International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Analyzing the Impact of Lexicon Based Features for Emotion Classification

Affreen Ara [1], Rekha V.*[2]

Submitted:13/03/2024 **Revised**: 28/04/2024 **Accepted**: 05/05/2024

Abstract: Emotions are psychological states that are frequently represented through actions, words or text. Emotion analysis is a method for deciphering a text to identify the feelings conveyed within it. Identification of emotion(s) contained in music lyrics is a complex process. The emotion model plays a key role in the design of emotion identification algorithms. Several text features are defined and used with machine learning algorithms for labelling lyrics based on emotion. Most of these features are defined following natural language processing concepts. Emotion lexicons play an important role in mapping words that appear in lyrics with discrete and continuous emotions. In this work, we analyze the impact of features derived from lexicons in identifying the underlying emotion of lyrics. Experiments are carried out with emotion-annotated datasets and different lexicons. Classification models are built with the lexicon features. The results obtained highlight the impact of Lexicon based features on classification accuracy. For the design of robust and efficient emotion classifier, the lexicon features need to be combined with other text based features.

Keywords: *Emotion Classification, Emotion, Lexicon, Lexicon Features*

1. Introduction

The rapid transition from analogue to digital in the music industry has resulted in a significant growth in the volume of music consumption. The enormous collection of music in online repositories has transformed the way people access music. For example, the selection of songs can be based on emotion. The progress in digital technology helps in recognizing emotions associated with music albums or lyrics thereby help listeners to choose songs that align with their feelings. With the amount of music being created and consumed increasing daily, processing this vast amount of music effectively is becoming a challenge. This requires the assistance of Music Information Retrieval (MIR)[1], an interdisciplinary science of extracting information from music. MIR facilitates classification, recommendation systems, and music generation.

Emotion recognition [2] is a challenging task due to the vast differences in music emotion and emotion models. Emotions are psychological states that are frequently represented through actions, words or text. Emotion analysis is a method for deciphering a text in order to identify the feelings conveyed within it. There is no straightforward way to find the most effective approach for the problem, because different approaches may work better for specific purposes. Music lyrics are a rich source of emotional and affective information. The feeling so the writer or performer are expressed in song lyrics. They induce extensive emotions from the listener. Emotion analysis from lyrics heavily depends on Natural Language Processing concepts. In this paper we study the effectiveness of lexicon-based features for emotion identification from music lyrics.

The most prominent way of performing emotion analysis is by employing knowledge and machine-learning methods. In the former, labels are assigned to words in texts using an emotion dictionary or lexicon. A lexicon associates words or expressions with emotion labels [3]. The foundation of this strategy is an emotion vocabulary. The latter requires a labeled dataset to build emotion identification models [4]. To determine the relationship between emotion and music, sentiment analysis and emotion identification are used. These techniques employ Natural Language Processing and text analysis methods to analyze the relationship between certain music parameters and emotions.

General-Purpose Emotion Lexicons like WorldNet-Affect[5] and SentiWordNet [6] are widely used techniques for emotion identification from the text. There are two types of affective lexicons –sentiment and emotion lexicons. Sentiment lexicons primarily capture words' polarity or sentiment orientation, indicating whether they convey a positive, negative, or neutral sentiment. Emotion lexicons focus on identifying and categorizing different emotions expressed in text, such as happiness, sadness, anger, fear, surprise, and more. Sentiment dictionaries are Word Net-Affect, SentiWordNet, and Vader[7].In contrast, Affective Norm for English Word (ANEW)[8], NRC Valence Arousal and Dominance (NRC VAD) [9], NRC

[1] *Christ University, Karnataka – 560029, INDIA*
ORCID ID : 0000-0001-6322-1001
[2] *Christ University, Karnataka – 560029, INDIA*
ORCID ID : 0000-0001-5275-5235

** Corresponding Author Email: affreen.ara@res.christuniversity.in*

Emotion Intensity Lexicon[10], and EmoWordNet [11] are examples of emotion dictionaries. Researchers have recently used Affective Norms for English words, NRC VAD as an emotion lexicon for dimensional models, and NRC Emotion Intensity and EmoWordNet for categorical models.

It is essential to decide on an emotion model before we study music emotion. Every model has a different approach to interpreting and quantifying emotion. Emotion models are of three types: discrete, dimensional, and miscellaneous. The discrete model is associated with the theory of primary emotion. It states that all basic emotions are derived from innate emotions such as anger, fear, disgust, and happiness. Some examples of discrete models are Hevner's adjective[12], Watson Tellegen and Clarke [13], and Plutchik Model [14]. The dimensional model maps emotion states to 2D/3D space. Russell's model [15] is two-dimensional, with valence being the degree of positivity. Valence ranges from negative to positive state. Arousal is the degree of calmness that ranges between calm and excited state .Third-dimension dominance [16] is described as dominant, controlling, influential, etc.; submissiveness is described as submissive, influenced, controlled, etc. Thayer model [17] and Pleasure Arousal and Dominance three-dimension model are other examples of Dimension Models.

In this paper different types of lexicon features are explored to study their effectiveness in identifying emotion from music lyrics. The study explores the impact of lexicon-based features on the classification accuracy of emotion in music lyrics. Feature extraction is centered on the mapping of words in lyrics to a chosen lexicon. Five lexicons are considered for the study and multiple features are defined. The extracted features are employed to build emotion classification models to analyze the discriminative power of them. A concise review of related works is given in the next section.

The paper examines different approaches to address the multi-emotion classification problem.

2. Literature Review

Bandhakavia et al.[18]revised earlier work on Domain-Specific Emotion Lexicon (DSEL) for feature extraction to classify text into emotion classes using machine learning techniques. Tengetal.[19]propose a context-sensitive lexicon- based technique built on a weighted-sum model. They use are current neural network to study the sentiment strength, intensification, and negation of lexical sentiments to find the sentiment value of phrases. In the model operations specifics, word weight is used as a hidden variable. Bruyne et al.[20] propose an expanded, unified lexicon with 30,273distinct words and used Bi-LSTM architecture. Eight pre-existing English emotion lexicons are combined into one larger joint emotion lexicon, using a multi-view variation auto encoder. Results show that the latent space's selected dimension could be associated with emotion dimensions existent in the source. The addition of lexical features improves the performance of simple word embedding models.

Tao et al.[21]propose a convolution neural network(CNN) to integrate lexicons with a rethinking mechanism. It can model all the characters tied with the probable words that match the sentence in parallel. The proposed model can improve the networks by adding a feedback layer that sends the high-level features back to the network. Cheng [22] study use CNN-LSTM (convolution neural networks-long short-term memory) network study to classify emotions. A multi modal ensemble learning method based on stacking is proposed in this work, it combines music audio and lyrics. The model achieves accuracy of 78%. Sebastian et al .[23]aims to investigate how understanding lyrics affects the way people perceive emotions in music. The study focuses on a small collection of songs that have already been annotated with emotions. Agarwal et al.[24]propose a transformer-based system for building an XLNet network for emotion classification with music lyrics. A Deep Learning system is created by Delbouys et al.[25], combining Convolution Neural Network, Long Term Short Memory ,and Fully Connected layers for emotion classification using music lyric and audio source. Abdillah et al. examines different approaches to address the multi-emotion classification problem.

In this literature, authors reported several text-based features extracted from music lyrics for the identification of underlying emotions. Emotion identification is closely linked to the emotion model. In this context emotion lexicon plays a crucial role to identify the emotion embedded in lyrics. This paper analyzes the effectiveness of lexicon-based features in emotion identification.

3. Methodology

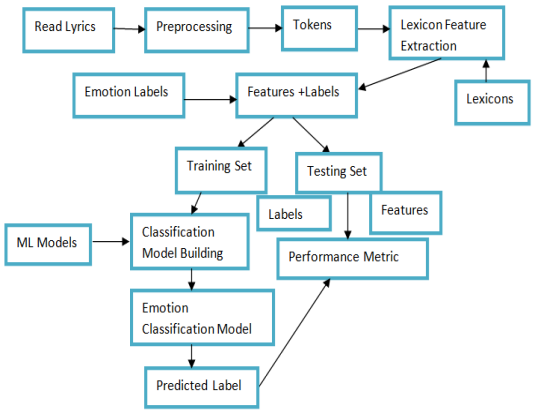

Fig 1-Emotion Classification Methodology

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.

3.1 Datasets

The lyrics (English) are taken from the Moody Lyrics [27] and Ricardo Malheiro [4] datasets. The Ricardo Malheiro ,et al (2017) dataset contains 771 emotion-labeled lyrics extracted from the AllMusic platform and validated by experts. The dataset contains words annotated with Russell quadrants Q1,Q2, Q3 and Q4. We used 1938 songs from the Mood Lyrics. The Mood Lyrics sentiment annotated dataset songs is annotated with four Russell Quadrants of Russell's 2D model with output classes as sad, relaxed, anger and happy.

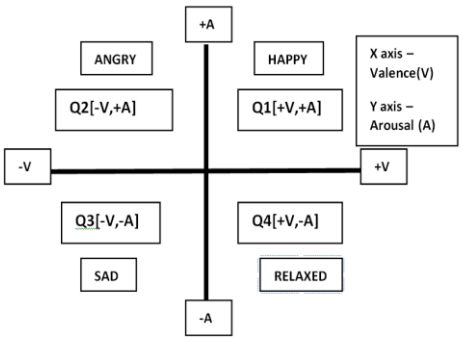

Fig 1-Emotion Classification Methodology

The Russell Model has dimensions of valence and arousal, the x-axis represented by valence (V) and the y-axis represented by arousal (A). The Q1 quadrant has high valence and high arousal [+V,+A], Q2 quadrant - [-V,+A], Q3 quadrant -[-V,-A] and Q4 quadrant - [+V,-A].

We combined the two datasets resulting in 2680 music lyrics annotated with Russell Quadrants. Duplicate lyrics are removed and output classes are made uniform by changing the labels of Mood lyrics into Ricardo Malheiro dataset labels. The output class angry is renamed to Q2, happy becomes Q1, sad becomes Q3 and relaxed becomes Q4. . The output class is Q1, Q2, Q3 and Q4. Dataset labeled D1 is the Ricardo Malheiro dataset, D2 is the Mood Lyrics dataset and D3 is the combined dataset containing 2680 lyrics.

3.2 Lexicon Datasets

In this section an overview of lexicon is given. This is followed by the definition of a set of lexicon features.

The five lexicons used in this work are L1: “Norms of valence, arousal and dominance for 13,915 English lemmas”, L2: “NRC Valence Arousal and Dominance”, L3: “NRC Affect Intensity”, L4: “EmoWordNet lexicon” and L5:” Synesketch”.L1 and L2 are dimension lexicons based on the Russell Model. Remaining are discrete lexicons containing intensity scores or emotion weight for each emotion.

The Norms of valence, arousal, and dominance for 13,915 English lemmas database [28] contains affect annotations for 13,915 words. The values of valence, arousal and dominance**,** range from 1-10. The lexicon also contains standard deviation values for valence, arousal and dominance.

The NRC VAD Lexicon [9] has affect annotations for English words with human ratings of valence, arousal, and dominance. It has more than 20,000 English words and uses Best–Worst Scaling [29] method to get fine-grained scores to solve the consistency problems that arise from traditional rating scale methods of annotation. The values for valence, arousal and dominance values range from 0-1.

The NRC Affect Intensity Lexicon [10] contains 60,000 words with intensity scores for eight basic emotions - anger, fear, anticipation, trust, surprise, sadness, joy, and disgust.

The EmoWordNet is a new [11] version of Depeche mood [30]. It is created from crowd sourcing news articles from Rappler.com. It has 67k words and 58k synsnets. annotated with 8 emotion (afraid, amused , angry, don’t care , happy ,inspired and sad.

The Synesketch lexicon [31] comprises of 5123 English words annotated manually with emotion weights. It uses Ekman’s six basic emotions (anger, joy, surprise, sadness disgust and fear). The values range from 0 to 1. The values range from 0 to 1.

3.3 Lexicon features

This section describes lexicon features used in the work. The features and defined based on “Valence Arousal Dominance”, and “Emotion intensity”, or “emotion weight” .

3.3.1 VAD Features for defined with respect to Lexicon L1 (Norms of valence, arousal, and dominance), and L2 (NRC VAD lexicon) using Valence , Arousal, and Dominance (VAD) values .

Mean Valence*:* It is the mean value of Valence V of all tokens extracted from lyrics L.

$$\mu_{V=}\frac{\sum Vw}{|W|}\ where\ Vw\ \neq 0 \quad (1)$$

Vw_{w} is the value of Valence of word w, for w $\in$ lyric L and |W| is the count of words W where valence is non zero

Mean Arousal: It is the mean value of Arousal A of all tokens extracted from lyrics L.

$$\mu_{A=}\frac{\sum Aw}{|W|}\ where\ Aw\ \neq 0 \quad (2)$$

Aw_{w} is the value of Arousal of word w, for w $\in$ lyric L and |W| is the count of words W where Valence is non zero.

Mean Dominance*:* It is the mean value of Dominance D of all tokens extracted from lyrics L.

$$\mu_{D=}\frac{\sum Dw}{|W|}\ where\ Dw\ \neq 0 \quad (3)$$

Dw_{w} is the value of Dominance of word w, for $\in$ lyrics L and |W| is the count of words W where Valence is non zero.

Minimum VAD : It is the minimum values of Valence V, Arousal A, and Dominance D of all tokens extracted from lyrics L where VAD are non-zero.

Maximum VAD*:* It is the maximum values of Valence V, Arousal A and Dominance D of all tokens extracted from lyrics L where VAD are non-zero.

Standard Deviation*:* Standard deviation is calculated for Valence V, Arousal A and Dominance D values of all tokens extracted from lyrics where VAD values are non-zero.

Standard deviation for Valence, Arousal and Dominance are calculated using the following:

$$\sigma_V = \frac{(\sqrt{(V-\mu_V)\text{^}2}}{|W|} \quad (4) \qquad \sigma_A = \frac{(\sqrt{(V-\mu_A)\text{^}2}}{|W|} \quad (5) \qquad \sigma_D = \frac{(\sqrt{(V-\mu_D)\text{^}2}}{|W|} \quad (6)$$

Dimension Lexicon Feature calculation is shown in Table 6. Each token in lyrics is mapped to lexicon L2 attributes. Table 6 illustrates how VAD features are assigned to lyric L1.

Table 6 VAD Features for Lyric L1

Lyric	*Token*	*Valence*	*Arousal*	*Dominance*
L1	w1	**8.2**	**2.1**	**3.4**
L1	w2	3.6	**4.3**	4.8
L1	w3	**2.5**	2.9	**7.3**
L1	w4	3.3	3.8	4.4

Lyric L1 contains four tokens with valence, arousal and dominance values. For minimum valence, the lowest value is selected from the valence column; for maximum valence, the highest value is selected. Mean and Standard deviation (SD) for valence is calculated using eq (1) and eq(5). Lyrics L1 feature calculation for valence attribute:

Mean (valence) = (8.2+3.6+2.5+3.3)/4=4.4

Min (valence) =2.5

Max (valence) =8.2

Standard Deviation (valence) =2.23

The same calculation is applied to arousal and dominance values. Table 7 showss a representation of Valence, Arousal, and Dominance (VAD) features for Lyric L1

Table7 Numerical Features sample for Lexicon L1 and L2

Features	*Valence*	*Arousal*	*Dominance*
Mean	4.4	3.27	4.95
Min	2.5	2.1	3.4
Max	8.2	4.3	7.3
SD	2.33	0.97	1.65

Each lyric is assigned lexicon feature. For each VAD value of lyrics Mean, Max, Min and SD is calculated.

Lyric L1 has twelve lexicon features (Mean_Valence, SD_Valence, Min_Valence, Max_Valence, Mean_Arousal, SD_Arousal, Min_Arousal, Max_Arousal, Mean_Dominance, SD_Dominance, Min_Dominance, Max_ Dominance.

3.3.2 Emotion Intensity or Weight Features

Discrete Features for Lexicons L3 (Affect Intensity lexicon), L4 (EmoWordNet) and L5 (Synesketch) are defined below.

Average Emotion Intensity Score(Φ_I^E): For each discrete emotion we calculate the average intensity of all tokens extracted from lyrics L.

$$\Phi_I^E = \frac{\sum I_W^E}{|W|_{I_W^E \neq 0}} \quad (7)$$

I_W^E is the intensity of word w for emotion E

Word Count per Emotion with a Threshold(WC_T^E): It is defined as the number of words W in the lyrics L having

intensity I greater than a set threshold T for a given emotion E.

$$WC_{T=}^{E} \; Count(W \mid I_W^E \geq T) \quad (8)$$

I_W^E is the intensity of Emotion E for word w and T is the threshold.

Normalized Word Count per Emotion:(CF^F) It is the count of emotion per class divided by total Emotion count |E|.

$$C^E = Count(W|Ew = E) \quad (9)$$

$$CF^F = \frac{C^E}{|E|} \quad (10)$$

Discrete Lexicon Feature calculation is shown in Table 8. Each token in lyrics L1 is mapped to lexicon L5 attributes. Table 8 illustrates how Weight Features are associated with music lyric L1.

Table 8 Word Count per Emotion with a Threshold(WC_T^E) and Average Emotion Intensity Score(Φ_I^E) Calculation

Token	***Happiness***	***Sadness***	***Anger***	***Fear***	***Disgust***	***Surprise***
w1	0.35	0	0	0	0	0
w2	**0.61**	0	0	0	**0.67**	0
w3	**0.80**	0	0	**0.80**	0	0
w4	**0.70**	0	**0.45**	0.20	0.	0
Φ_I^E	0.615	0	0.11	0.25	0.167	0
$WC_{0.40}^E$	3	0	1	1	1	0

In Lyric L1, four tokens are associated with emotions (happiness, sadness, anger, fear, disgust, surprise values), erived from the Synesketch lexicon L5. The Average Emotion Weight (Φ_I^E) is determined by averaging the scores for each emotion (happiness, sadness, anger, fear, disgust, surprise) within Lyric L1. It is shown in Table 8.

Table 8 shows the Word Count per Emotion with a Threshold (WC_T^E) for emotion (happiness, sadness, anger, fear, disgust, surprise). E is emotion and T threshold.

For calculation of the condition $WC_{0.40}^{Happiness}$, the $I_W^E \geq T$ Emotion, E = happiness and T=0.40 is satisfied for w2, w3, w4 shown in Table 8 column *Happiness.*

For token w2, I_W^E =0.61 which is greater than 0.40. Three tokens satisfy same condition, so $WC_{0.40}^{Happiness} = 3$. For emotion sadness and surprise, no tokens satisfy condition so $WC_{0.40}^{Sadness}$ and $WC_{0.40}^{surpise}$=0. Word Count per Emotion with a Threshold feature (WC_T^E) is applied to Affect Intensity, Emo Word Net and Synesketch lexicons. A threshold value of 0.40 is used for EmoWordNet and Synesketch lexicon whereas for Affect Intensity lexicon threshold value of 0.60 is taken.

Table 9: Normalized Word Count per Emotion: CF^F (Discrete Features for Lyric L1)

Word	***Anger***	***Anticipation***	***Joy***	***Disgust***	***Surprise***	***Trust***	***Sadness***	***Fear***
w1	0	**0.45**	0	0	0	0	0	0
w2	**0.45**	0	0	0	0	0	0	0
W3	0	0	0	**0.72**	0	0	0	0
CF^F	01-Aug	01-Aug	0/8	01-Aug	0/8	0/8	0/8	0/8

In the Table 9, for token w1, the total count of *Anticipation* class is one and total number of classes is eight, so $CF^F = 1/8$. Total no. of feature vector for $CF^F$is eight.

3.4 Performance metric

There are several performance metrics used to evaluate the performance of a classification model. Accuracy is an important metric; it is the degree of precision with which the classification system allocates items or instances to their correct classes.

Overall Accuracy: The overall accuracy of the model is the ratio of correctly predicted instances to the total instances.

The Confusion Matrix is a table used in classification to evaluate the performance of a machine learning model. It provides a detailed breakdown of the model's predictions compared to the actual classes.

Table 10 Confusion Matrix for 4 output classes Q1,Q2,Q3 and Q4

		Actual			
		Class Q1	***Class Q2***	***Class Q3***	***Class Q4***
Predicted	Class Q1	TP_Q1	FP_Q1	FP_Q1	FN_Q1
	Class Q2	FN_Q2	TP_Q2	FN_Q2	FP_Q2

Class Q3	FP_Q3	FP_Q3	TP_Q3	FP_Q3
Class Q4	FN_Q4	FP_Q4	FP_Q4	TP_Q4

Table 10 shows confusion matrix for a 4 class problem. It has 4 classes of Russell Model Q1, Q2 ,Q3 and Q4 displayed in 4X4 grid. The actual values are presented in x axis and predicted values across y axis. Accuracy score is calculated considering total True Positive, False Positive.

True Positives (TP_i): Total no of instances of class i that are correctly predicted as class i.

5. Result and Discussion

False Positives (FP_i): Total no of instances belonging to class i that are incorrectly predicted as class i.

False Negatives (FN_i): Total no of instances of class i that are incorrectly predicted as not belonging to class i.

This matrix can also assess the model's performance for each individual class. From these values various metrics for each class, can be calculated such as precision and F1 score.

		Actual			
		Class Q1	***Class Q2***	***Class Q3***	***Class Q4***
Predicted	**Class Q1**	**47**	40	37	26
	Class Q2	23	**52**	18	17
	Class Q3	42	36	**34**	14
	Class Q4	20	25	13	**89**

Table 11 Example for Confusion Matrix

Table 11 shows numerical values for confusion matrix.

True Positive (TP_i)= True Positive_Q1 + True Positive _Q2+ True Positive _Q3+ True Positive _Q4

=47+52+34+89 =217

False Positive (FP_i)= False_Positive_Q1 + False_Positive _Q2+ False_Positive _Q3+ False_Positive _Q4

=(40 + 37 + 26) + (23 + 18 + 17) + (42 + 36 + 14) + (20 + 25 +13) =311

Accuracy is calculated as ratio of TP divided by sum of TP and FP.

= (TP/TP+FP)*100= 217/(217+311)*100 = 40%

Table 12: Classification Accuracy of Lexicon L1 for µVAD ,σVAD, MINVAD ,MAXVAD

Feature Set	**µVAD**			σVAD			**MINVAD**			MAX*VAD*		
Datasets	***D1***	***D2***	***D3***	***D1***	***D2***	***D3***	***D1***	***D2***	***D3***	***D1***	***D2***	***D3***
Gradient Boost	40.05	45.63	**44.44**	39.52	46.66	41.42	41.42	30.00	41.84	34.10	40.09	36.56
Random Forest	41.09	***48.32***	42.16	**41.90**	***50.00***	43.51	40.00	33.33	40.59	37.47	46.31	39.19
Decision Tree	36.69	39.60	39.93	31.90	40.00	36.82	34.29	33.33	34.31	32.56	38.26	39.37

Table13: Classification Accuracy of lexicon L2 for μVAD, σVAD ,MINVAD ,MAXVAD

Feature Set	***Mean_VAD(μVAD)***			σVAD			***MIN*VAD**			MAX*VAD*		
Datasets	***D1***	***D2***	***D3***	***D1***	***D2***	***D3***	***D1***	***D2***	***D3***	***D1***	***D2***	***D3***
Gradient Boost	40.05	45.63	**44.40**	36.95	42.28	**44.40**	40.56	48.32	40.29	34.10	40.93	36.56
Random Forest	**41.09**	48.32	42.16	38.24	***49.66***	42.16	38.56	47.65	37.87	37.47	*46.31*	36.19
Decision Tree	36.69	39.60	39.93	30.23	40.94	38.62	33.33	40.27	36.94	32.56	38.26	39.37

Table 14 : Classification Accuracy of Lexicon L3 for WC^E and CF^F

Feature Set	$\boldsymbol{WC_T^E}$			$\boldsymbol{\emptyset_I^E}$			CF^F		
Datasets	***D1***	***D2***	***D3***	***D1***	***D2***	***D3***	***D1***	***D2***	***D3***
Gradient Boost	46.58	***47.20***	44.72	45.96	49.07	45.34	***54.03***	50.93	***50.31***
Random Forest	29.41	29.41	29.41	29.41	41.18	23.53	41.17	35.29	35.29
Decision Tree	43.8	42.70	43.26	47.19	***50.0***	37.64	51.68	***51.12***	49.44

Table 15: Classification Accuracy o f Lexicon L4 and L5 using $\boldsymbol{WC^E}$, and Φ^E

	WC_T^E *(L4)*			Φ_I^E ***(L4)***			WC_T^E $(L5)$			Φ_I^E(L5)		
Datasets	***D1***	***D2***	***D3***	***D1***	***D2***	***D3***	***D1***	***D2***	***D3***	***D1***	***D2***	***D3***
Gradient Boost	36.36	38.46	42.59	35.02	38.92	34.14	39.32	39.45	38.98	38.54	40.13	39.35
Random Forest	**41.56**	28.21	**47.04**	36.09	**40.94**	34.89	40.36	**44.22**	40.87	**41.15**	41.50	**41.05**
Decision Tree	35.93	30.77	36.67	35.29	39.60	32.65	34.11	42.18	35.40	28.39	31.97	38.40

The classification accuracy for the four features and the different classifiers are given in Table 12. For data set D1 and D2, Standard Deviation with Random Forest classifier is more effective. For D3 (combined data set), Mean value with Gradient Boost is giving better value.

For the Lexicon L2, the mean, Standard Deviation, Min and Max features are derived for each data set. The classification accuracy for the four features and the different classifiers are given in Table13. For the dataset D1 Mean with Random Forest is the best option. For database D2, Standard Deviation with Random Forest classifier is more effective. For D3, both Mean and SD, with Gradient Boost provide the best accuracy.

For the Lexicon L3, the Word Count per Emotion with a Threshold, Average Emotion Intensity Score and Normalized Word Count per Emotion features are derived for each data set. The classification accuracy for the three features and the different classifiers are given in Table 14. For all the datasets, Normalized Word Count per Emotion is the best feature, with Gradient Boost (D1, D3) and Decision Tree (D2).

For the Lexicon L4, the Average Emotion Intensity Score and Word Count per Emotion with a Threshold features are derived for each data set. The classification accuracy for the two features and the different classifiers are given in Table 15. For the dataset D1and D3, Word Count per Emotion with Random with a Threshold and Random Forest classifier gives the best accuracy. For D2, Average Emotion Intensity Score Forest classifier is more effective.

For the Lexicon L5, the Average Emotion Intensity Score and Word Count per Emotion with a Threshold features are derived for each data set. The classification accuracy for the two features and the different classifiers are given in Table 4. For the dataset D2, Word Count per Emotion with a Threshold with Random Forest is the best option. For D1 and D3, Average Emotion Intensity Score with a Threshold and Random Forest classifier is more effective.

The study is further extended by combining the different features. The classification results for different combinations of features derived from lexicon L3 is summarized in Table 16. Experiments are conducted with datasets D1 and D2 only.

For classification Gradient Boost and Random Forest algorithms are selected. For dataset D1. the combination Word Count per Emotion with a Threshold and Average Emotion Intensity Score with a Random Forest resulted in highest accuracy of 57.04%. For Dataset D2, the combination of Word Count per Emotion with a Threshold and Average Emotion Intensity Score with Random Forest classifier is best result of 61.74%.

Table 16: Classification Accuracy of feature combinations for Lexicon L3

Features	***Gradient Boost***		***Random Forest***	
	D1	***D2***	***D1***	***D2***
WC^E+ Φ^E T I	46.23	42.60	***57.04***	***61.74***
WC^ECF^F T+	50.09	29.41	23.52	41.18
Φ^E+CF^F I	50.93	50.31	41.18	23.53

Table 17: Classification Accuracy of feature combinations for Lexicon L1

Classifiers	**Gradient Boost**		**Random Forest**	
Lexicon Features for L1	***D1***	***D2***	***D1***	***D2***
MINVAD+μVAD	42.11	44.19	52.34	56.39
MINVAD+μVAD	44.70	41.86	*57.71*	58.39
MINVAD+σ_{VAD}	39.90	38.50	49.66	50.34
MAXVAD+σ_{VAD}	39.27	36.18	44.29	38.93

μVAD+σVAD	44.1	42.38	45.63	53.02
MAXVAD+MINVAD	41.18	42.12	51.67	58.39
MAXVADMINVAD+μVAD	45.73	44.44	51.67	58.39
μVAD+MINVAD+σVAD	43.69	41.60	53.36	*59.73*

Table 18 Classification Accuracy of feature combinations for Lexicon L3 and L1

Classifiers	**Gradient Boost**		**Random Forest**	
	D1	***D2***	***D1***	***D2***
Dataset				
MINVAD + MAXVAD+Φ^E+WC^E	52.17	55.28	52.17	55.28
MINVAD+μVAD+CF^F+WC^E	52.79	56.52	52.79	56.52
T MIN$_{VAD}$+MAX$_{VAD}$+σVAD+Φ^E+WC^E	***57.14***	54.04	***57.14***	54.04
I T MINVAD+ MAXVAD+μVAD+WC^E+CF^F	54.65	55.28	54.65	55.28
MINVAD+MAXVAD+μVAD+CF^F+Φ^E	***57.14***	54.04	***57.14***	54.04
MINVADMAXVAD+μVAD+CF^F+Φ^E	51.55	***57.14***	51.55	57.14
MIN$_{VAD}$+MAX$_{VAD}$+μ$_{VAD}$+CF^F+σVAD+Φ^E	54.65	54.66	54.65	54.66
MINVAD + MAXVAD+WC^E+Φ^E+CF^F T I	56.52	54.05	56.21	54.04
μVAD+MINVAD+WC^E+Φ^E+CF^F T I	55.15	55.90	51.5	55.90
σVAD+MIN$_{VAD}$+μ$_{VAD}$+WC^E+Φ^E+CF^F T I	55.27	54.66	55.27	54.66
μVAD+MINVAD+WC^E+Φ^E+CF^F T I	52.17	56.52	52.17	56.52
σVAD+MIN$_{VAD}$+μ$_{VAD}$+WC^E+Φ^E+CF^F+MAX$_{VAD}$ T I	56.52	54.66	56.51	54.66

The classification results for different combinations of features derived from lexicon L1 is summarized in Table 17. For dataset D1 combination of min and mean of VAD with a Random Forest resulted in the highest accuracy of 57.71%. For Dataset D2 combination of mean, min, and,-Standard deviation of VAD and Random Forest classifier is the best result of 59.73%.

The classification results for different combinations of features derived from lexicon L1 and L3 is summarized in Table 18. For dataset D1, the combination of (min, max, mean of VAD), Average Emotion Intensity Score and Word Count per Emotion with a Threshold feature using Gradient Boost and Random Forest resulted in highest accuracy of 57.14%. For and Random Forest resulted in highest accuracy of 57.14%. For D2 the combination of (min, mean) of VAD, Word Count per Emotion with a Threshold Score, Normalized Word Count per Emotion and (Gradient Boost and Random Forest) is best result of 57.14%.

Each Lexicon feature extracted is capable of giving classification accuracy above 40%. This is remarkable, as a single feature. For the Lexicons Norms of valence, arousal, and dominance and NRC Valence Arousal and Dominance the Mean and Standard deviation features are more discriminative. The choice of classifier depends on the data set. The accuracy obtained for D3, the combined dataset, is relatively low. The combination of Normalized Word Count per Emotion and Average Emotion Intensity Score is the best for NRC Affect Intensity Lexicon.

Further the combination of the lexicon features considerably improves emotion recognition accuracy.

he study carried out is to analyze the impact of lexicon-based features on the classification accuracy of emotion in music lyric. The choice of feature and classifier depends on the Dataset and the Lexicon. The classification performance with multiple Lexicon features is to be investigated. The work is to be extended by incorporating standard NLP features as well as Lyric oriented features for the design of a more robust classification model.

5. Conclusion

The work presented is an extensive experimental study on the impact of Lexicon features on the classification of Music Lyrics based on Emotion. Five Lexicons and Two annotated datasets are employed for the experiments. A third dataset is created by combining the two. Two of the Lexicons are Valence-Arousal Dominance based. The other three incorporate the intensity of discrete emotions. Lexicon features are defined depending on its type. Three different classifiers are used for the study. The results highlight the effectiveness of the defined features in identifying the emotions. The result would help in designing more robust Music Lyric classifiers by combining multiple Lexicon features, Natural Language Features and Lyric specific features.

References

[1] R. Raieli ,The current status of MIR systems, Editor in Chandos Information Professional Series, *Multimedia InformationRetrieval*,2013, Pages 175-193, ISBN

[2] A. R. Murthy, K. M. A. Kumar,; A Review of Different Approaches for Detecting Emotion from Text, *IOPConf.Ser.:Mater.Sci.Eng.***1110**012009,2021

[3] S. Buechel, S. Rucker & U. Hahn, "Learning and evaluating emotion lexicons for 91 languages" , Proceedings of the58th AnnualMeetingof42theAssociation forComputationalLinguistics,2020 (pp.1202–1217)

[4] R. Malheiro, R. Panda, P. Gomes, R. P. Paiva , "Emotionly-Relevant Features for Classification and Regression of Music Lyrics" , *IEEE Transactions on Affective Computing*, January 2018

[5] C.Strapparava, A.Valitutti,"Wordnet-Affect:A naffective extension fof WordNet" , *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004 , pages 1083–1086, Lisbon,Portugal

[6] A. Esuli. F. Sebastiani , "Senti wordnet: A publicly available lexical resource for opinion mining" , *Proceedings of the 5th Conference on Language Resources and Evaluation (LRECÕ06*)

[7] C. J. Hutto & E. Gilbert, "VADER: A Parsimonious Rule based Model for Sentiment Analysis of social media Text", *Eighth InternationalConferenceonWeblogsandsocialmedia*(ICWSM-14).AnnArbor MI ,June 2014

[8] P. J. Lang, M .Bradley," Affective norms for English words(ANEW): Instruction manual and affective ratings", Technical report,1999

[9] S. M. Mohammad, "Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words" , *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018 ,(Volume 1: Long Papers), pages 174–184,Melbourne, Australia. Association for Computational Linguistics

[10] S. M. Mohammad," Word Affect Intensities" , *Proceedings of the 11thEdition of the Language Resources and Evaluation Conference* (LREC-2018), May 2018, Miyazaki, Japan

[11] G. Badaro, H .Jundi, H. Hajj. and W. El-Hajj, "EmoWordNet: Automatic Expansion of Emotion Lexicon Using English WordNet" ,*Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, June 2018, New Orleans, Louisiana

[12] K. Hevner (1935), "The affective character of the major and minor modes in music ", *American Journal of Psychology*, 41, 103–118

[13] D. Watson, L. A. Clark, A. Tellegen, "Development and validation of brief measures of positive and negative affect: the PANAS scales *",* *JPersSocPsychol.*1988June,54(6):1063-70.Doi:10.1037//0022-3514.54.6.1063.PMID:3397865

[14] R. Plutchik,,"A general psych evolutionary theory of emotion" ,Emotion: Theory ,research ,and experience,1980,1(3),3–33

[15] J. A. Russell, "Core affect and the psychological construction of emotion", *Psychological review*, 2003, 110(1):145.

[16] A. Mehrabian, "Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament" , *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[17] J. Thayer & R .D. Lane ,"A model of neurovisceral integration in emotion regulation and dysregulation " , *Journal of affective disorders*,2000, 61 3, 201-16

[18] A. Bandhakavi, N. Wiratunga, D. Padmanabhan, S.

Massie, “Lexicon based feature extraction for emotion text classification” , *Pattern Recognition Letters,* Volume 93, 2017, Pages 133-142,ISSN

[19] Z Teng, D . Vo, Y. Zhang , ”Context-Sensitive Lexicon Features for Neural Sentiment Analysis” , *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*,2016,pages1629–1638, Austin, Texas. Association for Computational Linguistics.

[20] L. D. Bruyne, P. Atanasova, I. Augenstein , “Joint Emotion Label Space Modelling for Affect Lexica” , *Computer Speech & Language*,2022, Volume 71, 2022, 101257,ISSN 0885-2308

[21] G. Tao , M. Ruotian, Z. Qi. Z. Lujun, J. Yu-Gang Jiang, H. Xuanjing(2019), “CNN-Based Chinese NER with Lexicon Rethinking” ,*Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, IJCAI 2019, Macao, China, August 10-16, 2019.

,4982-4988.10.24963/ijcai.2019/692.

[22] C. Changfeng ,Q. Li, “A Multimodal Music Emotion Classification Method Based on Multifeature Combined Network Classifier” ,*Mathematical Problems in Engineering*, vol. 2020, Article ID 4606027,11 pages, 2020. https://doi.org/10.1155/2020/4606027

[23] J. S. G. Cañón, .,E. Cano., H. Boyer,., & H. G. Gutiérrez,.(2020). Joyful for you and tender for us: The influence of individual characteristics and language on emotion labeling and classification .In *Cumming J, Ha Lee J, McFee B, Schedl M , Devaney J, McKay C, Zagerle E, de Reuse T, editors. Proceedings of the 21st International Society for Music Information Retrieval Conference; 2020 Oct 11-16;Montréal, Canada.[Canada]: ISMIR; 2020.*. International Society for Music Information Retrieval (ISMIR).

[24] Y. Agrawal , R V Shanker ,V Alluri, Transformer-based approach towards music emotion recognition from lyrics. In *European Conference on Information Retrieva*l (pp. 167-175). Springer, Cham.(2021, March)

[25] R. Delbouys , R Hennequin, F Piccoli F, J Royo-Letelier , M Moussallam, Music Mood Detection Based on Audio and Lyrics with Deep Neural Net. arXiv 2018, arXiv:1809.07276

[26] J. .Abdillah ,I.Asror, Y. F. A .Wibowo . ;Emotion Classification of Song Lyrics Using Bidirectional LSTM Method with GloVe Word Representation Weighting. *J. RESTI* 2020, 4, 723–72

[27] E. Cano, M. Maurizio ,”Moody Lyrics: A Sentiment Annotated Lyrics Dataset”, 2017 *International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, Hong Kong, March, 2017. pp.118-

[28] A. B Warrier , V. Kuperman and M. Brysbert .”Norms of valence, arousal and dominance for 13,915 English norms”. Behavior Research Methods , 2013, 45,1191-1207

[29] J. Louviere T Flynn , and A. Marley . “Best-Worst Scaling :Thery, Methods and Applications, *Cambridge: Cambridge University Press,*2015

[30] U. Krcadinac, P. Pasquier , J. Jovonovic and V. Devedzic,” Synesketch: An Open Source Library f JulySept.2013,doi:10.1109/T- *AFFC.*2013.18

[31] N. S. Chauhan Decision Tree Algorithm Explained, https://www.kdnuggets.com/2020/01/decision-treealgorithm-explained.html (accessed Feb 2023)

[32] A . Nagpal , Decision Tree Ensemble Bagging and Boasting Towards data science https://towardsdatascience.com/(accessed Feb 2023

[33] What is Random Forest Foresthttps://www.ibm.com/topics/random-forest(accessed Feb 2023),