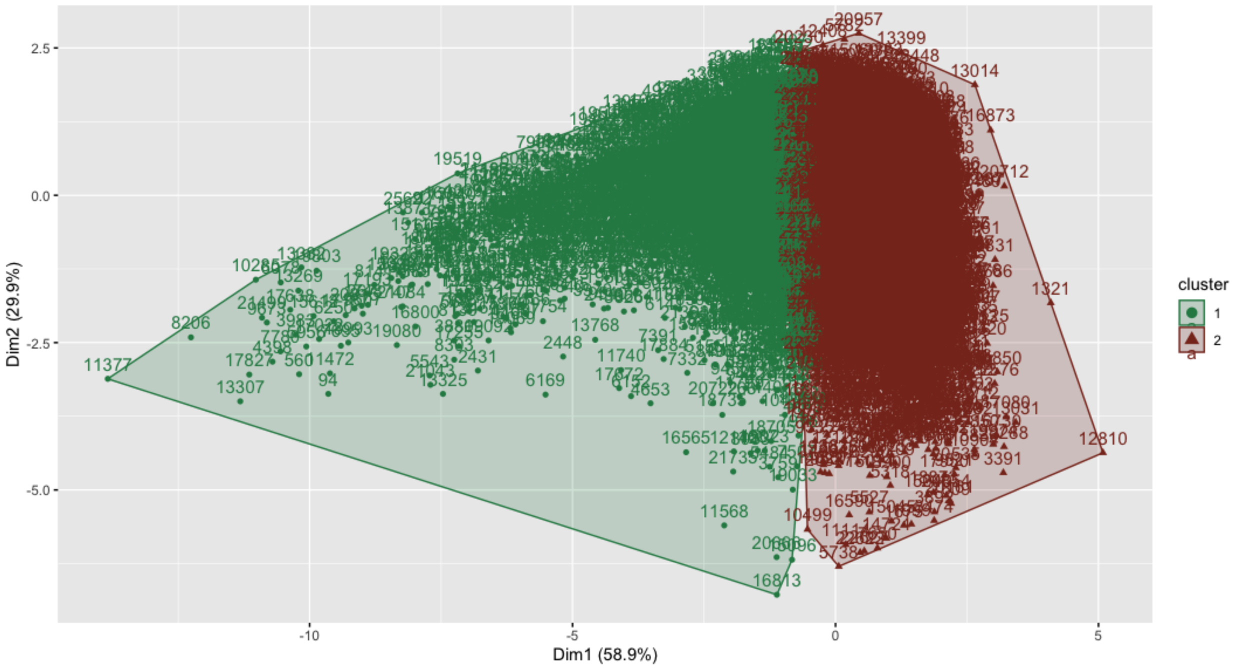
Another approach we used to determine the natural groups that the reviews could be clustered into was k-means cluster analysis. In order to determine whether sentiment scores had any effect on how reviews were classified, we appended the positive and negative sentiment scores from the NRC lexicon to the reviews data set. After scaling the numerical data, we used the Gap Statistic method to determine the optimal number of clusters to use for our k-means clustering. This method compares the total intracluster variation for different values of k with the expected values that would fall under a distribution with no obvious clustering.

Figure 9. Optimal Number of Clusters using Gap Statistic Method



Based on the visualization above, since the bend in the plot bends at k = 2, the optimal number of clusters to use for our data is two. Using this parameter for the kmeans() function with nstart = 25, we obtained the following visualization of the clustering for our reviews.

Figure 10. K-Means Clustering with k = 2



Based on the distribution of observations in each cluster, we can infer that the reviews have been clustered into reviews with positive sentiments and reviews with negative sentiments.

Table 1. Median Scores Based on Clusters

Cluster	Score	Negative NRC Score	Positive NRC Score
1	8	32	46
2	7.1	17	28

Examining the median scores across each genre, we can see that reviews in cluster 1 (positive sentiment) tend to have a higher median score than those in cluster 2 (negative sentiment).

V. Bigrams

Bigrams in music are often words that help enhance the beat of a song or get a point across, such as “la la.” However, we wanted to explore how bigrams have an effect on reviews of albums. We found the most frequent bigrams in the reviews and looked into discovering how negative negations, particularly looking at “not,” skewed the distribution of positive and negative ratings. After doing an extensive analysis, we found that more positive phrases were preceded by the word “not,” ultimately implying that there were a higher number of positive reviews than once thought. After this analysis, we constructed a Markov Chain to visualize the grammar of our bigrams as a Markov Chain. These core terms are what connect many of the reviews with album ratings.

Figure 11. Most Frequent Negation Bigrams

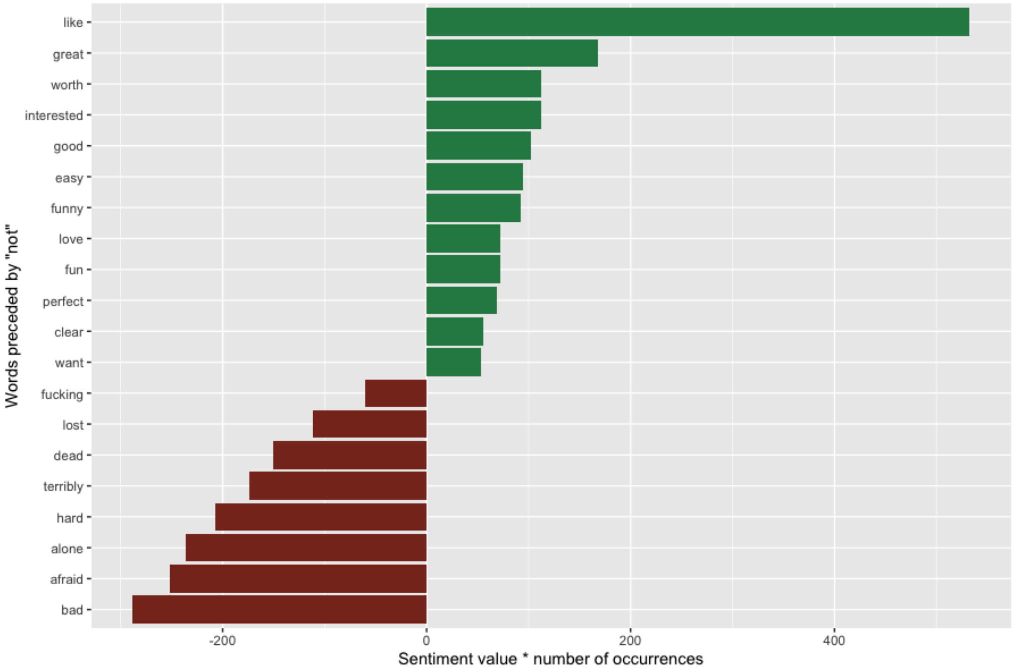
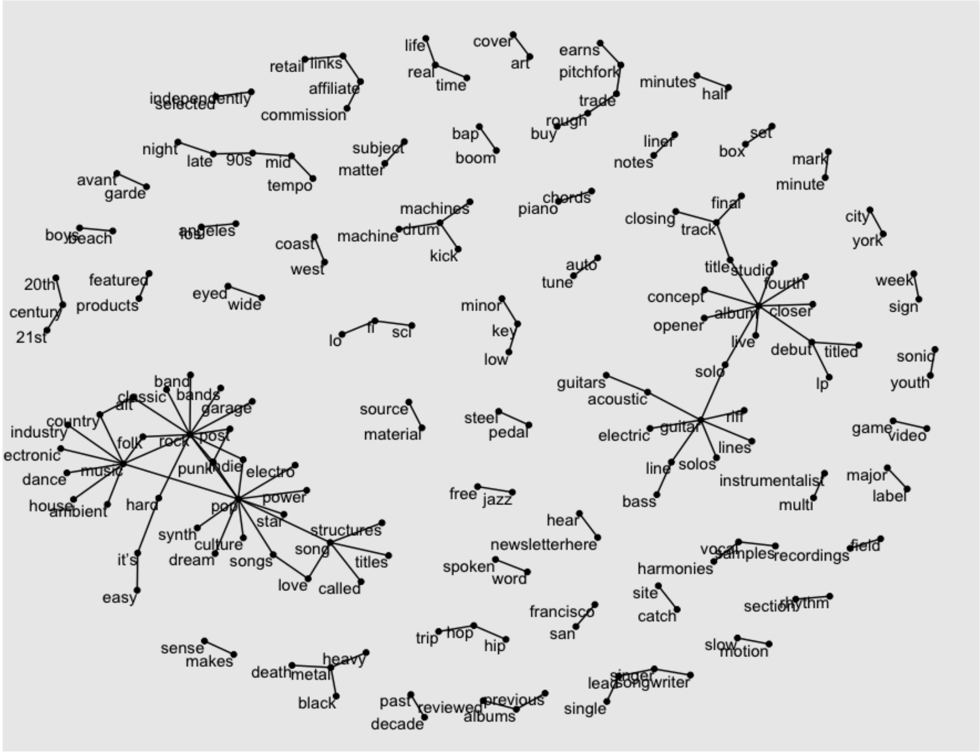


Figure 12. Bigrams Network



VI. Predictive Model

Table _. Baseline Accuracy

0	1
0.5223389	0.4776611

The baseline accuracy was about 47.8%. After using random forest with ntrees = 50, we achieve a baseline accuracy of 63.8% when predicting whether an album will be high scoring or not. This increase in accuracy implied that our model was successful, but could be improved. However, as we tried various models, we ran into the issue of our models becoming too computationally expensive to execute, so we ended up using this model as our final predictive model.

Discussion

While conducting our research, we were able to make several fascinating discoveries using text mining techniques tailored to album reviews, through the use of sentiment analysis, Latent Dirichlet Allocation, exploratory data analysis, bigrams, and topic modeling. We also used several text mining cleaning methods, including removing stopwords, whitespace, punctuation, and counting the total words post cleaning. For example, word removal allowed us to see that “dark” and “death” were the most frequent negative words in reviews, where before it would have been “sad.”

Overall, we were able to find frequent terms that were in both highly rated and negatively rated albums. There were more positive rated albums than negative after our analysis, showing that genre variety had little effect on what type of albums were rated highly or not.

To further examine this topic, we think that conducting a trigram analysis would be extremely helpful. Also diving deeper into the data by inspecting the lyrics of the albums in the dataset and comparing them to the reviews would help show how reviews correlate with the lyrics of songs and their artists. We also believe that finding more predictive models to compare to our random forest model could help enhance our results and ultimately find a way to use reviews from any dataset to predict an album’s rating.

There is a lot to be taken from music and song lyrics, but reviews ultimately shape the digital world we see today across streaming platforms, personalized data, and algorithm-based recommendations.

References

Almohalwas, Akram. All Lectures pertaining to “Statistics 133 Introduction to Text Mining Using R.” 4 Jan. 2024 - 11 Mar. 2024. University of California, Los Angeles.

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). “quanteda: An R package for the quantitative analysis of textual data.” *Journal of Open Source Software*, *3*(30), 774. doi:10.21105/joss.00774 (URL: https://doi.org/10.21105/joss.00774), <URL: https://quanteda.io>

Cabral, Luis, et al. “Music Reviews and Music Demand: Evidence from Pitchfork and Last.Fm.” *Lus X Cabral: INDEX_HTML*, Apr. 2021, http://luiscabral.net/economics/workingpapers/pitchfork%202021%20JIE.pdf.

“New Albums & Music Reviews.” *Pitchfork*, pitchfork.com/reviews/albums/. Accessed 10 Mar. 2024.

Robinson, Julia Silge and David. “2 Sentiment Analysis with Tidy Data: Text Mining with R.” *A Tidy Approach*, www.tidytextmining.com/sentiment. Accessed 15 Mar. 2024.

Zhang, Zhiyong. “Text Mining for Social and Behavioral Research Using R.” *PsychStat*, 23 Dec. 2019, books.psychstat.org/textmining/.