INSTITUTE OF COMPUTER SCIENCE
College of Arts and Sciences
University of the Philippines Los Baños

SENTIMENT ANALYSIS ON SONGS BASED ON SONG LYRICS USING NAÏVE BAYES ALGORITHM

AYESSA AMOR NOSOTROS HERNANDEZ

A SPECIAL PROBLEM SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF SCIENCE IN COMPUTER SCIENCE

August 2023

The special problem attached hereto, entitled

"SENTIMENT ANALYSIS ON SONGS BASED ON SONG LYRICS USING NAÏVE BAYES ALGORITHM"

prepared and submitted by **AYESSA AMOR N. HERNANDEZ** in partial fulfillment of the requirements for the degree of BACHELOR OF SCIENCE IN COMPUTER SCIENCE is hereby accepted.

PROF. CONCEPCION L. KHAN
Special Problem Adviser

(Date Signed)

PROF. REGINALD NEIL C. RECARIO
Director - OIC
Institute of Computer Science

(Date Signed)

Permission is given for the following people to have access to this special problem:

Available to the general public	Yes
Available only after consultation with author/SP adviser	No
Available only to those bound by a confidentiality agreement	No

TABLE OF CONTENTS

ABSTRACT

Sentiment Analysis on Songs based on Song Lyrics using Naïve Bayes Algorithm

AYESSA AMOR N. HERNANDEZ
University of the Philippines Los Baños
August 2023

Adviser:
Concepcion L. Khan

Music induces basic to complex emotions such as happiness, sadness, and nostalgia. These emotions can be classified into categories like positive or negative using sentiment analysis. Existing studies on mood classification mostly focus on the audio features of a song while the lyric features are ignored. A few studies on lyrics mood classification, on the other hand, pointed out the need to explore other classifiers like Naïve Bayes and improve its performance, using a larger dataset. In this study, a Naïve Bayes classifier model was created to identify whether a song is positive or negative based on its lyrics. The model which produced exceptional results with 95.02% accuracy and 94.42% precision was trained and tested using a dataset containing 1,810 song lyrics. Feature extraction techniques such as N-grams (tri-grams) and TF-IDF were applied after preprocessing the data.

Chapter I

INTRODUCTION

A. Background of the Study

Music makes people feel different emotions which affect their mood. It may induce basic to complex emotions such as happiness, sadness, or nostalgia (Juslin et al., 2014). These emotions or moods can be classified using sentiment analysis.

A listener's music preference might depend on features like melody or lyrics. This study, however, focuses on the lyrical feature of a song. The three experiments conducted by Stratton and Zalanowski (1994) show that lyrics *"appear to have greater power to direct mood change than music alone and can imbue a particular melody with affective qualities"*.

The sentiment is *"an attitude, thought, or judgment prompted by feeling"* ("Definition of Sentiment," 2023). Sentiment analysis is a Natural Language Processing (NLP) technique to extract sentiments from texts like lyrics. Some types of sentiment analysis are fine-grained sentiment analysis which focuses on the polarity of the sentiments and the emotion detection type which focuses on the specific emotion of a particular sentiment. According to Kapoor (2021), sentiment analysis can be used in fields involving brand monitoring, integrated analysis, public relations, marketing, data mining, and political analysis.

B. Statement of the Problem

Existing studies on mood classification mostly focus on the audio features of a song while the lyric features are ignored. However, the lyric features of a song have proved its usefulness in mood classification. Other researchers who conducted studies in lyrics mood classification pointed out the need to explore other classifiers such as Naïve Bayes since this classifier is one of the top classifiers for text classification. When modeling the Naïve Bayes classifier, there is an observation of a need to use a larger dataset when training and testing while producing higher accuracy scores.

Lyrics are hard to classify because lyrics are mostly abstract, and emotions are conveyed indirectly. There is a need to discover a combination of feature extraction techniques that will help to produce higher accuracy scores.

There are no existing open-source sentiment-classified datasets with readily available lyrics that can be used for training and testing mood classifiers due to copyright issues thus requiring the need to collect song lyrics to use.

C. Significance of the Study

The mood classification process on song lyrics is still lacking in exploring other classifier algorithms such as Naïve Bayes. Sentiment analysis on song lyrics can be used not only in the music field but also in psychology. Furthermore, this study can contribute to eventually solving the existing challenges concerning this topic.

a. Music Information Retrieval (MIR)

Music Information Retrieval (MIR) is a multidisciplinary research field that aims to develop systems and processes to retrieve information from music. Input data for MIR can be images, symbolic formats, digital audio, and metadata while outputs can be tasks like information retrieval, classification or estimation, and sequence-labeling (Burgoyne et al., 2015).

Perception of music is studied in the field of MIR and is usually applied in music recommendation systems and music search engines (Fell et al., 2020, Chapter 1).

b. Music Emotion Recognition (MER)

Music Emotion Recognition (MER) is a computational task that aims to automatically recognize emotions from music.

With the advancement of brain science, music emotion recognition (MER) has received a lot of attention in academia and business. For example, recommendation systems, automatic music composition, psychotherapy, and music visualization have all made extensive use of MER (Cui et al., 2022).

c. Music Psychology

Music psychology is a field of study that aims to understand and explain human behavior and experience toward music. Research in music psychology has applications in a variety of fields, including music composition, performance, education, criticism, therapy, and studies of human attitude, performance, intelligence, creativity, and social behavior.

D. Objectives

The main objective of this study is to evaluate the performance of the Naive Bayes classifier algorithm on sentiment analysis based on song lyrics, thus the specific objectives are as follows:

1) Collect song lyrics from the internet
2) Apply data cleaning and data preprocessing (tokenization, feature extraction techniques, etc.) to the dataset
3) Train and test the dataset
4) Evaluate the performance of the classifier

E. Scope and Limitations

This study focuses on the evaluation of the Naïve Bayes classifier algorithm and thus other classifier algorithms (e.g. SVM) are not used. Also, the dataset used in this study is limited to English song lyrics only.

Chapter II

REVIEW OF RELATED LITERATURE

A. Sentiment Analysis using Naïve Bayes Algorithm

a. Mobile phone reviews and Tweets

Sentiment analysis is commonly used in analyzing and understanding product or service reviews. Sentiment analysis on mobile phone reviews by Shaheen (2019) was done using different supervised learning techniques. These are the SGD Classifier, Gradient Boosting Classifier, Multinomial NB, NB-SVM, LSTM, CNN, and Random Forest. The Random Forest classifier, with 85% accuracy, outperformed other classifiers while the Multinomial NB and the NB-SVM classifiers showed good results with 70.55% and 73.51% accuracy respectively. Samuel et al. (2020) did a sentiment analysis of the public's COVID-19 sentiments using Tweets. The study used the Naive Bayes and the logistic regression classification methods, and a strong accuracy of 91% was observed with the Naive Bayes method.

b. Song lyrics

A study by Buzic and Dobša (2018) showed the success of Naïve Bayes in lyrics classification when they classified if a song was by Metallica or Nirvana. The model got a precision score of 93% with a Laplace estimator of 0.06.

In a study by Raschka (2016), they manually labeled their 1200 dataset (1000 for training and 200 for testing) into happy and sad categories. Their Naïve Bayes model got an 80% accuracy in training and 54.5% accuracy in testing. In another study by Tan et al. (2019) on music mood recognition using lyrics, they used the Naïve Bayes algorithm to train 120 songs and test 60 songs and their model got an 85% accuracy score. These studies showed the need to (1) produce higher accuracy scores in testing while (2) using a larger dataset.

B. Mood Classification using Lyrics

Despite the semantic richness of lyrics, Music Information Retrieval (MIR) has focused only on using the audio track to retrieve music information (Fell et al., 2020, Chapter 1). However, a study by Hu and Downie (2010) proved the significant importance of lyric features in mood classification. The study shows that lyric features like CW n-gram, GI-lex, and Affe-lex worked significantly better than audio features while text stylistic features performed the worst among all feature types. In another study by Shukla et al. (2017) where the performances of combined lyric feature types were evaluated, CW n-gram+GI-lex+ANEW+Affe-lex+FW+TextStyle performed the best with an accuracy of 63.8%.

Shukla et al. (2017) proved in their study the usefulness of lyrics in mood classification by utilizing linguistic lyric features and text stylistic features of song lyrics. They pointed out the need to explore other classification models in mood classifications, other than using SVM or Support Vector Machine.

Chapter III

METHODOLOGY

A. Flowchart

The figure below shows the flowchart of the steps done in this study:

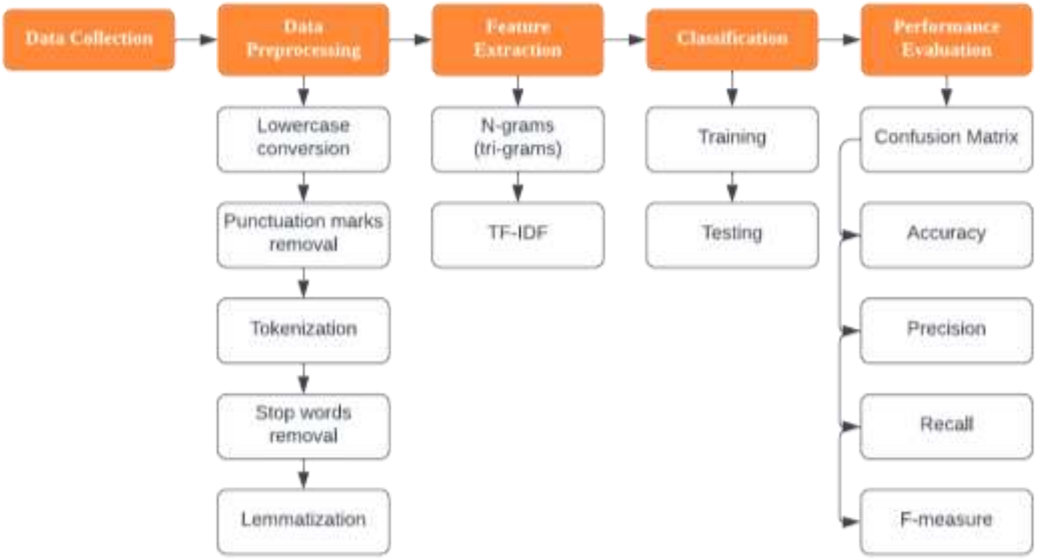


Figure 1. *Flowchart*

B. Data Collection

The collection of lyrics was done using a *python* library called *lyricsgenius*. It retrieves lyrics from Genius.com using the Genius API. The artist and song title are needed to search for a song using this library. It is easy to use and easy to understand, and many projects from online sources proved the work of this library.

A sentiment-classified dataset from MoodyLyrics which consists of 2000 songs with information such as the artist and the song title is used to search for a song. It is a balanced dataset with 1000 songs for each mood — positive and negative.

However, while collecting the data, there are song lyrics not found and song lyrics that are not in English resulting in an unbalanced dataset with 1810 song lyrics.

C. Data Cleaning

Data cleaning was done using regular expressions or *regex* to eliminate unnecessary punctuation and words in the data. The lyrics were also transformed to lowercase.

D. Data Preprocessing

Data preprocessing is a crucial step in data analysis. It is being done to give more meaning to the data by removing meaningless words or reducing different forms of words to their single form.

In this study, data preprocessing was done using the Natural Language Toolkit (NLTK). It is a collection of libraries for NLP which supports tokenization, stemming, lemmatization, and classification.

a. Tokenization

Tokenization is the process of breaking down a text into subpieces or words. It is being done for the classification model to easily understand and process the data.

In this study, tokenization was done using the `word_tokenize()` function.

b. Stop words removal

Stop words are words that are commonly used but are unimportant in the analysis, such as “the”, “is”, and “and”. Stop words removal is done to eliminate these meaningless words which results in a clearer analysis.

In this study, an NLTK dictionary of stop words is used to remove stop words. Words like “yeah”, “ohh”, “ahh”, and “oooh” were also removed.

c. Lemmatization

In this study, lemmatization was done instead of stemming. Stemming is the process of chopping off the ends of a word to get the word’s stem without looking at its meaning. On the other hand, lemmatization uses a vocabulary of words and removes inflectional endings only, and then returns the base of the word which is more suitable for this study. The *WordNet Lemmatizer* of NLTK was used for lemmatization in this study.

Table 1. *Sample Data Preprocessing*

Before Cleaning and Preprocessing	I walked through the door with you, the air was cold But something 'bout it felt like home somehow And I left my scarf there at your sister's house And you've still got it in your drawer, even now
After Cleaning	i walked through the door with you the air was cold but something bout it felt like home somehow and i left my scarf there at your sister s house and you ve still got it in your drawer even now
After Preprocessing	walked door air cold something bout felt like home somehow left scarf sister house still got drawer

E. Data Visualization

Data visualization is usually done when analyzing large data to easily understand what the data is saying. In this study, data visualization was done using *Seaborn* library for bar graphs, the *wordcloud* library for wordclouds, and *matplotlib* for plotting these graphs.

F. Feature Extraction

Feature extraction or text vectorization is the process of analyzing similarities between pieces of text by converting text data into a vector of features. The most common feature extraction techniques are the Bag-of-Words (BoW) and the Term Frequency-Inverse Document Frequency (TF-IDF). The BoW technique, as the name suggests, is the process of making a bag of words without considering the grammar and the order of the words and then getting the frequency of each word. On the other hand, the basic idea of TF-IDF is that a word that frequently occurs in a document but rarely in the entire corpus is more informative than a word that frequently occurs in both the document and the corpus (GeeksforGeeks, 2023).

In this study, however, the ordering of the words in the lyrics should be considered thus the use of *n-grams* is used instead. *N-grams* is the same as BoW where the frequency of words is taken note of. The difference is that *n-grams* considers the ordering of the words in a text.

There are three common *n-grams* used in NLP — unigram (*n*=1), bigram (*n*=2), and tri-gram (*n*=3).

Table 2. *Sample implementation of N-grams*

<table>
<tr><td colspan="2">Text:
“This is a text.”</td></tr>
<tr><td>Unigram</td><td>“This”, “is”, “a”, “text”</td></tr>
<tr><td>Bigram</td><td>“This is”, “is a”, “a text”</td></tr>
<tr><td>Tri-gram</td><td>“This is a”, “is a text”</td></tr>
</table>

In this study, feature extraction was done using the *CountVectorizer* and *TfidfVectorizer* of *sklearn*.

G. Classification

A subfield of artificial intelligence (AI) and computer science called machine learning focuses on using data and algorithms to simulate how humans learn, gradually increasing the accuracy of the system. The two most common machine learning methods are supervised learning, which uses labeled datasets to train and predict the outcomes accurately, and unsupervised learning, which uses machine learning algorithms to cluster unlabeled datasets (What Is Machine Learning? | *IBM*, n.d.). This study uses supervised learning because it is more appropriate for mood classification.

a. *Naïve Bayes Classifier Algorithm*

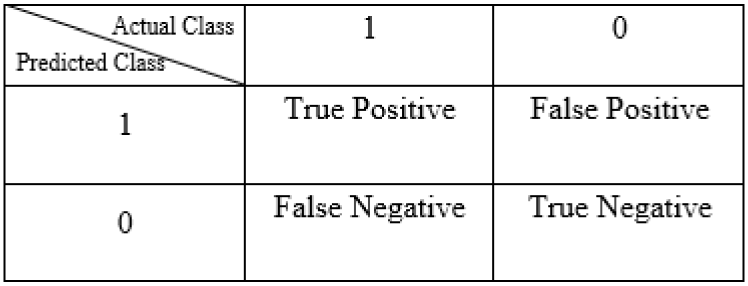
Naïve Bayes classifier is a supervised machine learning algorithm that assumes that all features are independent of each other. Common applications of text classification using Naïve Bayes are spam filtering, sentiment analysis, and recommendation systems.

Because this study deals with a small dataset, the Naïve Bayes algorithm is the most suitable classifier (Althnian et al., 2021). In this study, the *sklearn* library was used for training and testing. 66.66% of the dataset is used for training and 33.33% is used for testing.

H. Performance Evaluation

To evaluate the performance of the model, the following metrics were used:

a. *Confusion Matrix*



Actual Class Predicted Class	1	0
1	True Positive	False Positive
0	False Negative	True Negative

Table 3. *Confusion Matrix*

A confusion matrix is used to evaluate a performance of a classification model.

Four potential outcomes in a confusion matrix:

i. *True positive*: the correct prediction of the presence of a condition

ii. *True negative*: the correct prediction of the absence of a condition

iii. *False positive*: the incorrect prediction of the presence of a condition

iv. *False negative*: the incorrect prediction of the absence of a condition

b. *Evaluation Metrics*

i. Accuracy

Accuracy measures how often the classifier makes the correct prediction. It is the correct predictions divided by total predictions.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

ii. Precision

Precision measures how correct the prediction of the classifier is. It is the total number of correctly classified classes divided by the total number of predicted positive classes.

$$precision = \frac{TP}{TP + FP}$$

iii. Recall

Recall (also called Sensitivity) measures actual observations that are predicted correctly. It is the total number of correctly classified classes divided by the total number of actual positive classes.

$$recall = \frac{TP}{TP + FN}$$

iv. F1-score

The F1-score or F-measure is the harmonic mean of precision and recall.

$$f1 - score = \frac{2 * precision * recall}{precision + recall}$$

Chapter IV

RESULTS AND DISCUSSION

A. Exploratory Data Analysis

The following visualizations are presented to get an idea of what the data tells us:

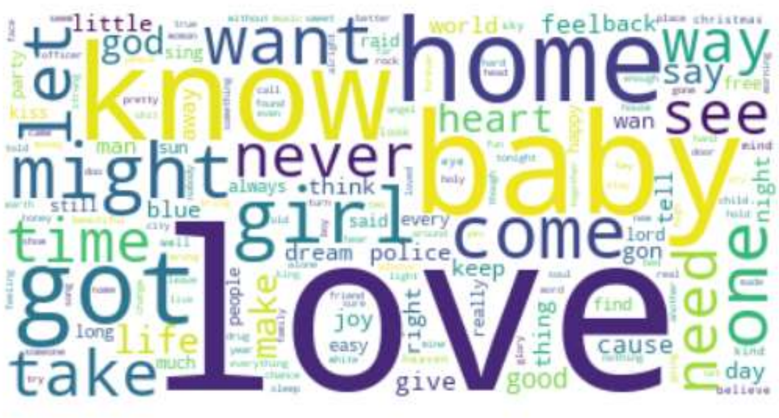


Figure 2. *A word cloud of the songs with a positive mood*

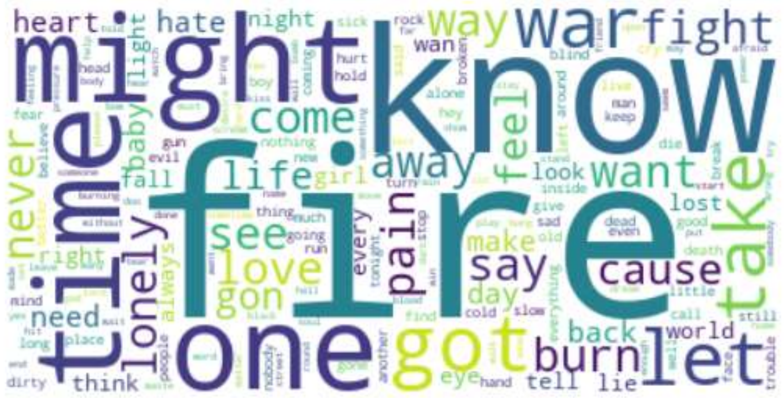


Figure 3. *A word cloud of the songs with a negative mood*

Figure 2 shows that words like *love* and *baby* are often used in positive songs while **Figure 3** shows that *know* and *fire* are often used in negative songs. With this, the difference between the lyrics of each mood is shown.

B. Results and Discussion

The following evaluation shows the performance of the Naïve Bayes classifier in identifying the mood of a song based on its lyrics.

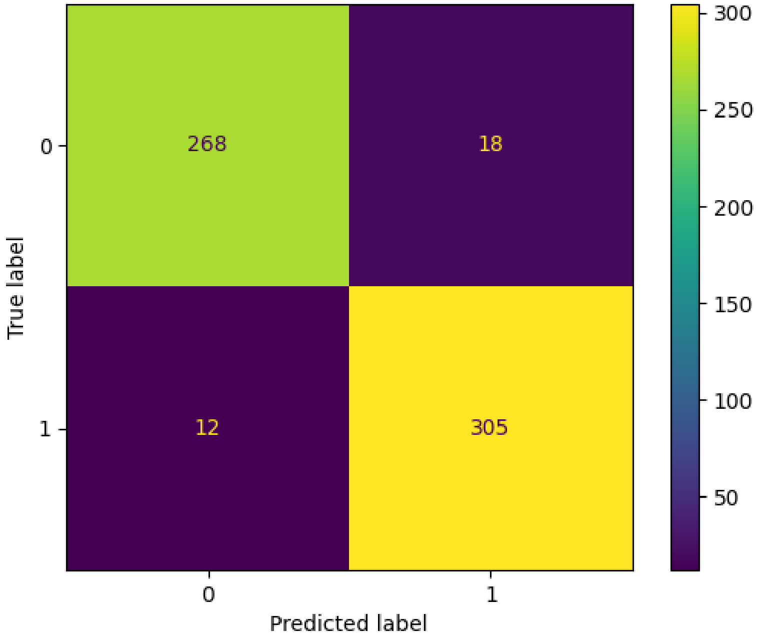


Figure 4. *Confusion Matrix of the NB classifier*

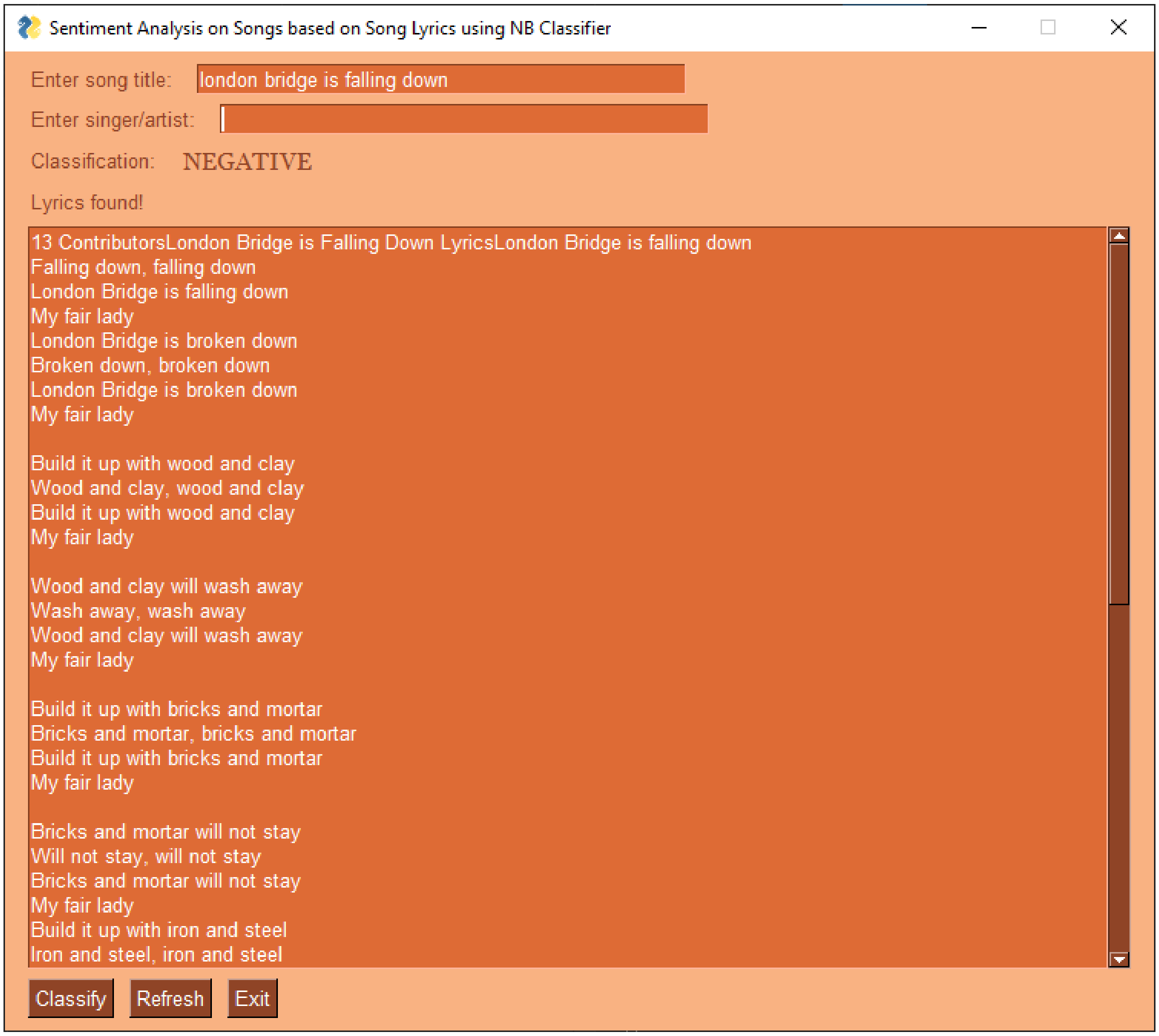
Out of the 603 song lyrics for testing, 323 songs for the positive class and 280 songs for the negative class, the classifier correctly classified the positive class with 94.42% precision while getting a 95.71% precision for the negative class. This implies that the classifier is effective for classifying both positive and negative classes which is what is needed for mood classification.

The table below shows the accuracy, precision, recall, and f1-score of the classifier:

Accuracy	95.02%
Precision	94.42%
Recall	96.21%
F1-score	95.31%

Table 4. *Evaluation scores of the classifier*

In **Table 4**, the F1-score shows that the classifier, despite being trained with an unbalanced dataset, obtained incredible results. **Table 4** shows that the overall performance of the classifier is exceptional but should always take into account that the results are dependent on the sentiment-annotated dataset, MoodyLyrics, which is used in this study.



Sentiment Analysis on Songs based on Song Lyrics using NB Classifier

Enter song title: london bridge is falling down

Enter singer/artist:

Classification: NEGATIVE

Lyrics found!

```
13 ContributorsLondon Bridge is Falling Down LyricsLondon Bridge is falling down
Falling down, falling down
London Bridge is falling down
My fair lady
London Bridge is broken down
Broken down, broken down
London Bridge is broken down
My fair lady

Build it up with wood and clay
Wood and clay, wood and clay
Build it up with wood and clay
My fair lady

Wood and clay will wash away
Wash away, wash away
Wood and clay will wash away
My fair lady

Build it up with bricks and mortar
Bricks and mortar, bricks and mortar
Build it up with bricks and mortar
My fair lady

Bricks and mortar will not stay
Will not stay, will not stay
Bricks and mortar will not stay
My fair lady
Build it up with iron and steel
Iron and steel, iron and steel
```

Figure 5. *Sample run UI*

Chapter V

CONCLUSION AND RECOMMENDATIONS

A Naïve Bayes classifier was created to evaluate its performance on mood classification of songs based on the song lyrics. The classifier showed high evaluation scores after being trained and tested on a dataset consisting of 1,810 song lyrics.

While collecting song lyrics from Genius.com using the *lyricsgenius* library, there are song search requests made by the library that appears to be successful but do not qualify as correct, for example, returning a book passage or a list of songs instead of the song lyrics. Because of these errors, data cleaning can be challenging for much larger datasets.

With proper data pre-processing and feature extraction techniques, it can be concluded from this study that the Naïve Bayes algorithm serves as a powerful tool in song lyrics mood classification. However, despite observing exceptional results in this study, this work can be improved by using a much larger dataset to see the stability of the effectiveness of the classifier.

References:

Burgoyne, J. A., Fujinaga, I., & Downie, J. S. (2015). Music Information Retrieval. In *John Wiley & Sons, Ltd eBooks* (pp. 213–228). Wiley. https://doi.org/10.1002/9781118680605.ch15

Buzic, D., & Dobša, J. (2018). *Lyrics classification using Naive Bayes*. https://doi.org/10.23919/mipro.2018.8400185

Çano, E., & Morisio, M. (2017). MoodyLyrics. *MoodyLyrics: A Sentiment Annotated Lyrics Dataset*. https://doi.org/10.1145/3059336.3059340

Cui, X., Wu, Y., Wu, J., You, Z., Xiahou, J., & Ouyang, M. (2022). A review: Music-emotion recognition and analysis based on EEG signals. *Frontiers in Neuroinformatics*, *16*. https://doi.org/10.3389/fninf.2022.997282

Fell, M., Cabrio, E., & Gandon, F. (2020). *Natural language processing for music information retrieval : Deep analysis of lyrics structure and content* [PhD Thesis]. Université Côte d'Azur.

Johnwmillr. (2020). *GitHub - johnwmillr/LyricsGenius: Download song lyrics and metadata from Genius.com 🎶🎤*. GitHub. https://github.com/johnwmillr/LyricsGenius

Juslin, P. N., Harmat, L., & Eerola, T. (2014). What makes music emotionally significant? Exploring the underlying mechanisms. *Psychology of Music*, *42*(4), 599–623. https://doi.org/10.1177/0305735613484548

Kapoor, N. (2021, December 30). Types of Sentiment Analysis and Its Uses - The Startup - Medium. *Medium*. https://medium.com/swlh/types-of-sentiment-analysis-and-its-uses-ad733535c895

Raschka, S. (2016, November 1). *MusicMood: Predicting the mood of music from song lyrics using machine learning*. arXiv.org. https://arxiv.org/abs/1611.00138

Samuel, J., Ali, G. G. M. N., Rahman, M., Esawi, E., & Samuel, Y. (2020). COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. *Information*, *11*(6), 314. https://doi.org/10.3390/info11060314

Shukla, S., Khanna, P., & Agrawal, K. C. (2017). *Review on sentiment analysis on music*. https://doi.org/10.1109/ictus.2017.8286111

Stratton, V. N., & Zalanowski, A. H. (1994). Affective Impact of Music Vs. Lyrics. *Empirical Studies of the Arts*, *12*(2), 173–184. https://doi.org/10.2190/35t0-u4dt-n09q-lqhw

Tan, K. R., Villarino, M. L., & Maderazo, C. (2019). Automatic music mood recognition using Russell's twodimensional valence-arousal space from audio and lyrical data as classified using SVM and Naïve Bayes. *IOP Conference Series, 482,* 012019. https://doi.org/10.1088/1757-899x/482/1/012019