Enhancing Music Emotion Classification Using Multi-Feature Approach

Affreen Ara[1], Rekha V[2]
Christ University, Bengaluru, India[1]
Department of Computer Science and Engineering, Christ University, Bengaluru, India[2]

***Abstract*—Emotions are a fundamental aspect of human expression, and music lyrics are a rich source of emotional content. Understanding the emotions conveyed in lyrics is crucial for a variety of applications, including music recommendation systems, emotion classification, and emotion-driven music composition. While extensive research has been conducted on emotion classification using audio or combined audio-lyrics data, relatively few studies focus exclusively on lyrics. This gap highlights the need for more focused research on lyric-based emotion classification to better understand its unique challenges and potentials. This paper introduces a novel approach for emotion classification in music lyrics, leveraging a combination of natural language processing (NLP) techniques and dimension reduction methods. Our methodology systematically extracts and represents the emotional features embedded within the lyrics, utilizing a diverse set of NLP techniques and integrating new features derived from various emotion lexicons and text analysis. Through extensive experimentation, we demonstrate the effectiveness of our approach, achieving significant improvements in accurately classifying the emotions expressed in music lyrics. This study underscores the potential of lyric-based emotion analysis and provides a robust framework for further research in this area.**

Keywords—Emotion classification; music lyrics; feature extraction; lexicon features

I. Introduction

Music can evoke strong emotions in listeners, such as happiness, sadness, anger, or excitement [1]. Analyzing the emotional content of music lyrics provides insights into why certain songs are popular and how they affect our moods. Emotion analysis employs natural language processing (NLP) and machine learning techniques to identify and extract emotional information from text. This analysis involves examining written or spoken language to understand how different songs and genres influence emotional responses. Emotion analysis of music lyrics is applicable in various fields, including marketing, social media analysis, mental health, and music therapy [2]. For instance, music therapy leverages specific songs and genres to help individuals' process emotions and improve well-being, which is also useful for diagnosing and monitoring mental states in mental health settings.

NLP, a branch of artificial intelligence, plays a pivotal role in emotion analysis by enabling computers to understand and interpret human language [3]. In emotion analysis, NLP techniques such as emotion detection are utilized to identify and classify emotions expressed in lyrics. The process typically begins with pre-processing the music lyric text, involving tokenization, normalization, and stop-word removal. This is followed by feature extraction, where relevant features such as stylistic, semantic, and lexicon-based attributes are identified and extracted from the text. Machine learning algorithms are then trained on these extracted features to identify the emotions embedded within music lyrics.

Emotion lexicons, which include vocabulary terms associated with one or more emotions, are crucial resources in NLP, sentiment analysis, and emotion detection [4]. These lexicons, such as the Affective Norm for English Words (ANEW) [5], enable emotion analysis algorithms to classify the emotional tone of text based on word associations. ANEW is grounded in the Russell Model's [6] two-dimensional circumplex model, which includes valence (pleasantness) and arousal (activation). While word-based lexicons are widely used, phrase-based and context-based lexicons offer a more nuanced understanding of emotional content. However, the subjectivity of emotions, polysemy, and contextual dependence can pose challenges in accurately assigning emotions to text. To overcome these limitations, emotion analysis often combines lexicons with other features, such as sentiment analysis, syntactic analysis, and semantic analysis. This integrated approach provides a more comprehensive and accurate understanding of the emotional content of text data.

Dimension reduction [7] is a technique used in data analysis and machine learning to reduce the number of features or variables in a dataset while preserving essential information. This technique is vital for uncovering underlying emotional patterns in lyrics, allowing for more accurate and computationally efficient machine learning models. By applying dimension reduction methods like Principal Component Analysis (PCA), researchers can effectively interpret and classify the rich emotional expressions found in music, providing valuable insights into the intersection of art and emotion.

Machine learning algorithms, such as decision trees, k-Nearest Neighbor (KNN), Naive Bayes, and Support Vector Machines (SVM), are commonly used for emotion classification in lyric text. The choice of algorithm depends on factors such as dataset size, feature representation, and computational efficiency. While most existing research in emotion analysis adopts a multimodal approach, combining both audio and lyrics to enhance emotion prediction accuracy, there is a notable gap in research focusing solely on lyrics as a textual representation of emotions. This research addresses this gap by providing a comprehensive analysis of emotions in music lyrics and designing new lexicon features that enhance emotion detection capabilities.

Our previous research [8] investigated the effectiveness of various lexicon-based features for emotion classification using music lyrics, achieving an accuracy of 53%. We extracted lexicon features from the NRC Emotion Intensity Lexicon, NRC VAD, EmoWordNet, and Synesketch lexicon. Extensive experiments were carried out using both lexicon and hybrid lexicon features for emotion classification in music lyrics; the current study extends this approach by incorporating a broader set of features, such as stylistic, lexical, and textual attributes, along with traditional NLP methods. Furthermore, we use Principal Component Analysis (PCA) for dimensionality reduction to enhance classifier performance by capturing the most informative aspects of the data.

This paper aims to enhance emotion classification accuracy in music lyrics by integrating various text features, including lexicon-based, syntactic, stylistic, and semantic attributes. Our contributions are threefold: (1) We create robust hybrid feature sets by combining lexicon-based features from multiple emotion sources to augment the emotional representation in lyrics; (2) We also use feature engineering techniques to extract syntactic, semantic, and stylistic information from the lyrics, thereby capturing emotional complexity of music lyric text; (3) We also use Principal Component Analysis (PCA) to reduce the dimensionality of the feature space in order to improve computational efficiency and model interpretability. By conducting extensive experimentation with classifiers such as Decision Trees, Random Forest, and Gradient Boosting, we demonstrate that our approach significantly enhances classification accuracy, achieving up to 98% accuracy. The remainder of this paper is structured as follows: Section II reviews the related work, Section III explains the method used, Section IV presents the experimental results, Section V discusses the findings and their implications, and lastly, Section VI concludes the paper with potential directions for future research.

II. Literature Review

This section presents literature review from papers of emotion analysis from music lyrics and text analysis.

Revathy V.R. [9] study uses knowledge from the MER dataset to train the Music4All dataset, aiming to label lyrics with emotions. The research employs a transfer learning approach and the Sentence Transformer model for emotion prediction. The LyEmoBERT model demonstrated superior performance compared to existing methods on the Music4All dataset. Sujeesha et al. [10] research develops a multimodal music mood classification system using transformers, incorporating both audio and lyrics. The study compares the system's performance with a Bi-GRU-based model, finding that the transformer-based model with transfer learning achieves higher accuracy. Priyanka et al. [11] propose a mood categorization of songs based solely on lyrics using TF-IDF feature extraction and the Random Forest algorithm. Their findings highlight the model's ability to accurately predict "happy" and "sad" emotions. Yudhik Agrawal [12] study proposes a deep neural network architecture using the XLNet Transformer model for emotion classification in music lyrics. The model employs multi-task learning through weight sharing, enhancing convergence speed and reducing over fitting. Fika Hastarita Rachman et al. [13] work introduces a method to recognize song emotions by combining lyrics and audio using the Thayer emotion model. They extract psycholinguistic and stylistic features from lyrics and audio waveforms, using various classifiers to perform emotion classification. Cong Jin et al. [14] propose a Bi-LSTM network with dilated recurrent skip connections to improve the model's ability to capture long-sequence information in lyrics. The model includes an attention mechanism to enhance the recognition of important words, improving semantic extraction performance. Study presents MoodNet[15], a deep convolution neural network (CNN)-based architecture designed to determine emotions from audio and lyrics. The model is evaluated using the MIREX Multimodal Dataset and the Million Song Dataset. Shahrzad Naseni et al. [16] research explores the connection between song lyrics and mood using two transformer-based approaches: natural language inference and next sentence prediction. The study focuses on lyric classification tasks to understand how lyrics and acoustics contribute to song mood. Leroto Parisi [17] study examines various feature vector representations like BERT, ELMo, and fastText embeddings combined with deep learning mechanisms to predict emotions conveyed in song lyrics. Yingjin Song and Daniel Beck [18] work introduces a two-stage BERTLex-State Space Model framework for sequence-labeling emotion intensity recognition tasks. The framework is aimed at predicting emotion dynamics in song lyrics without requiring supervision at the song level.

The literature review discusses various approaches to emotion analysis in music lyrics and text analysis. It highlights the use of datasets like MER and Music4All for training models to label lyrics with emotions, employing techniques such as transfer learning and transformer-based models. Notable works include the LyEmoBERT [9] model, which outperformed existing methods, and a multimodal music mood classification system that integrates audio and lyrics. Other studies, such as those by Yudhik Agrawal et al. [12], explore the use of deep neural networks like XLNet for emotion classification, utilizing multi-task learning for improved model performance. Fika Hastarita Rachman et al. [13] work focuses on combining psycholinguistic and stylistic features from lyrics with audio features using various classifiers for emotion recognition. Cong Jin et al.[14] introduce a Bi-LSTM network with dilated skip connections and attention mechanisms to better capture long-sequence dependencies in lyrics. A Bhattacharya, et al. propose the MoodNet architecture, a CNN-based model that analyzes both audio and lyrics for emotion detection. Additionally, Shahrzad Naseni [16] and colleagues use transformer-based models for lyric classification tasks, demonstrating the effectiveness of advanced NLP techniques. The review also emphasizes the gap in research focusing solely on lyrics for emotion classification, advocating for more targeted studies to address unique challenges in this area. It covers the calculation of emotional metrics, machine learning algorithms, feature reduction techniques like PCA, and the importance of lexicons in analyzing emotional content in lyrics.

III. Method for Music Emotion Classification for Lyric Text

The system takes music lyrics as input. These lyrics undergo a cleaning process to remove any unwanted elements and are then processed and converted into individual tokens. Feature extraction is performed on the lyrics, extracting various relevant features from the textual content, including leveraging emotion lexicons. To reduce the dimensionality of the feature space and remove irrelevant features, dimension reduction techniques are applied. Classification models are then constructed using the extracted features, both with and without dimension reduction, to accurately identify and classify the emotions embedded within the lyrics (Fig. 1).

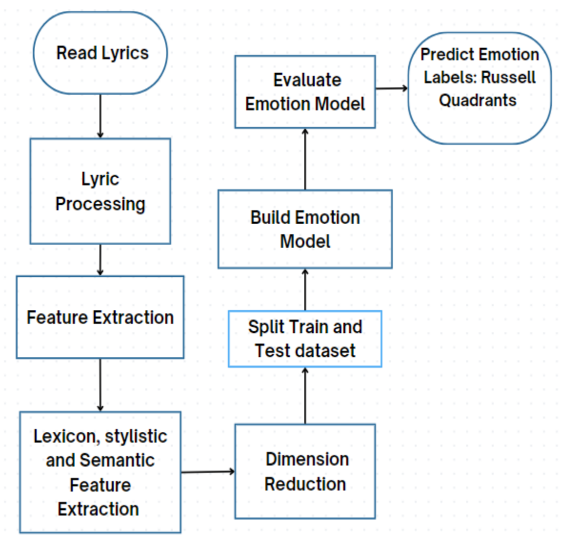

Fig. 1. Method for music emotion classification for lyrics text.

A. Datasets

In this research we have used MER and Mood Lyrics dataset. The MER Dataset [4] contains 771 song lyrics extracted from the AllMusic platform. All songs are equally distributed among the four emotion quadrants of the 2-D Russell's circumflex emotion model. The dataset is generated by linking mood tags from AllMusic to words in the ANEW dictionary. The values of A and V are assigned to each word from the ANEW dictionary. Each song is categorized within a particular quadrant if all the corresponding AllMusic tags fall within that Russell quadrant. We have also used 1935 song from Mood Lyrics [19] dataset containing English songs. The Russell Model is used to annotate each song with its four Quadrants with output classes (sad, relaxed, angry, and happy). The dataset is constructed by merging three lexicons, Word Net, WorldNet-Affect, and the ANEW dictionary, to assign Valence and Arousal scores to every word in the lyrics. Emotion annotation is performed by calculating the combined Valence and Arousal values for each lyric. Based on the VA (Valence-Arousal) values, each lyric is assigned to a specific quadrant in Russell's 2-dimensional model.

B. Feature Extraction

Text feature extraction involves transforming textual data into a numerical format suitable for use in various natural language processing (NLP) tasks. The goal is to capture the essential information and characteristics of the text in a way that machine learning algorithms can process and understand. This process includes both lexical and syntactic features, as well as stylistic elements. Techniques like Term Frequency-Inverse Document Frequency *(TF-IDF)* [20] and Bag of Words (BOW) are commonly used to capture the lexical and syntactic properties of text. The Bag of Words *(BOW)* approach represents a document by breaking it down into individual words and creating a feature vector based on the frequency of each word, ignoring word order. TF-IDF is a statistical method used in natural language processing to determine the importance of a word within a document relative to a collection of documents. It's a versatile technique that can be applied in various NLP tasks, such as text classification, information retrieval, and sentiment analysis. TF-IDF calculates a score for each word in a document based on its frequency within the document and its rarity across the entire corpus. This enables it to identify the most significant terms that contribute to the document's meaning and differentiate it from others.

In addition to statistical measures, stylistic features such as the use of slang, part-of-speech *(POS)* distributions, and verb usage help capture the linguistic style and emotional tone of the text. Style features capture the linguistic characteristics of text data, such as slang words, part-of-speech distributions, and verb usage. They are related to length of sentence, length of paragraph. We implemented length *(L)* features (word count, character count, sentence count, average word length, average sentence count). Word count, character count, sentence count, average word length and average sentence length are extracted from song lyrics. Slang words are informal and non-standard language often used in music lyrics. Counting the frequency of slang using online dictionaries can provide insight into a song's lyrical style and cultural references. We used slang count *(SL)* using Online Slang Dictionary. An example of a slang word from an online urban and slang dictionary is "lit". "Lit" is an adjective that means exciting, excellent, or highly enjoyable. It is used to describe something fun, energetic, or impressive. Slang words can vary in popularity and usage over time.

Part of speech tags *(POS)* represents the frequency of various word types (e.g., noun, verb, adjective). It can help capture the syntactic structure of a sentence. Each sentence is converted to form – list of words, list of tuples. Each tuple is represented in the form *(word, tag)*). The tag is part-of-speech, which signifies whether the word is a noun, adjective, verb. For example, some common POS tags include proper noun (NNP), noun (NN), verb (VB) and adjective (JJ), coordinating conjunction (CC), and personal pronoun (PRP). We use 21 features of POS tags using the NLTK dictionary in this work—an example of POS Tagging. Example - (Dirty, NNP), (old, JJ), and (river, NN).

The frequency of BE verbs *(BE)* in positive, negative, and interrogative sentences can provide insight into the emotion tone and sentiment conveyed by the lyrics. "Be" verbs indicate a state of being. Verbs must match subjects. Present Sentence, Negative Sentence, and Interrogative Sentence are three forms of BE Verbs. BE Verbs frequency is considered in this work, along with negative words such as "not" and question mark count. Present Sentence are sentences that start with “I am”, “You are” , “He is”, “She is” , “It is”, “We are” , “You are” and “They are”. Negative Sentence have not included with Present sentences. Examples are “I am not”, “You are not”, “and He is not” and so on. Interrogative Sentences start with “Am I?”, “Are You?”, “Is he?”, etc. By combining these features, including lexical, syntactic, and stylistic elements, classification models can gain more nuanced insights into the text. This comprehensive feature set can improve the accuracy of text classification tasks by leveraging the strength of different approaches.

C. Lexicons Feature Extraction

There are six lexicon dictionaries used in this work. The Norm of Valence, Arousal, and Dominance lemmas [21] has affective words for valence, arousal, and dominance with values ranging from 0 to 1. The NRC Affect Intensity Lexicon [22] contains words with intensity scores for Ekman's basic emotions (Anger, Fear, Anticipation, Trust, Surprise, Sadness, Joy, and Disgust). The intensity score varies between 0 and 1. A score of 1 indicates the intensity of emotion is high, whereas a score of 0 indicates the intensity of the word is low. The EmoWordNet [23] lexicon comprises of words associated with eight emotions (fear, anger, joy, sadness, and surprise) with scores ranging from 0 to 1. The Synesketch lexicons [24] contain English words annotated manually with emotion weights. It uses Ekman's six basic emotions (anger, joy, surprise, sadness, disgust, and fear).The Dictionary of Affect [25] in Language comprises 8743 words annotated in 3 dimensions: pleasantness, activation, and imagery. Pleasantness is similar to valence and activation to arousal. Imagery is a word that creates a mental picture. A score of 1 indicates the difficulty in forming a mental and score of 3 indicates it is easy to form a mental picture. The DepecheMood [26] is a high coverage and high precision emotion lexicon using distributional semantics, with numerical scores associated with more than one emotion (Afraid, Amused, Annoyed, Don’t care, Happy, Inspired and Sad), obtained from crowd sourcing, it has scores ranging from 0 to 1.

1) Lexicon features for intensity emotion weight: To extract lexicon features from lyrics, the process begins with preprocessing. This involves cleaning the text to remove any non-alphabetic characters, punctuation, and extra spaces. Next, the cleaned text is tokenized, meaning it is split into individual words (tokens). Each token is then matched with its corresponding emotion dimension/intensity score from the relevant lexicons. Finally, aggregate metrics such as the mean, maximum, minimum, and standard deviation are computed based on the scores of the mapped tokens for each lyric text. It's important to note that one word can be associated with more than one emotion, depending on the type of lexicon used. This systematic approach ensures the accurate extraction of lexicon features, which can then be used for further analysis.

In previous research, we introduced emotion lexicon features using intensity/emotion weight [8] Eq. (1) to Eq. (11). The work is expanded to include Dictionary of Affect and DepecheMood Lexicon, and hybrid feature combinations. This section details the lexicon features extracted from NRC Emotion Intensity, Synesketch, EmoWordNet, and DepecheMood; these are categorical lexicons. The categorical lexicon comprises words associated with specific emotions, each of which is linked with an intensity score.

a) Average emotion intensity/weight: Definition: It is the average emotion intensity/weight of lyrics l for emotion E. The formula for average emotion intensity is shown in Eq. (1):

$$\text{Average Emotion Intensity} = \frac{\sum \mathrm{I(W,E)}}{\mathrm{N}} \tag{1}$$

Where I (W, E) is the intensity of the word W for emotion E and N is total number of word in the lyrics. This metric provides a general sense of how strongly an emotion is expressed throughout the lyrics. It helps in understanding the overall emotional tone.

b) Maximum emotion intensity/weight: Definition: It is the maximum intensity/emotion weight of lyrics l for emotion E. This is defined in Eq. (2):

$$\text{Maximum Emotion Intensity} = \mathrm{Max}\big(\mathrm{I(W,E)}\big) \tag{2}$$

Where I (W, E) is the intensity of word W for emotion E and N is total number of word in the lyrics. This metric identifies the peak emotional intensity in the lyrics, showing the highest level of emotional expression.

c) Minimum emotion intensity/weight: Definition: It is the minimum intensity/emotion weight of lyrics l for emotion E, as defined in Eq. (3):

$$\text{Minimum Emotion Intensity} = \mathrm{Min(I(W,E))} \tag{3}$$

Where *I (W, E)* is the intensity of word W for emotion E and N is total number of word in the lyrics. This metric highlights the least intense emotional expression in the lyrics.

d) Threshold emotion word count(TEWC): Definition: It is a metric that quantifies the number of words in the lyrics, denoted as 'W, that have an intensity score greater than a predetermined threshold limit, 'TH', for a specific emotion, ‘E’. This is shown in Eq. (4):

$$\mathrm{TEWC\ (E, I, TH)} = \Sigma\ (\mathrm{W} \in \mathrm{l})\ [\mathrm{I(W, E)} > \mathrm{TH}] \tag{4}$$

Where *TEWC (E, l, TH),* represents the count of words per emotion E with a threshold for a given emotion 'E' in the lyrics 'l'. I (W, E) represents the intensity of word 'W' for the specific emotion 'E' and W ∈ I. From Eq. 4, [I(W, E) > TH] is a function that evaluates to 1 if the intensity of the word 'W' for emotion 'E' is greater than the threshold 'TH'. [I (W, E) > TH] evaluates to 0 if condition is not met. This formula calculates the total count of words in the lyrics that have an intensity surpassing the specified threshold TH for the given emotion ‘E’. Emotion Intensity score is the strength of emotional expression in the lyrics. Words surpassing the threshold are likely to have a substantial impact on the overall emotional tone. By counting words that exceed the threshold, TEWC helps identify which parts of the lyrics contribute significantly to a particular emotion. TEWC can enhance music recommendation systems by considering emotional relevance. Researchers can compare TEWC across different songs, genres, or artists to understand variations in emotional content.

e) Emotion proportion within class: Definition: It measures the proportion of a particular emotion within a specific class relative to its overall occurrence in all classes of lyrics l for emotion E, as shown in Eq. (5):

EPC = (Count of emotion within class) / (Total count of emotion across all classes)

$$EPC = (CE) / |E| \tag{5}$$

Where CE denotes the count of a specific emotion within a particular class and |E| signifies the total count of the specific emotion across all classes. This formula quantifies the relative proportion of a specific emotion within a class by dividing the count of that emotion in the class by its total count across all classes.

TABLE I. SAMPLE CALCULATION FOR LEXICON FEATURES DERIVED FROM NRC EMOTION INTENSITY LEXICON WITH JOY, SADNESS AND TRUST

Word	Joy	Sadness	Trust
joyous	0.9	0	0.23
elated	0.8	0	0
cheerful	0.95	0	0
TEWC (E, I, 0.6)	3	0	0

Let's consider the Table I provided for emotions "Joy," "Sadness," and "Trust" and their corresponding intensity score.

- Using equation (1), the average Intensity for the emotion "Joy" is calculated as (09 + 0.8 + 0.95) =0.883.
- The minimum intensity score for the emotion "Joy", using Eq. (3) is 0.8 and the maximum intensity score for the emotion "Joy" calculated using Eq. (2) is 0.95.
- For Threshold Emotion Word Count (TEWC) calculation we apply Eq. (4). TEWC (Joy, 0.9, 0.6) calculates the number of words in the lyrics with intensity greater than the threshold value (0.6) for the emotion "Joy". Using Eq. (4), TEWC (Joy, 0.9, 0.6) gives result of three and TEWC (Trust, 0.23, 0.6) gives result of zero.

2) Lexicon features for dimension emotions: This section describes lexicon features extracted from Norm of Valence, Arousal, and Dominance lemma and Dictionary of Affect. The Russell three dimension model contains dimensions of valence, arousal and dominance.

a) Average valence: It is the sum of all Valence scores divided by the total no of words (N)) in lyrics. This is shown in Eq. (6):

$$\text{Average Valence} = \frac{\Sigma V}{N} \tag{6}$$

Where Valence V is non zero and N is total number of word in the lyrics. The average valence score represents the average emotional positivity or negativity expressed in the lyrics. This metric provides a general sense of how strongly valence is expressed throughout the lyrics.

b) Average arousal: Definition: It is the sum of all arousal values (A) divided by the total no of words (N) in lyrics. This is defined in Eq. (7):

$$\text{Average Arousal} = \frac{\Sigma A}{N} \tag{7}$$

Where Arousal A is non zero and N is total number of word in the lyrics. The average arousal score represents the average calmness or excitement expressed in the lyrics. This metric provides a general sense of how strongly an arousal is expressed throughout the lyrics.

c) Average dominance: Definition: It is the sum of all dominance values (D) divided by the number of words (N) in lyrics. This is defined in Eq. (8):

$$\text{Average Dominance} = \frac{\Sigma D}{N} \tag{8}$$

Where Dominance D is non zero and N is total number of word in the lyrics. The mean dominance score represents the average level of control or power expressed in the lyrics. This metric provides a general sense of how strongly dominance is expressed throughout the lyrics.

d) Standard deviation: The standard deviation provides a measure of the dispersion or variability of the scores around the mean, indicating how consistent the emotional expression is across the lyrics. It is calculated for Valence V, Arousal A and Dominance D values of all tokens extracted from lyrics where VAD values are non-zero. Given set of dimension scores X={x1, x2, x3….an} for lexicon L. Standard deviation for score x is shown in Eq. (9):

$$\sigma = \frac{1}{N}\sum_{i=1}^{n}(x_i - x_{mean})^2 \tag{9}$$

Where σ is Standard deviation, if we apply this to valence, arousal and dominance score from lyrics. For given set of score for S= {s_1, s2…..Sn} for valence, arousal and dominance. This is defined in Eq. (10) and (11):

$$\sigma = \frac{1}{N}\sum_{i=1}^{n}(s_i - s_{mean})^2 \tag{10}$$

$$s_{mean} = \frac{1}{N}\sum_{i=1}^{n}(s_i) \tag{11}$$

Where N is the total number of score and i_s is the score of (valence, arousal, or dominance). $S_{mean}$is the mean (average) of all the scores.

e) Average imagery: Definition: It is sum of all imagery scores divided by total number of words (N) in lyrics. This is defined in Eq. (12):

$$\text{Average Imagery} = \frac{\Sigma Im}{N} \tag{12}$$

Where Imagery Im is non zero and N is total number of word in the lyrics. The mean imagery score represents the average level of mental picture evoked by the lyrics. Higher mean imagery indicates that, on average, the lyrics are more capable of creating clear mental pictures for the listener. The average scores for Pleasantness and Activation are determined in a similar manner. Additionally, the minimum and maximum scores for Pleasantness, Activation, and Imagery are also calculated. We use a consistent method to ascertain the mean, minimum and maximum values for arousal, dominance, activation, imagery, and pleasantness in lyrics. These metrics are crucial for understanding the range and variation of emotional and sensory expressions. Minimum values indicate the least intense expressions, reflecting calmness, weakness, passivity, lack of vividness, and negative affect. In contrast, maximum values capture the most intense peaks of power, energy, vividness, and positive affect. For example, minimum arousal corresponds to low emotional intensity, while maximum arousal signifies high intensity. Similarly, minimum valence reflects negative emotions, whereas maximum valence indicates positivity. This method offers a nuanced view of lyrical dynamics for deeper interpretation.

For emotion classification, we use notations as Lyrics Length Features: *L*, Slang Count: *SL,* POS tags: *POS*, BE Verbs: *BE*, Norm of Valence, Arousal and Dominance features: *AF1*, Depeche Mood features: *BF1*, EmoWordNet features: *BF2*, Synesketch features: *BF3*, Emotion Intensity features: *BF4*, Dictionary of Affect features: *BF5*.

D. Dimension Reduction

The main goal of dimension reduction [7] is to simplify the dataset without losing critical information. PCA achieves this by identifying the principal components of the data, which are linear combinations of the original features that capture the most variance. The variance values associated with each principal component reflect the amount of information that component carries; higher variance indicates more information. These variance values are calculated using Eigen values derived from the covariance matrix of the data. The first principal component accounts for the greatest variance, with each subsequent component capturing progressively less. By selecting only the most significant principal components, we can effectively reduce the dimensionality of our dataset while retaining essential information.

E. Classifiers

Random Forest [27] and Gradient Boosting are popular machine learning algorithms used for classification and regression tasks. Random forest is an ensemble method that combines multiple decision trees to make predictions. It works by constructing multiple decision trees using subsets of the training data and random feature subsets. The predictions from each tree are then combined to make a final prediction. Random Forest is known for its accuracy and ability to handle high-dimensional datasets. Gradient boosting is another ensemble method that works by iteratively adding weak learners, typically decision trees, to a model to improve its performance. Gradient boosting uses a loss function to determine the error of the current model and update the model by fitting a new tree to the residuals of the previous model. Gradient boosting is known for its high accuracy and ability to handle complex datasets but can be computationally expensive and prone to over fitting. The choice of algorithm depends on the specific problem and dataset at hand.

F. Performance Metric

Classification accuracy is a metric used to evaluate the performance of a machine learning model in classification tasks. It measures the proportion of correctly predicted instances (both true positives and true negatives) to the total number of instances in a dataset. In simple terms, classification accuracy represents how often the model makes correct predictions.

IV. Experiments and Results

We conducted extensive experiments for classification, both with and without dimension reduction, using the Mood Lyrics dataset (D1) and the MER dataset (D2). We employed Random Forest and Gradient Boost classifiers for the classification task. Due to the large number of experiments, it is not feasible to present all the results here. The features notation used for classification is given in the Lexicon sub section. The Emotion output classes are four Russell Quadrants (sad, relaxed, angry, and happy). Classification experiments were conducted using datasets D1 and D2, employing Random Forest and Gradient Boost classifiers. These experiments utilized stylistic, lexical, and length features derived from lyric text, with the feature acronyms detailed in Method subsection C.

A. Emotion Classification Accuracy by Russell Quadrants

For Table II, the combination of (Bag Of Words, Norm of Valence, Arousal and Dominance features, BE Verbs, Slang Count, Dictionary of Affect features and POS tags) and (Bag Of Words, POS tags, Norm of Valence, Arousal and Dominance, Dictionary of Affect, Synesketch features, Emotion Intensity, Slang Count, BE Verbs) features gives an accuracy of 93.16% for Gradient Boost, using Dataset D1.

TABLE II. EMOTION CLASSIFICATION BY QUADRANTS FOR DATASET D1, USING BAG OF WORDS (BOW), STYLE AND LEXICON FEATURES

Feature Sets Combinations	Random Forest	Gradient Boost
BOW + AF1 + BE + SL + BF5 + POS	64.5	93.16
BOW + AF1 + BF4 + BE + SL + BF5	61.49	80.68
BOW + AF1 + SL + BF5	57.76	91.92
BOW + AF1 + BF3 + B5 + BF4	50.93	50.93
BOW + AF1 + BF4 + SL + BF5	50.31	53.41
BOW + POS + AF1 + BF5 + SL	52.17	50.31
BOW + POS + AF1 + BF5 + BF3 + BF4 + SL + BE	54.03	93.16
BOW + POS + A1 + B4 + B5 + B3	54.03	90.68
BOW + POS + A1 + B4 + B5 + B3 + SL	57.76	63.35

In the Table III, the combination of (TF-IDF, Norm of Valence, Arousal and Dominance features, Emotion Intensity features, Slang Count , Dictionary of Affect features) and (TF-IDF, Norm of Valence, Arousal and Dominance features, Slang Count: and Dictionary of Affect features)for Gradient Boost classifier gives an accuracy of 91.92%, using Dataset D1.

TABLE III. EMOTION CLASSIFICATION BY QUADRANTS FOR THE DATASET D1, USING TF-IDF, STYLE, AND LEXICON FEATURES

Feature Sets Combinations	Random Forest	Gradient Boost
TF-IDF + AF1 + BF4 + BE + SL + BF5 + POS	60.24	91.30
TF-IDF + AF1 + BF4 + BE + SL + BF5	57.76	91.92
TF-IDF + AF1 + SL + BF5	57.96	91.92
TF-IDF + AF1 + BF3 + BF5 + EF4	55.90	54.65
TF-IDF + AF1 + BF4 + SL + BF5	56.52	47.20
TF-IDF + POS + AF1 + BF5 + SL	54.65	53.41
TF-IDF + POS + AF1 + BF5 + BF3 + BF4 + SL + BE	55.27	88.81
TF-IDF + POS + AF1 + BF4 + BF5 + BF3	57.14	90.68
TF-IDF + POS + AF1 + BF4 + BF5 + BF3 + SL	54.03	55.27

For Table IV, the combination of (TF-IDF, Norm of Valence, Arousal and Dominance features, Emotion Intensity, BE Verbs, Slang count, Lyrics Length and Synesketch) features give 78.49 % accuracy for Gradient Boost. The combination of (TF-IDF, POS tags , Norm of Valence, Arousal and Dominance features, Dictionary of Affect features, Synesketch features and Emotion Intensity features) and (TF-IDF, Norm of Valence, Arousal and Dominance features, Emotion Intensity features, BE Verbs, Slang Count , Lyrics Length Features and Synesketch features) give accuracy of 77.42% for Random Forest with Dataset D2.

TABLE IV. EMOTION CLASSIFICATION BY QUADRANTS FOR DATASET D2, USING TF-IDF, STYLE AND LEXICON FEATURES

Feature Sets Combinations	Random Forest	Gradient Boost
TF-IDF + POS + A1 + DAL + BF3 + BF4	65.59	69.89
TF-IDF + AF1 + BF4 + SL + BF3 + LF	63.44	65.59
TF-IDF + AF1 + BF4 + SL + BF3 + LF + DAL	64.51	66.67
TF-IDF + POS + AF1 + BF4 + SL + BF3	74.19	75.26
TF-IDF + POS + AF1 + DAL + BF3 + BF4	77.42	76.34
TF-IDF + POS + AF1 + DAL + BF3 + LF	68.81	75.26
TF-IDF + POS + AF1 + DAL + SL + BF4	67.74	73.11
TF-IDF + POS + AF1 + DAL + BF3 + BF4	65.59	69.89
TF-IDF + AF1 + POS + BF3 + BF2 + SL	77.42	75.26
TF-IDF + AF1 + POS + BF3 + BF2 + SL + L	72.04	68.81
TF-IDF + AF1 + BF4 + BE + SL + L + BF3	72.04	78.494

For Table V, the combination of (Bag of Words, POS Tags, Norm of Valence, Arousal and Dominance features, Dictionary of Affect features, Slang Count and Emotion Intensity) gives accuracy of 75.26% for Gradient Boost and 72.04% for Random Forest, using Dataset D2.

B. Principal Component Analysis (PCA)

PCA transforms the original data into a new set of variables, known as principal components, which are uncorrelated and ordered so that the first few retain most of the variation present in all the original variables. The variance helps determine how much information each principal component holds from the original dataset. Explained Variance shows how much of the data's variation a principal component captures. Cumulative Variance sums up the variances of components to show the total variation explained.

TABLE V. EMOTION CLASSIFICATION BY QUADRANT FOR THE DATASET D2, USING BOW, STYLE AND LEXICON FEATURES

Feature Sets Combinations	Random Forest	Gradient Boost
BOW + POS + AF1 + DAL + BF3 +BF4	72.04	69.89
BOW + AF1 + BF4 + SL + BF3 + L	68.81	62.36
BOW + AF1 + BF4 + SL + BF3 + L + DAL	64.5	64.5
BOW + POS + AF1 + BF4 + SL + BF3	70.96	67.74
BOW + POS + AF1 + DAL + BF3 + BF4	72.04	69.89
BOW + POS + AF1 + DAL + BF3 + LF	62.36	64.5
BOW + POS + AF1 + DAL + SL + F4	72.04	*75.26*
BOW + POS + AF1 + DAL + BF3 + BF4	43.08	49.46
BOW + POS + AF1 + DAL + BF3 + BF4	72.04	69.89
BOW + AF1 + POS + BF3 + BF2 + SL + L	58.06	62.36
BOW + AF1 + BF4 + BE + SL + LF + BF3	66.66	62.365

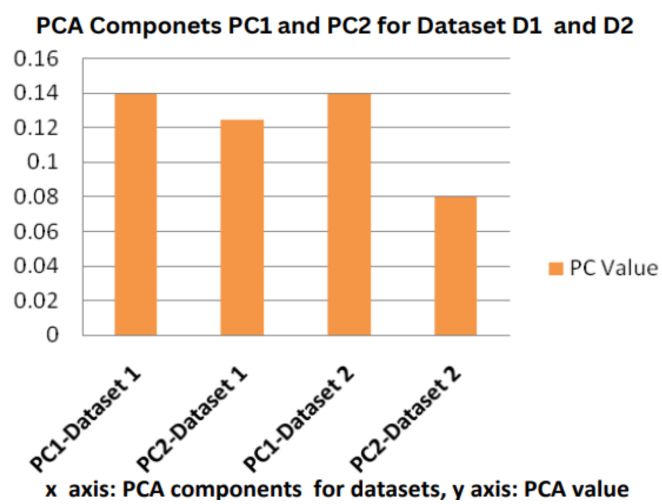

Fig. 2. PC1 and PC2 for dataset D1 and D2.

The bar chart in Fig. 2 shows the values of two main components, PC1 and PC2, for two datasets. The x axis shows PCA components for (D1 and D2); and y axis shows PCA value. In Dataset 1, PC1 explains a moderate amount of variance, showing it captures a significant part of the data's variability. PC2 has less variance than PC1, meaning the remaining variability isn't as focused in one direction. Dataset 2 has a slightly higher variance for PC1 than Dataset 1, indicating more of its variability is due to the first principal component. PC2's variance in Dataset 2 is even lower than in Dataset 1. Result indicates that Dataset 2's variability is more concentrated in the first component. While both datasets have similar PC1 values, Dataset 2 relies more on PC1, showing a more unidirectional variance. In contrast, Dataset 1 has a higher variance in PC2, indicating its variability is spread across more dimensions. This means Dataset 1 has a more balanced variance distribution between the components, while Dataset 2 leans more on PC1 for its variance explanation.

The PCA analysis of music lyrics for emotion classification highlights key features influencing the first two principal components across two datasets. In Dataset 1 (D1), PC1 is primarily influenced by character count, word count, sentence count, singular nouns (NN), and personal pronouns (PRP), pointing to detailed and intense emotional expression. PC2 is shaped by personal pronouns (PRP), adverbs (RB), present tense verbs (VBP), and general verbs (VB), indicating immediacy in emotional experiences. For Dataset 2 (D2), PC1 is similarly influenced by character count, word count, singular nouns (NN), personal pronouns (PRP), and sentence count, suggesting a focus on emotional intensity. However, PC2 is defined by personal pronouns (PRP), proper nouns (NNP), singular verbs (VBP), conjunctions (CC), and past tense verbs (VBD), which highlight introspection and narrative elements. By using 40 PCA components, the analysis effectively reduces dimensionality while capturing the nuanced features of emotional expression in lyrics, enhancing emotion classification accuracy.

Table VI shows result for emotion classification by quadrants for D1 using PCA, achieving 90% explained variance with 40 components. The Gradient Boost classifier performed the best with an accuracy of 98.13%. Table VII shows that the emotion classification by quadrants for Dataset 2 using PCA, achieving 85% explained variance with 40 components. The Gradient Boost classifier performed the best with an accuracy of 65.59%.

TABLE VI. EMOTION CLASSIFICATION BY QUADRANTS FOR D1 USING PCA

PCA Explained Variance	PCA Components	Classifiers	Accuracy
90%	40	Random Forest	96.89%
90%	40	Gradient Boost	98.13%
90%	40	Decision Tree	96.89%

TABLE VII. EMOTION CLASSIFICATION BY QUADRANTS FOR D1 USING PCA

PCA Explained Variance	PCA Components	Classifiers	Accuracy
85%	40	Random Forest	59.13%
85%	40	Gradient Boost	65.59%
85%	40	Decision Tree	47.31%

V. DISCUSSION

The classification experiments for emotion detection in lyrics using various feature sets and classifiers demonstrated notable differences in performance across datasets, classifiers, and feature combinations. The emotion detection experiments in song lyrics showed notable differences in performance across datasets, classifiers, and feature combinations. For Dataset D1, the Gradient Boost classifier consistently outperformed Random Forest, achieving the highest accuracy of 93.16% with a Bag of Words approach combined with stylistic, lexical, and length features, including Norm of Valence, Arousal, and Dominance, BE verbs, slang count, dictionary of affect features, and POS tags. This indicates that a comprehensive feature set is effective for capturing the nuances of emotion in lyrics. Using TF-IDF instead of BOW yielded similar results, with a slight decrease to 91.92% accuracy for Gradient Boost, showing that text representation (BOW or TF-IDF) has a marginal impact when combined with these features. In contrast, Random Forest did not perform well on D1, it achieved highest accuracy of 64.5%.

For Dataset D2, the results were less consistent. Gradient Boost achieved the best accuracy of 78.49% using TF-IDF, AF1, BE verbs, slang count, lyrics length, and Synesketch features, highlighting the importance of emotional content and syntactic complexity. Random Forest also performed relatively well on D2, with an accuracy of 77.42% using a different feature set. The result suggests shows that, it might be more effective for this dataset due to differences in data distribution or emotional expression. Principal Component Analysis (PCA) for dimensionality reduction further clarified the effectiveness of feature sets. For D1, using 40 PCA components, Gradient Boost achieved an accuracy of 98.13%, showing that PCA effectiveness in capturing the most informative aspects of the data. However, for D2, the highest accuracy with PCA was 65.59%, indicating that PCA might not capture the emotional nuances as effectively as in D1. Overall Dataset D1 generally achieved higher classification accuracy than D2. The difference in performance is likely due to differences in dataset characteristics. The results emphasize the need for comprehensive feature sets and suitable algorithms to capture emotional complexity in lyrics. PCA is effective for reducing dimensionality and enhancing classifier performance, especially for Dataset D1, though its impact varies with different datasets and feature representations.

The use of PCA highlights its value in reducing dimensionality and enhancing classifier performance when sufficient variance is retained, as seen with Dataset D1. However, the mixed results for Dataset D2 indicate that while PCA can improve efficiency, it may also overlook some nuances if the components do not adequately represent the data's emotional characteristics. In music emotion recognition from lyrics, our method has shown superior effectiveness compared to both transformer-based and traditional approaches. For Dataset D1, the Gradient Boost classifier achieved an accuracy of 93.16%, which improved to 98.13% with PCA.

This study builds upon our previous work [8] by incorporating a more diverse set of features and applying Principal Component Analysis (PCA) for dimensionality reduction. While the prior study primarily used lexicon-based features, the current work integrates a wide range of NLP, textual, stylistic, and lexical features. This expanded feature set includes POS tags, DAL, slang count, BE verbs, length-based attributes, and affective dictionaries. The use of the expanded feature set allows for a more comprehensive analysis of emotional nuances in song lyrics. The application of PCA for dimensionality reduction is another novel aspect, resulting in improved accuracy and model interpretability. The results of this approach are significant, with an accuracy of 98.13% achieved on Dataset D1 using PCA, compared to 93.16% without PCA. This performance surpasses previous methods in the field. For instance, Transformer-based models like XLNet [12] achieved 83% accuracy on the Mood Lyrics dataset, while traditional Bag of Words and TF-IDF approaches reported accuracies [29] between 65.49% and 67.98%. Malheiro et al. attained an F1 score of 73.6% for MER Lyrics [4] using the SVM classifier. This study achieved a remarkable accuracy of 98.13% on Dataset D1, surpassing previous research that employed multi-modal deep learning approaches by Pyrovolakis et al. (2022) [28] and Transformer-model [12]. Jiddy Abdullah [29] achieved classification accuracies of 76% and 83% using Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) classifiers, respectively, on the Mood Lyrics dataset. Malheiro et al. (2016) [4] utilized emotionally relevant features; this study's uses broader integration of stylistic, lexical, and content features which led to superior performance. Accuracy gains are crucial for fine-grained emotion detection applications, but they come with increased computational cost and complexity. While this may limit scalability, the benefits often outweigh the added resources, making the improvement worthwhile in many practical scenarios.

Despite these improvements, the study has limitations. These include potential issues with generalizability across different datasets and a lack of exploration into the impact of various hyper parameters for the Gradient Boost classifier and other dimension reduction methods. Additionally, the focus on accuracy may have overlooked other important metrics. This study shows that using various features with methods like PCA to reduce complexity effectively captures the emotions in song lyrics, improving upon the accuracy and representation issues of past approaches.

VI. Conclusion

In conclusion, this research significantly advances the field of emotion analysis in music lyrics by filling a critical gap with a focused study on the textual representation of emotions. The development of an NLP-based framework and the introduction of hybrid lexicon features have proven to be instrumental in enhancing emotion detection capabilities. The NLP-based framework and novel lexicon features significantly enhance emotion detection. Advanced classification techniques like Gradient Boost and Decision Trees, along with innovative feature extraction and dimension reduction, have achieved high accuracy in emotion classification. The superior performance of Gradient Boost classifiers highlights the importance of choosing appropriate algorithms for handling high-dimensional data. This research provides valuable insights into understanding emotional expressions in music. However, it also faces limitations, such as reliance on specific features and the resource-intensive nature of initial computations for large datasets. Additionally, model generalizability across diverse datasets remains a challenge. Future research directions could include exploring other emotion datasets, feature engineering and integrating deep learning approaches.

References

[1] A. Kawakami, K. Furukawa, and K. Okanoya, “Music evokes vicarious emotions in listeners,” Front. Psychol., vol. 5, p. 431, May 2014, doi: 10.3389/fpsyg.2014.00431.

[2] M. de Witte, A. da Silva Pinho, G.-J. Stams, X. Moonen, A. E. R. Bos, and S. van Hooren, “Music therapy for stress reduction: a systematic review and meta-analysis,” Health Psychol. Rev., vol. 16, no. 1, pp. 134–159, 2022, doi: 10.1080/17437199.2020.1846580.

[3] S. Zad, M. Heidari, H. James Jr, and O. Uzuner, “Emotion detection of textual data: An interdisciplinary survey,” in 2021 IEEE World AI IoT Congress (AIIoT), May 2021, pp. 255–261, doi: 10.1109/AIIoT52608.2021.9454192.

[4] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva, “Emotionally-relevant features for classification and regression of music lyrics,” IEEE Trans. Affective Comput., vol. 9, no. 2, pp. 240–254, 2016, doi: 10.1109/TAFFC.2016.2598569.

[5] M. M. Bradley and P. J. Lang, Affective norms for English words (ANEW): Instruction manual and affective ratings, vol. 30, no. 1, Univ. Florida, Tech. Rep. C-1, 1999.

[6] J. A. Russell, “Core affect and the psychological construction of emotion,” Psychol. Rev., vol. 110, no. 1, pp. 145–172, 2003, doi: 10.1037/0033-295X.110.1.145.

[7] K. N. K. Singh, D. Devi, H. M. Devi, and A. K. Mahanta, “A novel approach for dimension reduction using word embedding: An enhanced text classification approach,” Int. J. Inf. Manage. Data Insights, vol. 2, no. 1, p. 100061, 2022, doi: 10.1016/j.jjimei.2022.10006.

[8] A. Ara and V. Rekha, “Analyzing the impact of lexicon-based features for emotion classification,” Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 4, pp. 677–687, 2024.

[9] V. R. Revathy, A. S. Pillai, and F. Daneshfar, “LyEmoBERT: Classification of lyrics’ emotion and recommendation using a pre-trained model,” Procedia Comput. Sci., vol. 218, pp. 1196–1208, 2023.

[10] S. A. Suresh Kumar and R. Rajan, “ Transformer-based automatic music mood classification using multi-modal framework,” J. Comput. Sci. Technol., vol. 23, 2023.

[11] P. Padmane, K. Agrahari, R. Kesharwani, K. Mohitkar, S. Khan, and N. P. Kamale, “Prediction of song mood through lyrics,” Open Access Repository, 9(6), 137-141,2022

[12] Y. Agrawal, R. G. R. Shanker, and V. Alluri, “Transformer-based approach towards music emotion recognition from lyrics,” in European Conference on Information Retrieval, Mar. 2021, pp. 167–175, doi: 10.48550/arXiv.2101.02051.

[13] F. H. Rachman, R. Sarno, and C. Fatichah, “Hybrid approach of structural lyric and audio segments for detecting song emotion,” Int. J. Intell. Eng. Syst., vol. 13, no. 1, Feb. 2020, doi: 10.22266/ijies2020.0229.09.

[14] C. Jin, Z. Song, J. Xu, and H. Gao, “Attention-based Bi-DLSTM for sentiment analysis of Beijing Opera lyrics,” Wireless Commun. Mobile Comput., vol. 2022, p. 1167462, 2022.

[15] Bhattacharya and K. V. Kadambari, “A multimodal approach towards emotion recognition of music using audio and lyrical content,” arXiv preprint, vol. 1811, p. 05760, 2018.

[16] Naseri, S., Reddy, S., Correia, J., Karlgren, J., & Jones, R. (2022). The Contribution of Lyrics and Acoustics to Collaborative Understanding of Mood. arXiv:2207.05680, https://doi.org/10.48550/arXiv.2207.0568

[17] L. Parisi, S. Francia, S. Olivastri, and M. S. Tavella, "Exploiting synchronized lyrics and vocal features for music emotion detection," arXiv preprint arXiv:1901.04831, 2019.

[18] Y. Song and D. Beck, "Modeling Emotion Dynamics in Song Lyrics with State Space Models," arXiv preprint arXiv:2210.09434, 2022.

[19] E. Cano and M. Maurizio, "MoodyLyrics: A Sentiment Annotated Lyrics Dataset," in Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, Hong Kong, 2017, pp. 118–124.

[20] F. Karabiber , TF-IDF-Term Frequency- Inverse Document Frequencey , https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/

[21] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," Behavior Research Methods, vol. 45, pp. 1191-1207, 2013, doi: 10.3758/s13428-012-0314-x.

[22] S. M. Mohammad, "Word Affect Intensities," in Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018), Miyazaki, Japan, May 2018, doi: 10.48550/arXiv.1704.08798.

[23] G. Badaro, H. Jundi, H. Hajj, and W. El-Hajj, "EmoWordNet: Automatic expansion of emotion lexicon using English WordNet," in Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, 2018, pp. 86–93, doi: 10.18653/v1/S18-2009.

[24] U. Krcadinac, P. Pasquier, J. Jovanovic, and V. Devedzic, "Synesketch: An open source library for sentence-based emotion recognition," IEEE Transactions on Affective Computing, vol. 4, no. 3, pp. 312–325, 2013, doi: 10.1109/T-AFFC.2013.18.

[25] C. Whissell, M. Fourner, R. Pelland, D. Weir, and K. Makarec, "A dictionary of affect in language: IV. Reliability, validity, and applications," Perceptual and Motor Skills, vol. 62, no. 3, pp. 875-888, 1986.

[26] J. Staiano and M. Guerini, "Depechemood: A lexicon for emotion analysis from crowd-annotated news," arXiv preprint arXiv:1405.1605, 2014.

[27] Logistic Regression Vs. Random Forest Classifier https://www.geeksforgeeks.org/logistic-regression-vs-random-forest-classifier/

[28] K. Pyrovolakis, P. Tzouveli, and G. Stamou, "Multi-modal song mood detection with deep learning," Sensors, vol. 22, no. 3, p. 1065, 2022.

[29] J. Abdillah, I. Asror, Y. F. A. Wibowo, et al., "Emotion classification of song lyrics using bidirectional LSTM method with GloVe word representation weighting," Journal RESTI (Rekayasa Sistem Dan Teknologi Informasi), vol. 4, no. 4, pp. 723–729, 2020.