Analyzing Music Lyrics' Vocabulary, Themes, and Emotionality via NLTK to Predict Song Popularity

Jacob Farner
Computer Science, COMP331
Occidental College
Los Angeles, California
jfarner@oxy.edu

Leopold Ringmayr
Computer Science, COMP331
Occidental College
Los Angeles, California
lringmayr@oxy.edu

***Index Terms*—Emotionality and Sentiments, Logistic Regression, Lyric, Music Lyrics, Naive Bayes, NLTK Sentiment, Analysis**

I. Abstract

We explore the impact that lyrical content has on the commercial success of songs, and analyze trends in the subjects of songwriters lyrics in popular music over the span of 50 years, from 1965 to 2015. Different feature selection techniques including standard tokenization and TF-IDF are used in conjunction with multiple classification models including logictic regression, Naive Bayes, Decision Tree, and support vector machine (SVM) with varying degrees of success. In addition, this work sheds light on emerging themes in the lyrics over the last 6 decades of popular music as we examine thematic trends.

II. Introduction

In this paper, we examine the relationship between a song's lyrics and its placement on the Billboard Top 100 charts in order to determine whether or not lyrical content is a significant factor in a song's overall success. Beyond looking at whether this is a possible correlation, we delve into emerging themes and sentiments from the Billboard 100 songs over the span of fifty years.
While there has been extensive research regarding what makes a work of media stand out as a commercial blockbuster, especially in music and film, it seems that there is minimal work which analyzes the effect that lyrical content may have on a song's success. Rather, most research examines tonal qualities and musical features when looking at mass appeal, and lyrics are generally only used in providing a calculated sentiment for songs.

Applying natural language processing and sentiment analysis on music is compelling when approached from a variety of perspectives. First, there is significant economic motivation to this topic as this model aims to find patterns in lyrics that lead to hit songs. Although a song's success is defined by far more than lyrical content, artists and record labels could rely on this data and apply it to their writing in order to give their pieces better chances at appealing to mass audiences and subsequently rising on the charts. More song purchases and higher streaming numbers result in higher immediate revenue for song-makers, as well as improved name recognition and greater opportunities down the line. So, if an artist's goal with a particular piece of music was solely to make it onto the charts, they might increase the probability of this happening through "hit-worthy" vocabulary established through this model. It's difficult to measure direct financial impact a specific song has on an artist's revenue as most music is streamed today and artists gain the majority of their earnings from touring. However, artists with hits can charge more for ticket prices and perform more frequently and are generally granted more licensing opportunities. According to the International Federation of the Phonographic Industry (IFPI), the recording industry was valued at $19.1 billion in 2018, so top artists have massive earning potential.[3]

From a more human perspective, this project has potential to give a better understanding of how individuals listen to, interpret, and create music. There are numerous research works on how music impacts creativity, reasoning, and overall happiness, and writing music to target this specific type stimulation could be wildly beneficial for listeners and artists alike. This work specifically targets song lyrics and examines whether there is a possible relationship between the song lyrics and how popular the song is, indicating that certain lyrical patterns may resonate with listeners more consistently than others. In contrast to other existing work, this project deals specifically with the language processing of the song lyrics, and not with acoustic song elements (e.g. which instruments are used, tempo, key, time signature, etc.). Hence, this project explores whether one can find a hit-worthy vocabulary, i.e. whether higher chart positions can be predicted based on song lyrics (and the sentiments stemming from these lyrics) alone. Naturally, we are aware that other factors such as artist names or even record labels play a role as well.

III. Related Work

While there does not appear to be any work examining the relationship between lyrical content and a song's commercial success, there are many studies that have been conducted on lyrical sentiment and how music can affect individual's

emotions. We hoped that in this project, we might uncover a correlation, or at least a somewhat unique, consistent vocabulary that chart topping songs tend to use. One model that we used in our work is a Naive Bayes classifier, which are seen commonly throughout natural language processing projects.

One paper which suggests that there may be a correlation between lyrical content and a song's commercial success is *Automatic Prediction of Hit Songs*, in which the authors rely on support vector machines that take lyrical and acoustic information into account to predict whether or not a piece of music is likely to have commercial success.[4] Here, the authors conclude that there are commonalities between hit songs related to acoustic and lyrical data, although they admit that further research is necessary to solidify their claims. They had the highest rates of success with lyrical data for distinguishing hits and lower success with acoustic data. They also found that combining features of lyrical and acoustic data did little to improve the model. This paper also only considered songs that made it to the number one position on the charts between January 1956 and April 2004. These number 1 hits are then set apart from a general lyrics database containing approximately 47000 songs from 500 artists, which is a much larger and more diverse data set from our own. However, because so much goes into a song's general success between the size of the artist, label influence, and marketing, the lyrics on their own likely had little to do with the song's overall success.

In our eyes, music from artists that chart consistently is in an entirely different commercial class from the majority of music. Therefore, we choose to look only at songs that have reached the top 100 as they likely have similar sized names and marketing budgets across the board, comparatively. From here, we can look at songs in the top 100 that were pushed to a higher placement.

In *Naive Bayes Classifiers for Music Emotion Classification Based on Lyrics*, the authors demonstrate the effectiveness of a Naive Bayes classifier on analyzing lyrical sentiment, which indicated that a Naive Bayes approach would be an appropriate algorithm in the early stages of our model, and would a reasonable baseline at the very least. Naive Bayes also makes sense as a model logically, as it relies on the formula:

$$\mathrm{P(A|B_1.B_n)} = \mathrm{P(A)} \prod_{\mathrm{i}} \frac{\mathrm{P(A|B_i)}}{\mathrm{P(A)}}$$

This means that it assesses probability of a classification (A) given all values of a feature (B). This paper also presents an effective way to measure emotion through the Thayer emotion model, which we opted to use as well thanks to the Thayer emotion plane relying on energy and stress features which can be mapped to instrumental musical features, should we expand upon this work later on. The authors of *Naive Bayes Classifiers for Music Emotion Classification Based on Lyrics* also opted to simplify the emotion plane to two categories; positive and negative, which we ultimately did as well in order to make our model more cohesive given our dataset.

In this project, we also look at cultural references in order to determine whether or not there are patterns in subject matter consistent across high-charting songs. In the book, *Ancient Manuscripts in Digital Culture: Visualisation, Data Mining, communication*, the authors include a chapter titled Using Natural Language Processing to Search For Textual References, which describes how lexicons can be used in natural language processing in order to find categorical references in a piece of text.[2] Some artificial intelligent systems that process text use similar ideas to provide context clues to a system so that it can better understand what an individual is talking about. While we focused on modern lyrical content, this paper looks primarily at ancient religious and philosophical texts, relying largely on a bag-of-words approach combined with an n-gram model. While this paper looks at trends across the writings of great philosophers, we can apply similar methodology to identify trends in subject matter across high charting singles.

IV. Methodology

The main components for this classification problem are pre-processing of the raw CSV data, reading in and segmentation of the text data, feature engineering via NLTK and term frequency-inverse document frequency (TF-IDF) and classification through logistic regression, Naive Bayes, Decision Tree, and support vector machine(SVM) classification. Beyond the classification section, there is an emphasis on utilizing NLTK and common NLP techniques for exploratory purposes: Using TF-IDF and further word count tools, the most prominent phrases and lyrics are examined.

A. Project Data and Data Preparation

As we focused on the correlation between song lyrics and chart position, we needed a dataset which provides both. After considering various options, we settled on a dataset titled *50 Years of Pop Music Lyrics* hosted on Kaggle. This dataset contains 5100 entries which were scraped from Wikipedia entries on the Billboard top 100 from 1965 to 2015, giving us a wide range to work with. Dataset features include the songs' rankings, the artists performing each piece, the year each song was released, and the songs' lyrics. Because we looked primarily at the relation between lyrics and ranking, those two columns were most useful. As we are focusing on lyrical content, we chose to omit the artist name in our final model. On Kaggle.com, the dataset ReadMe notes that there are issues in the dataset surrounding collaborations between artists, but eliminating artist name form our model remedies this.

For this work, the text files are read in locally; the data is stored within a PyCharm project. The project data consists of 5,100 songs and their lyrics spanning 50 years with the following features: rank, song, artist, release year, and source, as well as a numerical identification column. A sample of the raw data can be seen below:

Fig. 1. Data Sample [1]

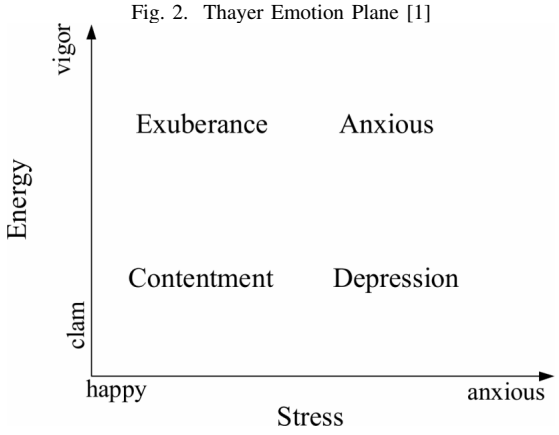
	# Rank	Song	Artist	# Year	Lyrics	Source
	1 100	4583 unique values	2473 unique values	1965 2015	NA 4% [empty] 1% Other (4632) 95%	1 79% 3 16% Other (2) 5%
					hanging about with anyone its not unusual to see me cry oh i wanna dieits not unusual to go out at any...	
65	65	shes about a mover	sir douglas quintet	1965	well she was walking down the street looking fine as she could be hey hey well she was walking down that street looking fine as she can be well you know i love you baby so yeah what id say shes about ...	5
66	66	shake	sam cooke	1965	shake shake shake shakelisterlisten while i talk to you i tell you what were gonna do theres a new thing thats going around and ill tell you what theyre puttin downjust move your body all around and ...	1
67	67	wonderful world	hermans hermits	1965	i dont know much about my history now	1

We relied primarily on the unique numerical IDs, lyrics, rank, and release year for our model. The numerical ID was used instead of song name, as there were many duplicate song names, resulting in only 4,583 unique values for this column. We also omitted the source column, as it was unclear what it represented. There was no mention of it on the Kaggle page, nor the git repository that contained the dataset's readme. We assumed that it referenced the websites that lyrics were sourced from and was irrelevant to our project, especially as it is not described anywhere. The data was originally presented in CSV format but was converted to a DataFrame using the Pandas Data Analysis Library for Python to allow for easier use. This helped with accessing and manipulating data rows and columns.

For building the classifier using Python's NLTK, the text messages are mapped to the corresponding label within the DataFrame.

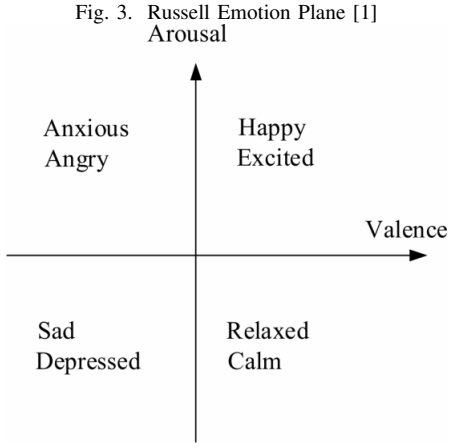
For our musical and lyrical analysis we also needed to determine an emotion model. As stated previously, we chose to use the Thayer emotion model, as its energy and stress features fit musical analysis well.

Fig. 2. Thayer Emotion Plane [1]



In general, the Thayer emotion plane is preferred when conducting sentiment analysis in music and is used often in natural language processing. Alternatively, the Russell emotion model opts for valence on the X axis and Arousal on the Y axis, and is also used in natural language processing models.

Fig. 3. Russell Emotion Plane [1]



While we considered the Russell Emotion model, we ended up using the Thayer emotion model for this work.

B. Data Cleaning and Feature Selection

Pre-processing of the data took place at different stages of the project and was a fundamental step for the overall feature extraction from the song lyrics. Missing data was an issue with this dataset; therefore, NaN values and empty values were handled using pandas' tools prior to any processing using NLTK. The individual songs were processed and cleaned using Python's NLTK, including removing stopwords, using word lemmas, converting to lowercase letters, and removing non-alphanumeric characters.

For feature selection, different features were extracted for the different models: To extract emotional and sentiment features of the song data, two tools were employed.

The first tool is the VADER Sentiment Analyzer, a library and lexicon for emotional analysis. For VADER, a positive, negative, and compound score are given based on the sentiments and emotions reflected in the text. The VADER Sentiment Analyzer generally has a higher performance on shorter texts. For this reason, to compute the VADER sentiment score, each song lyric text was divided into ten parts and the compound VADER score was calculated for each one. Then the final VADER sentiment score for the song was computed as an average of the ten parts; then this average value was converted into a categorical value and was stored in a new DataFrame column. By using this average score, the sentiment of the entire song would be determined more accurately.

Secondly, as an additional measure, the Text Blob scores for polarity and subjectivity were computed for each song as well. These values were also stored as new DataFrame columns, and used as possible features. TF-IDF values and vectors were computed for the songs and used as features for the Logistic Regression, Decision Tree, and SVM models.

C. Classification

For the classification portion, the sklearn library was utilized to build Logistic Regression, Decision Tree, and SVM models. Different approaches included using TF-IDF value as well as the sentiment scores as features. (Note: A POS-tagged approach was also tried; however, it was dismissed due to lower performance). For the Naive Bayes model, the NLTK Naive Bayes functionality was used. 10-fold cross-validation was employed to access the model performance and performance statistics accuracy, precision, recall were collected and averaged for this model.

D. Exploration of Themes/Lyrics and Visualization

For further analyses, new DataFrames were created for each decade of the original dataset. Using these smaller datasets for each decade, TF-IDF values were computed. Along with this, a custom lexicon was used to look at themes across the decades–to help identify emerging trends.

V. Results and Analysis

The first approach was a Naive Bayes model that took word frequency into account based on a vast lyrics lexicon, as well as VADER sentiment scores as features. This model, relying on 10-fold cross-validation, achieved an average accuracy of 52 percent over the ten folds, with folds above 54 percent. The average precision for the top charting class was 0.54 and the average recall for the top charting class was 0.41. For the lower charting class, the precision was 0.51 and the recall was 0.63 with an F-score of 0.56.

In addition to this Bayesian model, we chose to explore other statistical models through the sklearn python package, in order to see if there were any significant improvements based on our dataset. The first model we looked at was logistic regression, as it has appeared often in the works we've read on natural language processing. This model accepts data that has been cleaned and tokenized with features that have been selected through TF-IDF. For implementation we used the built in sklearn function, which returned the following:

Fig. 4. Logistic Regression Classifier

	precision	recall	f1-score	support
top 50	0.52	0.78	0.63	525
lower 50	0.51	0.25	0.34	495
accuracy			0.52	1020
macro avg	0.52	0.51	0.48	1020
weighted avg	0.52	0.52	0.49	1020

Logistic Regression accuracy Score: 0.5215686274509804

This model returned an accuracy of around 52%, with suboptimal classification metrics overall.

Another model that appeared in our related works specifically related to the lyric and chart correlation was support vector machine (SVM) classification. Again, this model accepted data that was cleaned and tokenized with features selected through TF-IDF. We used the built in sklearn function for SVM, which returned the following.

Fig. 5. SVM Classifier

	precision	recall	f1-score	support
top 50	0.53	0.74	0.62	525
lower 50	0.53	0.31	0.39	495
accuracy			0.53	1020
macro avg	0.53	0.52	0.50	1020
weighted avg	0.53	0.53	0.51	1020

SVM accuracy Score: 0.5294117647058824

This model performed marginally better than our logistic regression with an accuracy closer to 53% with similar classification metrics.

The last model we examined was a decision tree classifier, which we have personally had success with in similar projects. Again, this model used data that was cleaned and tokenized with features selected through TF-IDF. We used the built in sklearn function for our decision tree classifier, which yielded the following.

Fig. 6. Decision Tree Classifier

	precision	recall	f1-score	support
top 50	0.52	0.23	0.32	525
lower 50	0.49	0.77	0.60	495
accuracy			0.50	1020
macro avg	0.50	0.50	0.46	1020
weighted avg	0.50	0.50	0.46	1020

Decision Tree accuracy Score: 0.4950980392156863

This model performed marginally worse than all of our other models, at 49% with similar classification metrics to our other models. Besides these classification performance results, we report other results that shed light on themes and lyrics that are highly represented in the top Billboard charts. For example, an analysis of the most frequent terms across the different decades reveals very similar popular words and themes. These findings are visualized below in Figure 7 and Figure 8. In the word clouds, one sees that the most frequent terms in the 1960's include "love", "know", "time". These terms are also the most frequent in other decades such as the 1980s-1990s (Figure 8).

Beyond these word frequency findings, there also are the results from the TF-IDF analyses. TF-IDF scores for different words were calculated for the different decades. This analysis reveals similar trends as the intial word cloud. For the 1960's, the words "love" (0.052901), "baby" (0.036217),"got" (0.026179), "know" (0.025757) had the highest TF-IDF values. For the 1980's and 1990's, the most impactful features

Fig. 7. Word Cloud for 1964-1975 Song Lyrics

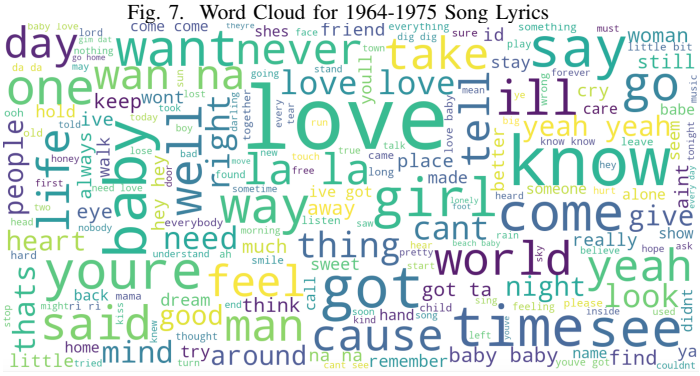
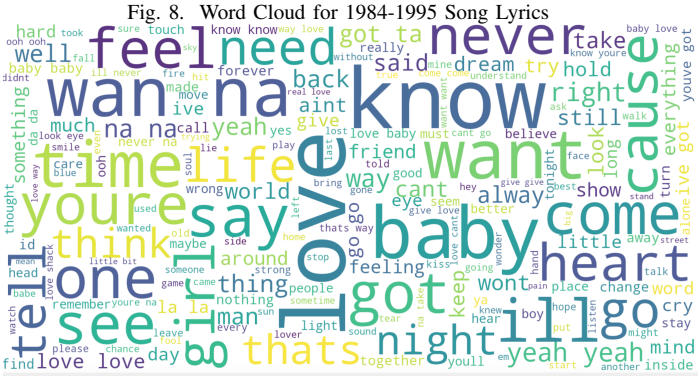


Fig. 8. Word Cloud for 1984-1995 Song Lyrics



were ”na” (0.043474), ”baby” (0.033555), ”want” (0.027853), ”got” (0.026458).

For the time span of 2005-2015, this trend is continued, with the highest TF-IDF valued terms being ”baby” (0.026551), ”girl” (0.025169), ”youre” (0.024727).

VI. Threats to Validity

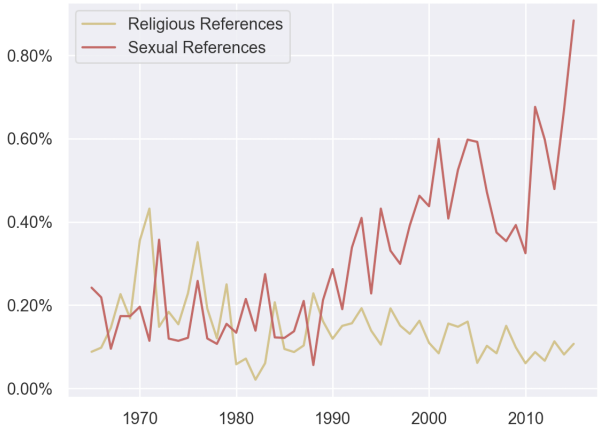
Song’s that chart highly generally have large production and marketing budgets behind them, as well as renowned artists. Our models do not include this information as variables, but they likely have significant impact based on market research we have conducted. We hoped to diminish the relative influence that these factors would have in our model by only using music that reached at top 100 or better on the charts. By doing this we can assume that all songs in the dataset likely have the same powers behind them, giving us a better idea of which lyrical patterns can push a song further. However, this may be criticized as there is likely significant homogenization across the majority of music that makes it to the top 100. A more robust model would include budget data for each song as well as artist metrics so that these songs could be tested against a more general sample of music.

In our exploratory works, we also found significant changes in lyrical content over the 50 years that our dataset spans. These general changes may be affecting our models, as we treat the entire dataset as a single collection rather than breaking up by decade. Our logic behind this decision, however, was that we wanted to see if there are intrinsic subjects that consistently resonate with people regardless of time period.

VII. Further Exploration

Unfortunately, our predictive models demonstrated little correlation between song lyrics and chart placement. However, we still believed that there were likely trends in popular music that would be apparent when looked at over the 50 year period. One exploratory aspect that we were interested in was generally subject matter in song lyrics, and examining how they change over time. Because this dataset spans 1965-2015, there were major cultural shifts which likely influenced what songwriters sang about. Primarily, we were interested in the mainstream acceptance of talking about what some would consider taboo subjects like sex, and any notable religious shifts over time. We felt that these two categories were opposites in many ways, and could effectively convey societal and cultural shifts in music than any other referenced subjects. For this, we built lexicons for each subject based off of related words and subjects on Thesaurus.com and Dictionary.com for the words ’God’ and ’Sex’. We also added in related slang for the latter. From here, song lyrics were treated as bags-of-words, and our program looked for number of occurrences of each word over time. We felt that these lexicons of keywords were an effective way to pick out references, as well as their frequency.

Fig. 9. Subject References Over Time



Here, a clear shift is visible in both topics. Over time, we see that on average, sexual references increase over time, while religious references tend to decrease. This could be explained by major cultural shifts over time, as society as a whole becomes more progressive, and topics like sex become more publicly appropriate to talk about. This could also be due to counter culture movements in music, like the psychedelic revolution of the 1960s and or rap in the 1990’s, for instance, push boundaries lyrically in order to make statements, political or otherwise. It should also be noted that this trend may be more significant than our model displays, as we had to omit various popular slang terms or euphemisms, as they were to close to normal vocabulary. For instance, the phrase ’doing it’ may be a sexual reference or an innocent lyric depending

on context. This graph also displays a significant decrease in religious references in popular music overtime, suggesting that religion may have actually become a more contentious subject overtime, perhaps due to decreased homogenization in religion in countries like the United States.

VIII. Conclusion

Based on the results from the models, there are different possible conclusions one is compelled to draw. We used different classification models and looked at different features to use for each model. The features we considered were: TF-IDF scores, word occurrence from the lexicon of all song lyrics, VADER sentiment scores, subjectivity and polarity scores from the Text Blob library, as well as our self-built lexicon. Based on our findings, paired with the fact that we explored a multitude of possibly significant features in different combinations, one conclusion would be that there may be no direct correlation between certain lyrics. Perhaps lyrics and sentiment scores derived from the lyrics are not sufficient indicators of what constitutes a hit. Related research often includes acoustic elements of songs in their analysis along with the lyrics.

However, this work did shed light on some possible themes and lyrics that tend to occur more frequently with hit songs. For example, we found similarities and differences across the different decades—this suggests that some themes are timeless and universally suited for "hit" songs. The model findings lend credibility to the idea that a hit song may not be reliant on certain lyrics.

Therefore, for future work, it would be interesting to consider acoustic elements along with the mere song lyrics to provide a more holistic picture of what constitutes a hit song. After all, there are many factors that play a role in the overall popularity of a song. One aspect–which requires a more comprehensive dataset–would be to use artist names as a model feature and assign a weighted score to each artist based on their history of creating hit songs. Additionally, other musical aspects such as beat, tempo, instruments, and genre play a role in the charting as well. Therefore, one might not be able to create a hit based on lyrical analysis alone.

Acknowledgment

Thank you to Professor Chen for a great semester of NLP, where we learned many aspects of natural language processing through toolkits like NLTK, as well as in-depth explanations of the models we use, so that we are capable of applying them to our other endeavors.

References

[1] 50 years of pop music lyrics: https://www.kaggle.com/rakannimer/billboard-lyrics

[2] Graham, Brett. "Using Natural Language Processing to Search for Textual References." In Ancient Manuscripts in Digital Culture: Visualisation, Data Mining, Communication, edited by Hamidović David, Clivaz Claire, and Savant Sarah Bowen, by Marguerat Alessandra, 115-32. LEIDEN; BOSTON: Brill, 2019. Accessed May 8, 2020. doi:10.1163/j.ctvrxk44t.11.

[3] "Global Recorded Music Sales Totalled US $19.1nbsp;Billion in 2018." IFPI, IFPI, www.ifpi.org/global-statistics.php.

[4] Dhanaraj, Ruth and Beth Logan. "Automatic Prediction of Hit Songs." ISMIR (2005).

[5] Cormack, G. "Content-based web spam detection." In Proceedings of the 3rd international workshop on adversarial information retrieval on the web (AIRWeb). New York, 2007.

[6] X. Hu, J. S. Downie, and A. F. Ehmann, "LYRIC TEXT MINING IN MUSIC MOOD CLASSIFICATION," Poster Session, p. 6, 2009.

[7] Y. An, S. Sun, and S. Wang, "Naive Bayes classifiers for music emotion classification based on lyrics," in 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), May 2017, pp. 635–638, doi: 10.1109/ICIS.2017.7960070.