Emotion Unmasked: A Transformer-based Analysis of Lyrics for Improved Emotion Recognition

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computational Linguistics by Research

by

R. Guru Ravi Shanker
2018114011
ramaguru.guru@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled 'Emotion Unmasked: A Transformer-based Analysis of Lyrics for Improved Emotion Recognition.' by R. Guru Ravi Shanker, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. Vinoo Alluri

To **family**, friends and

the almighty

Acknowledgments

I want to express my sincere thanks to Prof. Vinoo Alluri for her continuous support and guidance. She is more than a professor, like a mentor and easily approachable. Her continuous feedback and criticisms helped me improve. I also would like to thank Yudhik for mentoring me in various aspects during my initial days of my research. This work would not have been possible without their support and guidance.

It is my blessing to have such a supportive and understanding family. I cannote be thankful enough to them, especially my brother who guided me on every step of my life. Their continuous encouragement and freedom really helped me. I would also like to thank my grandparents and extended family members for wishing me well.

I am thankful to Shashank and Sai Avinash for making my 5th year exciting. I also like to thank Manikanta, Koushik, Sandeep, Tharun, Datta, Harish, Neeraj, Akash, Tushar, Sridhar, Snehith, Vinay, Umesh, Vishnu, Keshav for making awesome memories during my college days.

Special thanks to Jashn and Akshit for being my amazing roommates and well wishers during my research life.

Ella pugazhum iraivanukke

Abstract

Music is an important art form which has been present for centuries. Lyrics are a crucial part of a song conveying thoughts, messages and also a wide range of emotions. Individuals listen to songs to satisfy their emotional needs. The task of identifying emotions from a given music track has been an active pursuit in the Music Information Retrieval (MIR) community for years. Music emotion recognition has typically relied on acoustic features, social tags, and other metadata to identify and classify music emotions. The role of lyrics in music emotion recognition remains under-appreciated in spite of several studies reporting superior performance of music emotion classifiers based on features extracted from lyrics. In the first study, we use the transformer-based approach model using XLNet as the base architecture, which has not been used to identify emotional connotations of music based on lyrics. Our proposed approach outperforms existing methods for multiple datasets. We also used a robust methodology to enhance web crawlers' accuracy for extracting lyrics.

There are no datasets of Indian language songs that contain both valence and arousal manual ratings of lyrics. We present a new manually annotated dataset of Telugu songs' lyrics collected from Spotify with valence and arousal annotated on a discrete scale. A fairly high inter-annotator agreement was observed for both valence and arousal. Subsequently, we create two music emotion recognition models by using two classification techniques to identify valence, arousal and respective emotion quadrant from lyrics. Support vector machine (SVM) with term frequency-inverse document frequency (TF-IDF) features and fine-tuning the pre-trained XLMRoBERTa (XLM-R) model were used for valence, arousal and quadrant classification tasks. Fine-tuned XLMRoBERTa performs better than the SVM by improving macro-averaged F1-scores of 54.69%, 67.61%, 34.13% to 77.90%, 80.71% and 58.33% for valence, arousal and quadrant classifications, respectively, on 10-fold cross-validation.

In addition, we compare our lyrics annotations with Spotify's annotations of valence and energy (same as arousal), which are based on entire music tracks. We also compare the performance emotion recognition models on the original, translated, transliterated texts. The implications of our findings are discussed. We make the dataset publicly available with lyrics, annotations and Spotify IDs. We also used the XLM-R model to identify emotional connotations in Indian language song lyrics datasets of Hindi and Telugu. We also conducted a

perceptual validation study on misclassified lyrics. Lastly, we conduct a study to understand the individual differences between the preferences of music with and without lyrics.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Music plays an important role in our daily lives. It helps in regulating our moods, and individuals prefer songs which satisfy their emotional needs. Music Information Retrieval (MIR) is an interdisciplinary field that aims to retrieve information from music which helps in understanding the musical tastes and preferences of people. It has applications like music classification, recommendation systems, generation, and other related tasks. Automatic music emotion recognition is an active music classification task in the field of MIR. It is typically done using audio signals, social tags and song metadata.

Lyrics which usually contribute to musical enjoyment are largely neglected when compared to audio, social tags and metadata in MIR. They play a crucial role in eliciting emotions. Lyrics evoke strong emotions, especially negative ones. They also contribute to the musical reward for choosing songs. Lyrics are one of the major components of songs which attracts listeners. It can be used to convey complex ideas and messages. Though lyrics play an important role in music listening and creating, there have not been many studies focusing on lyrics as compared to audio in music emotion recognition tasks. Hence, leveraging the rich information lyrics can help improve MIR tasks.

1.2 Research Gap

Natural language processing (NLP) deals with processing and extracting relevant information from text. NLP techniques have been less used in lyrics text as compared to prose and other short texts like reviews and tweets. Despite comprising short, concise and obscure phrases, lyrics have been used as a tool to convey strong messages, thoughts and emotions. NLP in lyrics has been limited to using word-level features like Bag-of-words and TF-IDF features for the emotion recognition task. Recently, recurrent neural network-based approaches have been used in emotion recognition. The significant advancements in the field of NLP have led to

novel and improved ways to extract information from text. This can be used to bridge the gap between the task of music emotion recognition using lyrics and NLP. The advanced transformer models can help extract relevant lyrical features.

Lyrics emotion recognition in Indian languages is also not explored much as compared to other high-resource languages. There are very few publicly available datasets for lyrics emotion recognition in Indian languages. NLP on Indian language lyrics were also limited to the use of only word-level features. Transformer models were not used in processing Indian language lyrics.

Spotify provides emotion values for each musical track in its database, but it doesn't provide a method to calculate the values. These emotion values are popularly used in many studies. We need to check if these values are consistent with human perception of emotion or are biased towards a certain group of songs.

There has been a staggering increase in the consumption of music on online music streaming platforms. People listen to different types and genres of music. Lyrics may be an important factor for people to choose songs, but for some, it may only depend on the music. We need to understand whether certain personality traits affect our choice of music.

1.3 Research Objectives

Through this thesis, we aim

- To validate the robustness of the large context-aware transformer models in English, Hindi and Telugu language song lyrics on the task of lyrics emotion recognition.
- To create and publicly release a new manually valence-arousal annotated lyrics dataset, the first of its kind in the Telugu language.
- To check any differences in human perceived emotions and Spotify-retrieved emotion values?
- Are there any particular personality traits which prefer listening to music without lyrics and music with lyrics?

1.4 Key Contributions

The key contributions of the thesis are:-

- Context-aware transformer models like XLNet, XLM-R perform better than context-free models on the task of lyrics emotion recognition in three languages of English, Hindi and Telugu.

- Publicly released a new manually VA annotated dataset, the first of its kind in the Telugu language.
- Compared human perceived emotions with Spotify-retrieved emotion values and found associations between perceived arousal and Spotify arousal values.
- Conducted a perceptual validation study on misclassified lyrics and found about 50% of the misclassified lyrics were agreed upon by the new annotations. This implies the high subjectivity of emotion perception and the need for at least 15-20 people to annotate an excerpt.
- Found associations between individuals scoring with dominant Openness trait and their preferences for music without lyrics. Also, found association between individuals scoring with dominant Neuroticism trait showed affinity towards music with lyrics

1.5 Thesis Roadmap

The thesis is constructed in the following way

- Chapter 2 deals with the background of sentiment analysis and emotion recognition.
- Chapter 3 deals with improving existing lyric emotion recognition models using transformer models for English song lyrics.
- Chapter 4 deals with the creation of a new Tollywood lyrics dataset with valence and arousal annotations, the first of its kind in Telugu. The dataset also contains Spotify-retrieved emotion values, which are compared against human annotations and results are discussed. We also validate the use of the transformer model on Telugu and Hindi language song lyrics.
- Chapter 5 deals with a study to see individual differences in the preferences for music with lyrics and music without lyrics.
- Chapter 6 concludes with summary, limitations and future prospects of the studies.

Chapter 2

Background

2.1 Sentiment Analysis

Sentiment analysis deals with identifying affective connotations of text. It is a way to extract subjective opinions and attitudes from language. Before the availability of large text collection and online resources, sentiment analysis was based only on surveys and public opinions [32]. Historically, sentiment analysis was done on written paper documents to ascertain the attitude of the writer. Nowadays, it is done on social media posts [16], product reviews [25] on various e-commerce websites, news articles [59] and other subjective texts on the internet to determine the opinion polarity towards a product or any person.

It has gained widespread recognition in recent years, not just among researchers but also among various organizations [64], to determine the sentiment of the users towards their organization, entity or product. They also help in improving recommendation systems by providing sentiment features. It can also be used to flag certain types of posts on social media, market a product and other applications. For example, sentiment analysis was used by the Obama government to understand the sentiment of the people towards the government's policy announcements.

There are different levels of sentiment analysis. It is broadly done on a sentence, document, and aspect or feature level [64]. In sentence-level sentiment analysis, as the name suggests, analysis is done on individual subjective sentences that express a particular opinion without considering previous or next sentences. The document-level analysis involves determining the overall opinion of the document. The dominant sentiment across the whole document is used for analyzing the documents. Aspect-based sentiment analysis involves characterizing the sentiment of a particular target.

Document-level sentiment analysis is the most popular method of sentiment analysis, for example, on product reviews, news articles, tweets, lyrics and similar others. Typically, humans annotate a corpus, following which a model is trained to classify new samples [48, 9].

2.1.1 Different methods in Sentiment Analysis

The raw text doesn't provide any numeric information to the machine. The text has to be represented in a meaningful numeric way for machines to interpret. There are representations which take the order of the words into consideration (contextual representations) and which don't take the order of the words into consideration (context-free).

2.1.2 Context Free

2.1.2.1 Lexicon

There are various affect lexicons like Affective Norms for English Words (ANEW) [11], Sentiwordnet [23] and similar others which can be used for sentiment analysis. For each word in the lexicon they, have the corresponding manual annotated rating of sentiment or affect to it. These types of lexicons have a limited number of words and do not cover all the words. One way to deal with out-of-lexicon words is to use synonym information from lexicons like wordnet. We can also bring all the words in the text to their base form using stemming or lemmatization to improve the presence of words in the lexicon.

2.1.2.2 Bag of words

Bag of words [67] is one of the simplest text-to-vector techniques, which describes the frequency of a set of words in the document. It is similar to one hot encoding for different categories. For example, for the vocabulary set {a, the, car, girl, boy, driving, swimming}, "A boy is driving the car" is encoded as {1, 1, 1, 0, 1, 1, 0}. If two documents have similar encoding, it means that they have similar words in their documents. The drawbacks of these representations are that there may be frequently occurring common words which do not provide any meaning to the representation. Also, it doesn't take into account the syntactic structure of the sentence. The syntactic structure of the sentence is also not preserved. For example, the sentences "Take pity, not action" and "Take action, not pity" have different meanings but would have a similar bag of word representations.

2.1.2.3 Term Frequency-Inverse Document Frequency (TF-IDF)

The TF-IDF [53] is an improved statistical method that calculates the relevance of a word in a document from a set of documents. It penalizes the frequently occurring words in all documents, like articles and prepositions. It depends on two values, one term frequency (TF), which represents the frequency of the word in the particular document, and the other inverse document frequency (IDF), which represents the count of unique documents the word is present.

$$TF = \frac{(number\ of\ times\ term\ appears\ in\ document)}{(total\ number\ of\ terms\ in\ document)} \tag{2.1}$$

$$IDF = \log\left(\frac{(total\ number\ of\ documents)}{(number\ of\ documents\ containing\ the\ term)}\right) \tag{2.2}$$

$$TF - IDF = TF * IDF \tag{2.3}$$

This method also doesn't take into account the syntax of the sentence.

2.1.3 Contextual models

2.1.3.1 N-grams

N-grams are a continuous sequence of n-words. An n-gram can be of any length. For example, the 2 grams (bigrams) in the sentence "Let us make peace." is "Let us", and "us make" and "make peace". For each sentence of k words, there would be k - n + 1 n-grams. These n-grams can be used as features to represent a piece of text. N-grams take into account the previous words (other than unigram) while representing the text hence preserving the context. These features are popularly used in sentiment analysis and help in identifying negations. [27, 6]

2.1.3.2 Word Embeddings

Word embeddings are a type of word representation in a high-dimensional continuous vector space. The embeddings capture the context by taking information from the neighbouring words. There are different types of word embeddings, including word2vec [43], GloVe [51], FastText [10] and similar others. There are pre-trained models which are trained on large volumes of text. These pre-trained models are used for representing the text or sentences for downstream tasks like sentiment analysis.

2.1.3.3 Neural Network Based Representation

A recurrent neural network (RNN) is a sequential neural network which takes in sequential input, like words of a sentence. Prior information from the previous words is passed sequentially from one cell to another to the present word. Bi-directional RNNs are used to capture context from both sides. Popular RNNs are Long Short Term Memory(LSTM) [28] and Gated Recurrent Units(GRUs).

2.1.3.4 Transformer Models

The recent advancements in NLP have to do with the emergence of transformer models. These models use self-attention to represent the text. Self-attention is an attention mechanism which helps to involve and gather information from other parts of the text sentence. They do not involve recurrent or convolution layers. They are faster to train and are easily parallelizable. They help capture longer dependencies in the text compared to LSTMs. The encoder blocks in

the transformer models are used to encode the text for representation to perform downstream NLP tasks. Popular encoder transformer models are BERT [20], XLNet [66], RoBERTA [38] and many others.

2.1.4 Classification Models

Supervised classification techniques are used on top of the vector representations of the text to perform downstream tasks such as sentiment analysis. Popular classification techniques include Support Vector Machine and Naive Bayes.

2.1.4.1 Naive Bayes

Naive Bayes works on the principle of the Bayes theorem. It is a probabilistic model which calculates the probability of each output class given a set of input features. The algorithm assumes that all input features are independent.

$$P(c|x) = \frac{P(x|c) * P(c)}{P(x)} \tag{2.4}$$

where c is the class, and x is the input features.

2.1.4.2 Support Vector Machine

Support Vector Machine (SVM) [18] transforms the data to a higher dimensional plane. It divides two or more classes by finding optimal hyperplanes in the higher dimensional plane (Refer to figure 2.1). These hyperplanes divide the classes by maximizing the distance between the data points of the class and the hyperplane. The data points which are closest to the hyperplanes are called support vectors. It works for both linearly and non-linearly separable data.

2.1.5 Sentiment Analysis on different texts

Sentiment analysis has been done on a wide variety of texts. Sentiment analysis has been heavily done on shorter texts like reviews [11], tweets [16], sentences and others. Such short sentences do not require high contextual features for performing the analysis. Longer texts like prose, poem, articles and lyrics require longer contextual dependencies for representing the text as a whole. Creative texts like poems and lyrics require powerful contextual models for performing sentiment analysis. Even though song lyric messages are often heartfelt, exciting, and brief, lyrics' emotion recognition is difficult. So, it requires advanced contextual transformer models.

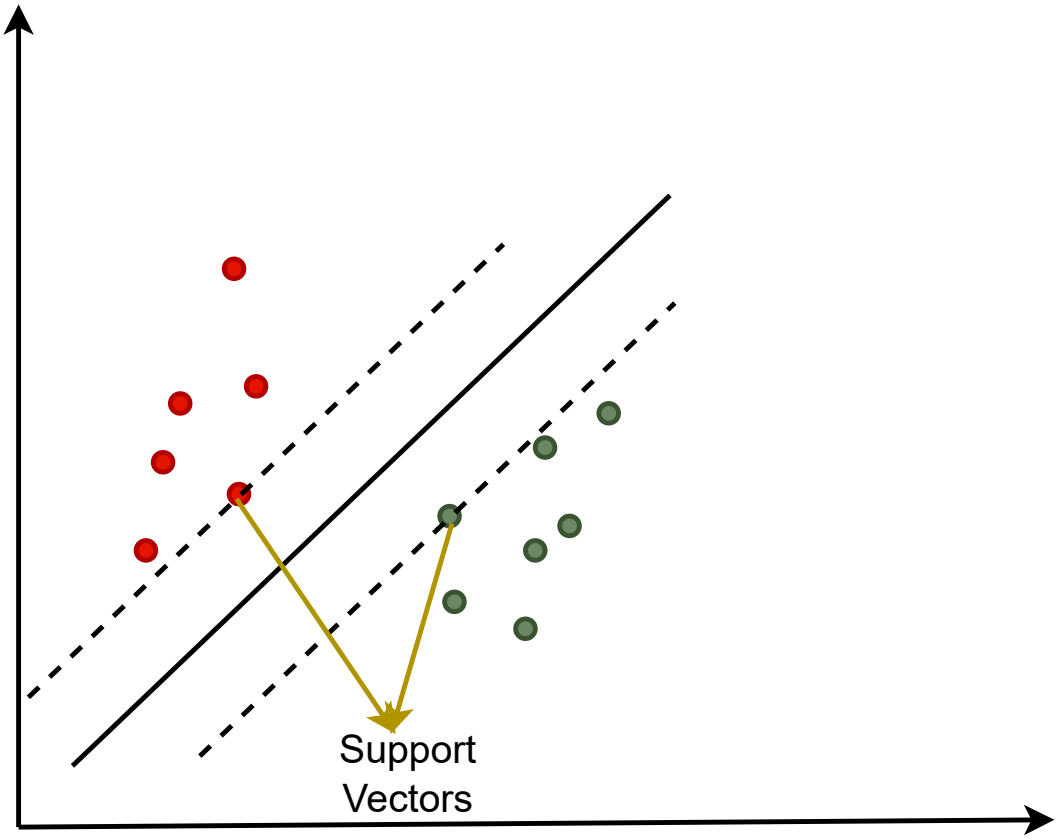

Figure 2.1 Support Vector Machine Model

2.1.6 Sentiment Analysis in Indian languages

India is a multilingual country with 22 official languages. Similar to the English language, there has been a lot of work done in the sentiment analysis of Indian languages. Sentiment analysis in Indian languages has been done using simple lexicon-based features [47] , TF-IDF features [9], word vector features [44] and also transformer-based features [63]. It has also been done on various types of text like tweets [54], reviews [25], news articles [59], poems and similar other texts. There are also many Indian languages like Konkani, Gujarati, Oriya and others which have not been explored, like Hindi, for sentiment analysis tasks.

2.2 Emotion Recognition

Emotion is a complex behavioural phenomenon involving many levels. Sentiment analysis deals with identifying the polarity, whereas emotion recognition identifies the specific emotions expressed by the excerpt.

Music helps us regulate our emotions and is one of the main reasons people listen to and interact with music. For example, a few songs help us to focus and be motivated. Music gives us a sense of identity and belongingness. It also helps us get amazing transportive experiences. Most of the previous works involve using traditional methods using audio, social tags and metadata. The first study on music and emotions was done by Gordon Burner in 1990 by correlating music and mood [13].

Audio-based music emotion recognition involves using signal processing techniques for extracting energy, rhythm, timbre, melody, tempo and other structural features [37, 15]. The patterns in acoustic cues help differentiate different emotion perceptions. [34] involves the use of social tags for the classification of music moods.

2.2.1 Emotion Taxonomy

It is also necessary to differentiate between perceived and induced emotions. Perceived emotions are the ones which can be perceived based on song cues, and induced emotions are the ones which are felt because of personal experiences. For example, the song "Happy Birthday" can be perceived as happy based on its musical cues, but someone might feel sad because the song reminds them of a close friend who is no more. Induced emotions are personalized and depend on personal experiences. Most of the studies deal with the prediction of perceived emotions since they are less subjective and not personalized.

There have been many cognitive models to account for the emotional experience. The datasets for emotion recognition follow one of the many available taxonomies. It is selected based on music theory and cognition studies for emotion modelling. The different types of taxonomies are broadly categorized into two types, viz. categorical and dimensional models.

2.2.2 Categorical Model

The model conceptualizes emotions into a predefined set of distinct classes. Paul Ekman proposed a basic emotion theory [21] according to which there are six basic emotions recognized by facial expressions. The six basic emotions, viz. happiness, sadness, fear, anger, disgust, and surprise, are distinct and universal. Plutchik's Wheel [52] of Emotions proposes eight primary emotions, viz. joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. All other emotions are a combination of primary emotions. This is the simplest way of categorizing emotions but is limited to a predefined fixed number of categories which results in poor resolution of human emotions. The number of emotions are too small when compared to richness in human emotion perception. Complex emotions cannot be represented as basic emotions, for example, envy, nostalgia and similar others. Cultural and individual differences in perceiving lead to ambiguity in emotions.

2.2.3 Dimensional Model

The model deals with mapping a music excerpt in a continuous emotion space rather than a finite set of mood categories. James Russel proposed the circumplex model [57] in the year 1980. According to the model, all human emotions can be distributed in a two-dimensional circular space with axes of valence and arousal. Valence represents pleasantness or unpleasantness, whereas arousal represents energy. These pairs of values can be used to represent any emotion. Ortony, Clore and Collins proposed the OCC model [5] of emotion, stating that emotions can be classified into three main categories of pleasure, arousal, and dominance. This model also proposes that emotions can be further classified based on the degree of pleasure or arousal being experienced.

The dimensional approach helps to capture complex emotions better than the categorical approach. It also has the ability to represent the complex gamut of emotions in a 2-D plane See figure2.2). The emotions which are inversely correlated are placed at the opposite ends of the circumplex model.

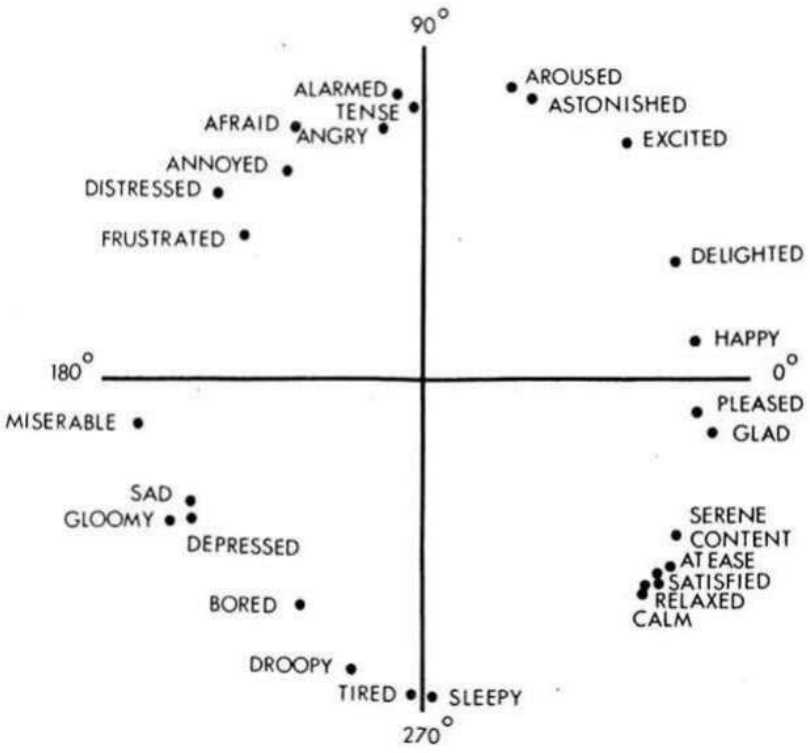

Figure 2.2 28 emotion words in Russel's 2-D plane [57]

Chapter 3

Lyrics Emotion Recognition in English

3.1 Introduction

In this chapter, we will see how context-aware transformer models perform well in the task of lyrics emotion recognition. We also employ a multi-tasking approach to learn the emotional quadrant, valence, and arousal. We also use a robust lyrics extraction method to upgrade the scraping of English lyrics.

3.2 Related Work

NLP has been effectively used in the field of MIR for topic modeling [31], identifying song structure via lyrics [24], and also mood classification [29]. But NLP approaches have been used limited to word-level features and embeddings [29, 30], and have not utilized long-term dependencies such as transformers [20, 66]. The MoodyLyrics dataset [14] was used by [2], who reported an impressive $\mathcal{F}1$-score of 91.00% using RNNs. RNN models like LSTMs operate on Markov's principle, where past information is used to predict a future state. Previous studies have also reported that emotion classifiers based on features extracted from lyrics perform better than those based on audio [29, 65]. Lyrics can be seen as narratives, which makes transformers a natural choice for mining affective connotations. In this study, we employ a multi-task setup, using XLNet [66] as the base architecture for the classification of emotions and evaluate the performance of our model on several datasets that have been organized based on lyrics' emotional connotations.

3.3 Lyrics Extraction

Most of the publicly available datasets do not provide lyrics due to copyright issues. They provide URLs from different websites, which become obsolete over time. However, the datasets

Table 3.1 Examples of Track name or Artist name mistakes in datasets

Track name	Artist name	Mistake
Bridge over Trobled Water	Simon and Garfunel	Spelling mistake in trackname. "Trobled" → "Troubled"
Penny lane	Beatles	Artist name is "The Beatles" as per the website
What a feeling	Irene Clara	Track's name is Flashdance... What a feeling, as per website

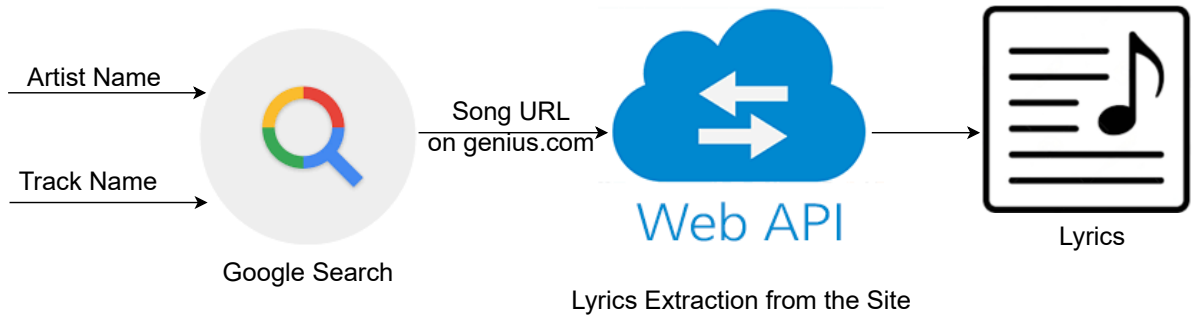

Figure 3.1 Lyrics Extraction Pipeline

provide the artist and track names to identify the song for which the lyrics have been annotated. All the existing APIs require the exact artist and track name for extracting the lyrics, which are often misspelt in the datasets. Refer to table 3.1 for examples. We develop a robust lyrics extraction method, which fetches the correct Genius [1] URL of the song using a google search of song and artist name. The lyrics are then scraped from the website using the URL. Refer to figure 3.1.

3.4 Datasets

3.4.1 MoodyLyrics

MoodyLyrics [14] dataset comprises of 2595 songs and their categorized emotions viz happy(Q1), anger(Q2), sad(Q3) and relaxed(Q4). The authors made a mixed lexicon by combining the existing lexicons of ANEW(Affective Norm of English Words), WordNet and WordNet-Affect to get the valence and arousal values for each word present. Then they calculated aggregated scores of valence and arousal for each song's lyrics and normalized between the range [-1, 1]. They distinctly placed the songs in each of the four categories by taking a threshold of valence and arousal above or below which(depending on the quadrant) the song is categorized to a

[1] https://genius.com/

particular emotion as per the quadrant in the Valence Arousal plane. We extracted 2576 song lyrics from the dataset using our lyric extraction method.

3.4.2 MER

This dataset [41] comprises 180 songs along with their valence arousal values. This is a human-annotated dataset done by 39 people obtaining an average of 8 annotations on a lyric. The annotators assigned a value in both valence and arousal scales based on the lyrics of the song on a 9-point scale, i.e. from -4 to 4. The values obtained from the average rating of valence and arousal per song were calculated. We were able to extract 177 songs using our lyric extraction method and manually added the remaining three songs to complete the lyric extraction for the 180-song dataset. Each song in the dataset was annotated by multiple people. Annotators were made to read the lyrics and identify the predominant emotion of the song. Multiple annotators manually assigned a value for the valence and arousal of a song based on the lyrics rather than using available lexicons for the values. The annotators assigned valence and arousal values for a song based on lyrics.

3.4.3 AllMusic

This dataset [41] comprises 771 songs along with the quadrant it belongs in the valence arousal plane. User mood tags from the All Music website were scraped from the website which was used in classification. Depending on the valence arousal values of the words obtained from ANEW, they identified 33, 29, 12 and 18 words in Q1, Q2, Q3 and Q4, respectively. They categorized a song into a specific quadrant if all the mood tags of a song lie in the same quadrant. To validate their classification, they made three annotators read the lyrics for each song, and if any one of the three annotators agreed with the classification, then they proposed that the song had been correctly classified. We were able to obtain 762 song lyrics from this dataset.

3.5 Architecture

We propose a deep neural network, when given lyrics, outputs the emotion quadrant along with the valence and arousal quadrants. Multi-task learning of valence, arousal and quadrant classification helps in faster convergence and reduces overfitting. We use XLNet [66] as our base network. It is a large bi-directional transformer with improved training methodology, training on larger data and more computational power. XLNet improves upon BERT [20] with added recurrence and its ability to learn longer sequences.

Figure 3.2 shows the overview of the proposed architecture. The XLNet transformer takes in lyrics and outputs hidden raw states, which are then passed on to the Sequence Summary

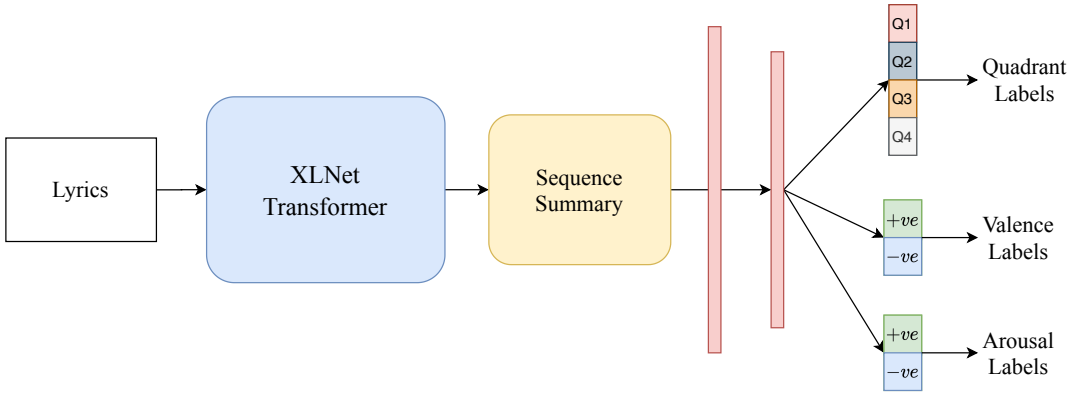

Figure 3.2 Model Overview

block, which computes a single vector summary of a sequence of hidden states. Further, it is passed through a fully connected layer which encodes the information to a vector of length 8. This layer does the complementary tasks of Quadrant, Valence and Arousal classification separately. The pre-trained XLNet model and the classification layers are trained for all three tasks.

We use the following loss function

$$L = (\lambda_1 * L_Q) + (\lambda_2 * L_V) + (\lambda_3 * L_A) \tag{3.1}$$

where L_Q,L_V, and L_A represents the classification loss on Quadrants, Valence, and Arousal, respectively.

Implementation Details

For implementation, we use the rich pre-trained (xlnet-base-cased) model. We fine-tune the large model with AdamW optimizer [39] with an initial learning rate of $2e^{-5}$. We use Cross-Entropy Loss for calculating loss. We use the max length of lyrics as 1024 and batch size 8. We also train single-task models to compare the performance.

3.6 Experiments

We report the average accuracy (also known as the micro-averaged F1-score), macro-averaged precision, recall and F1-score (F1). The macro-averaged F1-score ($\mathcal{F}_1$) is given by the formula 3.2.

$$F1_x = 2\frac{P_x R_x}{P_x + R_x}; \qquad \mathcal{F}_1 = \frac{1}{n}\sum_x F1_x \tag{3.2}$$

P_x, R_x and $F1_x$ are the standard precision, recall and $\mathcal{F}_1$-score of a particular class x, and n is the total number of classes.

Table 3.2 Classification by Quadrants; Trained on MER and validated on AllMusic dataset using our baseline network.

Approach	**Accuracy**	**Precision**	**Recall**	$\mathcal{F}_1$**-score**
Traditional NLP based technique [2]	-	-	-	73.60%
Lexicon based technique [3]	74.25%	-	-	-
Our Baseline	**76.31%**	**75.86%**	**75.21%**	**75.40%**

3.7 Results

We use the above-mentioned MoodyLyrics, MER and AllMusic datasets to compare our model's performance. For evaluating the performance, we use the same train test splits as done by previous researchers on the datasets to have fair comparisons. In the table, the baseline model refers to our multi-task model. Table 3.2 compares the performance of our baseline model trained on MER and validated on the AllMusic dataset. Table 3.3 shows the result of training and validating our baseline model on the MoodyLyrics dataset for the Quadrant classification task. Similarly, table 3.4 shows the result of training and validating our baseline model on the MER dataset.

Ablation Study:

We choose the Moodylyrics dataset to do an ablation study due to its large size and quadrant representativeness. We also compare our multi-task model with the single-task model. See table 3.5. Single-task models are trained on a single task of either Quadrant, Valence or Arousal classification. The multi-task model performance is comparable to our single-task models, which take a long time and has to be done on single tasks. We also compared the performance of the XLNet transformer model with the BERT transformer model keeping other parameters the same and observed a better performance of XLNet by 1.3% of $\mathcal{F}_1$-score.

Table 3.3 Results of classification by Quadrants on MoodyLyrics dataset.

Approach	**Accuracy**	**Precision**	**Recall**	$\mathcal{F}_1$**-score**
Naive Bayes [1]	83.00%	87.00%	81.00%	82.00%
BiLSTM + Glove [1]	91.00%	92.00%	90.00%	91.00%
Our Method	**94.78**%	**94.77**%	**94.75**%	**94.77**%

Table 3.4 Results of classification on MER dataset.

Classification	**Approach**	**Accuracy**	**Precision**	**Recall**	$\mathcal{F}_1$**-score**
Quadrant	Traditional NLP-based technique [2]	-	-	-	80.10%
Quadrant	**Our Method**	**88.89%**	**90.83%**	**88.75%**	**88.60%**
Valence	Traditional NLP-based technique [2]	-	-	-	90.00%
Valence	**Our Method**	**94.44%**	**92.86%**	**95.83%**	**93.98%**
Arousal	Traditional NLP-based technique [2]	-	-	-	88.30%
Arousal	**Our Method**	**88.89%**	**90.00%**	**90.00%**	**88.89%**

Table 3.5 Ablation Study on MoodyLyrics

Classification	**Accuracy**		$\mathcal{F}_1$**-score**	
	Multi-Task	Single-Task	Multi-Task	Single-Task
Quadrant	94.78%	**95.68%**	94.77%	**95.60%**
Valence	95.73%	**96.51%**	95.67%	**96.46%**
Arousal	94.38%	**94.38%**	94.23%	**94.35%**

3.8 Conclusion

In this chapter, we have seen the robustness of transformer models over context-free models on multiple datasets for the task of lyrics emotion recognition in English. This study can improve applications such as playlist generation of music with similar emotions. Also, music recommendation systems can derive from incorporating emotional connotations of lyrics with acoustic content-based and collaborative filtering approaches.

Chapter 4

Lyrics Emotion Recognition in Indian Languages

India is a diverse country with many languages. Over 10,000 original songs are produced every year in India in different languages. There have been very few studies on Music Emotion Recognition in Indian languages compared to English. Bollywood and Tollywood are two of the prominent film industries in India. We aim to improve the Music Emotion Recognition systems in the Indian languages of Hindi and Telugu by leveraging the use of a multilingual transformer model.

4.1 Introduction

Sentiment analysis deals with identifying affective connotations of text. Most studies involve a categorical approach in which text is classified as either positive or negative in valence or arousal. There have been lyrics datasets using a dimensional approach containing both valence and arousal values in English [14, 41]. However, the very few that exist in Indian languages have limited annotations to either valence or arousal but not both.

For example, BolLy is a Hindi dataset annotated for valence [9], similar datasets exist for Manipuri [19], Bengali [44], Telugu [25] while another study has annotated Telugu lyrics for arousal [55]. Also, most lyrics datasets in Indian languages are not publicly available [19, 9, 49]. We need both valence and arousal dimensions to capture the entire gamut of emotions. Telugu is a morphologically rich language which makes automatic emotion identification difficult in contrast to languages poor in morphology, such as English. In our study, we create a manually annotated lyrics dataset in Telugu, which contains average valence and arousal perceptual ratings.

Further, to identify valence, arousal, and quadrant from lyrics, we employ two methods, Support Vector Machine (SVM) [18] with TF-IDF [56] features and fine-tuning pre-trained XLM-RoBERTa (XLM-R) [17] model for emotion recognition on the dataset.

Recently there have been studies [36, 35] that use Spotify features to find cultural differences. Since lyrics' emotions are mostly congruent with music emotions, we aim to verify congruence

between the lyrical perceptual emotion values and Spotify-retrieved emotion values. This has implications in using Spotify features for music emotion recognition, especially in the context of culturally diverse music. We release the Telugu lyrics dataset with average valence and arousal values along with Spotify IDs [1].

4.2 Related Work

There are annotated corpora for sentiment analysis in English [40] as well as in other Indian languages such as Telugu [25], Hindi [60] amongst others. Mostly the task has been done on short context texts like tweets [3], and reviews [25] compared to long context text like poems and lyrics. Both categorical and dimensional approaches for emotion recognition are widely accepted for lyrics. MER [41] dataset contains 180 English songs with manually annotated valence and arousal discrete values from -4 to 4. The existing datasets present in the Indian language's lyrics are limited to either valence or arousal categories. The valence annotated lyrics dataset is present in Hindi [9], albeit limited to two categories, that is, positive and negative sentiment. Similar datasets are present in Telugu [25], and Manipuri [19]. Arousal annotated lyrics dataset is present in Telugu [55]. Hence, there is a need to create datasets that have both valence and arousal. The task of emotion recognition has evolved from using simple lexicon-based methods [45] to using transformer models [4].

Numerous studies use NLP techniques like bag-of-words [22], TF-IDF features [9], word vectors [44], and models like convolutional neural network [44], recurrent neural networks (RNN) [2] for sentiment analysis. Recently, fine-tuned transformer XLNet model performed better than RNN-based techniques [4] for lyrics emotion recognition task because of its ability to capture longer contexts. Hence, we also compare the performance of context-free (SVM with TF-IDF) and context-based (fine-tuning pre-trained XLM-R) on the dataset.

Spotify features based on music tracks like valence, energy, danceability, and instrumentalness, amongst others, have been popularly used in many studies [61, 35] for various tasks such as analyses on music listening habits on streaming platforms and exploring differences in mood perception, respectively. However, several studies have evidenced culture-specific differences in emotion perception, and to music, [35, 58]. [35] compared human mood perception and Spotify-retrieved emotions, including valence and energy(or arousal), across English, Korean and Brazilian participants. They found the correlation between Spotify features of valence and arousal(VA) and human ratings ($r = 0.49$ and 0.63 respectively for VA) less than the mean correlation of human ratings ($r > 0.93$ for both VA).

Although valence and energy features from Spotify rely on the entire music track and not just lyrics alone, owing to the typical congruence between emotions conveyed by lyrics and musical features, one can expect a moderately high correlation between Spotify features and

[1] https://developer.spotify.com/documentation/web-api/

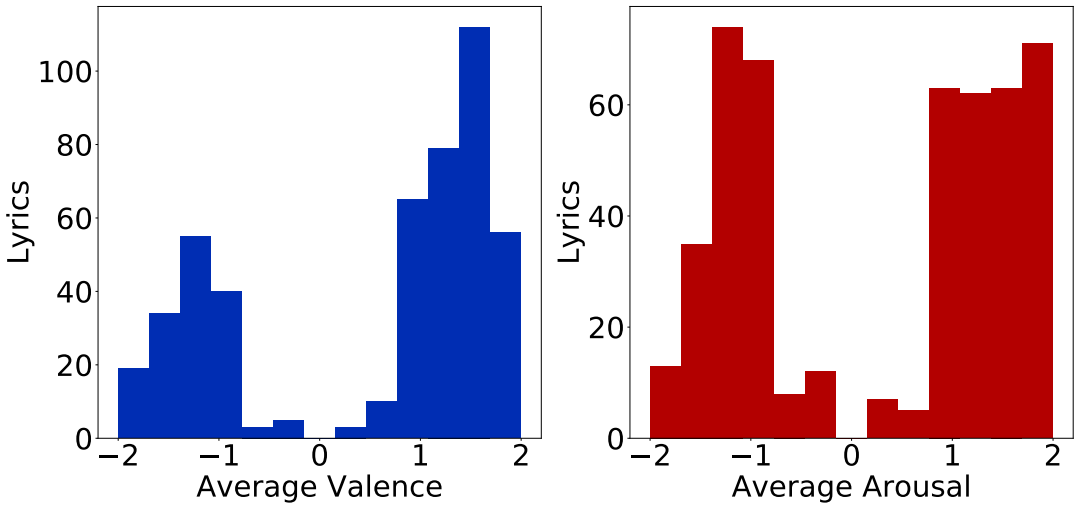

Figure 4.1 Histogram of Annotated Values

annotated features. Hence we additionally compare manual annotations of our dataset with Spotify's features.

4.3 The Dataset

4.3.1 Construction of the dataset

For the construction of the lyrics dataset, we chose Tollywood (Telugu film industry) songs from playlists that were available on Spotify. Telugu songs are generally positive, as noted by [55], and hence we initially observed around 70% of the songs annotated were positive on valence and high on arousal. To balance the number of songs in each quadrant, we chose songs from 9 diverse playlists, such as "Happy Vibes Telugu", "Sad Melodies Telugu", "Sleepy Telugu", and "Angry Telugu", amongst others.

Lyrics were scraped manually in their original script from Spotify, if available. For those not available, alternative websites such as LyricsTape[2], Lyrics Telugu[3] amongst others, were used for scraping the lyrics. Each track has an associated unique Spotify ID, based on which we removed 19 duplicates. Finally, a total of 481 song lyrics with their Spotify IDs were scraped. The dataset size is comparable to manually annotated lyrics datasets for both valence and arousal [41]. The Spotify features available for a track can be used to compare music features with lyrics. So we also provide Spotify ID for each song along with annotated valence and arousal of lyrics.

[2]https://www.lyricstape.com/
[3]https://lyricstelugu.in/

4.3.2 Annotation

We use Russell's circumplex model for annotating the lyrics based on valence and arousal. Three annotators, native Telugu speakers in the age group of 20-23, annotated the lyrics of each song for valence and arousal on a discrete five-point scale ranging from -2 to 2. The annotations were solely based on lyrics without listening to the song's audio.

To check the reliability and internal consistency of the dataset, we calculated Krippendorff's alpha [33] for ordinal data. As a result, acceptable alphas of 0.716 and 0.782 were obtained for valence and arousal, respectively, implying fair agreement among the annotators. Krippendorff's alpha (α) has been used in many studies [41] to assess inter-annotator agreement [33]. It can be used for all types of data, including ordinal and interval.

4.3.3 Dataset Release Information

The dataset has been made publicly available at `https://tinyurl.com/mu2zkjtc`. It contains average valence and arousal values for 481 Telugu songs, as shown in Figure 4.1. We also provide Spotify ID for each song which helps extract music-related features from Spotify for further analysis.

Out of the 481 songs, 325 lyrics were perceived as positive valence and 156 as negative valence. Similarly, 271 and 210 lyrics were positive and negative arousal lyrics, respectively. We achieved a final quadrant split, as shown in Figure 4.2. Quadrant-1 (Q1) consists of lyrics with positive valence and positive arousal, Quadrant-2 (Q2) with negative valence and positive arousal, Quadrant-3 (Q3) with negative valence and negative arousal and Quadrant-4 (Q4) with positive valence and negative arousal.

4.4 Experimentation

4.4.1 Methodology

In order to create automatic music emotion recognition models based on our annotated dataset, we use two supervised machine learning approaches, that is, context-based and context-free classification. To this end, we perform three classification tasks, namely valence classification (VC), arousal classification (AC) and quadrant classification (QC). Though our dataset has average valence and arousal annotations on a continuous scale, we chose to perform classification instead of regression owing to the bimodal distribution of the ratings (See Figure 4.1). VC involves the prediction of whether the valence is positive or negative for a given track's lyrics. Similarly, AC predicts whether the arousal is high or low. QC involves predicting the quadrant the lyrics belong to.

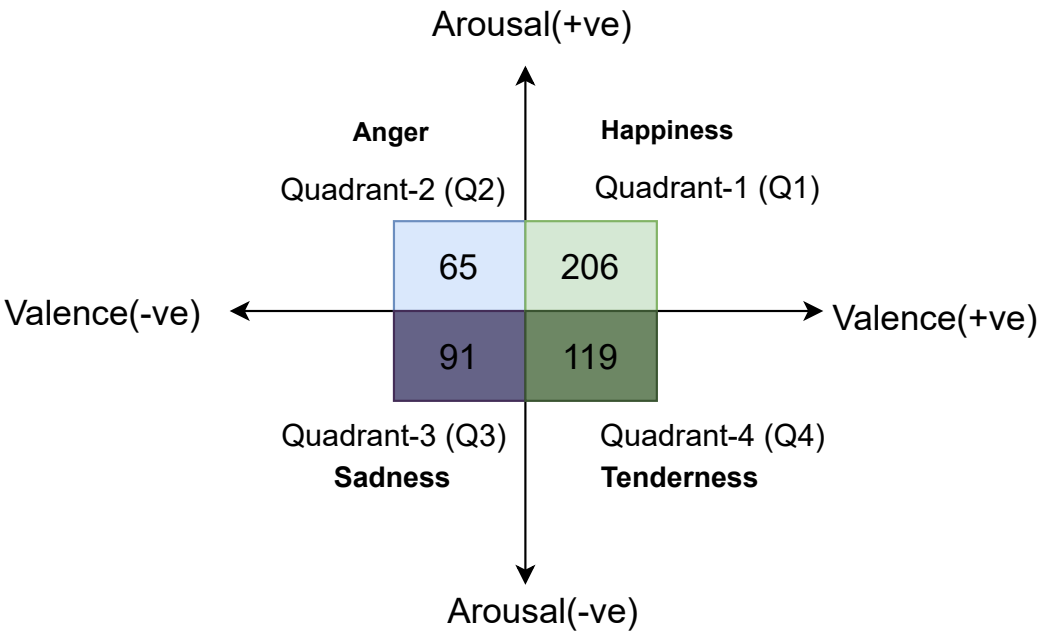

Figure 4.2 Song distribution in VA plane

SVM is a supervised learning algorithm that projects the input to a higher dimensional plane and finds hyperplanes to differentiate two or more classes. XLM-R is a multilingual pre-trained model for many South Asian languages, including Telugu, and Hindi, amongst others. It uses RoBERTa [38] transformer architecture as its base for pre-training, which improves BERT [20] by training on longer sequences and more data. It performs very well on downstream natural language processing(NLP) tasks like sentiment analysis and other natural language inference (NLI) tasks. It is also competitive with strong monolingual models on NLI tasks. It also helps to deal with codemix text [46]. Hence, we use both these methods for

For the SVM-based (context-free) classifications, we use the popularly employed TF-IDF features to represent lyrics to give as input to SVM for the classification task. SVM with linear kernel and TF-IDF have been implemented using 'scikit-learn' [50] library with default values. This is an elementary context-free method with basic features of TF-IDF for text classification.

For fine-tuned XLM-R model (context-based), we use the pre-trained(xlm-roberta-base) for the Sequence Classification task from the Hugging face [4] library to fine-tune our dataset. We use a learning rate of 2e-6 for VC and AC, and 4e-6 for QC with AdamW optimizer [39]. We also use the default max sequence length of 512 and a batch size of 8 for our implementation. We perform 10-fold cross-validation on our dataset and report the average accuracy (also known as the micro-averaged F1-score) and macro-averaged F1-score (F1). The macro-averaged F1-score

[4]https://huggingface.co/docs/transformers/model_doc/xlm-roberta

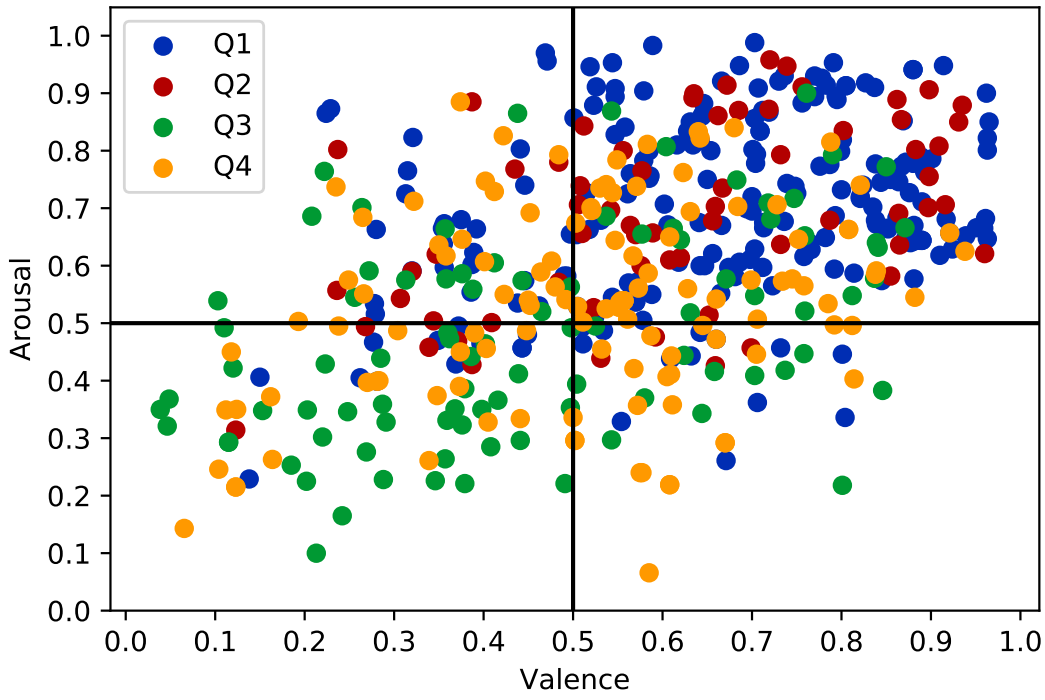

Figure 4.3 Distribution of Spotify-retrieved VA values by quadrants of annotated VA values

($\mathcal{F}_1$) is given by the formula 4.1.

$$F1_x = 2\frac{P_x R_x}{P_x + R_x}; \qquad \mathcal{F}_1 = \frac{1}{n}\sum_x F1_x \tag{4.1}$$

P_x, R_x and $F1_x$ are the standard precision, recall and F1-score of a particular class x, and n is the total number of classes.

4.4.2 Spotify Features

Spotify provides valence and arousal for each track in addition to several other features like danceability and instrumentalness, amongst others. Spotify doesn't provide the algorithm used to calculate its features but is based on musical tracks in their entirety. The valence and arousal values from Spotify are continuous and lie in the range of 0 to 1. Tracks with a valence or arousal greater than 0.5 are considered to be positive or high, and less than 0.5 are considered to be negative or low, respectively.

To examine the association between the average annotated VA values, and Spotify's VA values, we performed Spearman's correlation. In order to get quadrant-specific insights, we also perform Spearman's correlation between the same quadrant-wise.

Table 4.1 Results of valence classification (VC), arousal classification (AC) and quadrant classification (QC) tasks

	Method	Accuracy	$\mathcal{F}_1$
VC	SVM	69.43%	54.69%
	XLM-R	**80.88%**	**77.90%**
AC	SVM	68.60%	67.61%
	XLM-R	**81.51%**	**80.71%**
QC	SVM	48.63%	34.13%
	XLM-R	**62.58%**	**58.33%**

4.5 Results

4.5.1 Classification

As expected, we observe a far superior performance of the context-based fine-tuning XLM-R model compared to the context-free SVM with TF-IDF, as shown in Table 4.1. Nearly ten percentage of song lyrics contain few verses in English or Hindi. This is due to the changing trends of Indian songs in which lyrics are code mixed with other languages. This may explain the better performance of fine-tuned XLM-R [46] compared to SVM with TF-IDF features.

4.5.2 Spotify Analysis

As shown in Table 4.2, we observe a small yet significant correlation ($r = 0.167$, $p < 0.01$) between annotated and Spotify-retrieved valence and a moderate positive correlation ($r = 0.533$, $p < 0.01$) between annotated arousal and Spotify-retrieved arousal. Quadrant-wise correlation analyses revealed significant positive correlations for high arousal quadrants Q1 ($r = 0.475$, $p < 0.01$) and Q2 ($r = 0.269$, $p < 0.05$) with Spotify-retrieved arousal. For Q4 lyrics, valence annotations correlated ($r = 0.225$, $p < 0.05$) positively with Spotify-retrieved valence. We manually compared Spotify and annotated valence values and found that some songs with positive lyrics are identified as negative valence songs based on Spotify value and vice versa. One of the songs is "DJ Tillu" from "DJ Tillu", which is a song that has positively exciting lyrics and is identified as a negative valence song based on Spotify value. Similarly, "Daakko Daakko Meka" from "Pushpa: The Rise" has negative lyrics but is identified as positive based on Spotify value.

Though we are comparing lyrics annotations with music annotations, the agreement value is less than expected. From Figure 4.3, we can observe the difference in the classification of quadrants based on Spotify-retrieved values and annotated values. One possible explanation

Table 4.2 r values of Spearman's Correlation between lyric annotations and Spotify values. (*p < 0.05 and **p < 0.01)

Songs	**Valence**	**Arousal**
All	0.167**	0.533**
Q1	-0.041	0.475**
Q2	0.053	0.269*
Q3	0.027	-0.155
Q4	0.225*	0.073
+ve Valence	0.021	-
+ve Arousal	-	0.420**
-ve Valence	0.146 (p=0.067)	-
-ve Arousal	-	0.015

Table 4.3 Cross validation

	Training Set	**Annotated**	**Spotify**
Valence	Annotated	77.90%	64.22%
	Spotify	66.67%	73.38%
Arousal	Annotated	80.71%	78.11%
	Spotify	67.43%	78.8%

for low agreement between Spotify and human annotations for Telugu songs is the difference in the mapping of musical features, which could be attributed to cultural differences.

4.6 Cross Validation

We also performed cross-validation using the multilingual XLM-R. We trained the model on manual VA annotations and tested on the same in addition to testing it on Spotify VA. Similarly, we trained on Spotify VA and tested it on both.

We observed that the models trained on Spotify VA and the model trained on manual annotations of valence did not generalize. For arousal, the model trained on manual annotations was able to generalize well, achieving an f1-score of greater than 78% on both datasets. Refer to table 4.3.

Table 4.4 Comparison with different texts

Text	Method	Spotify Valence	Annotated Valence	Spotify Arousal	Annotated Arousal
సంగీతం	Original (XLM-R)	73.38%	77.90%	78.8%	80.71%
Music	Translated (XLNet)	72.15%	79.03%	76.58%	78.21%
Sangeetam	Transliterated (XLNet)	70.84%	70.20%	76.41%	76.25%

4.7 Comparing Different Texts

We also compared the performance of emotion recognition models on three types of text. The results are in the table 4.4) shown are from 10-cross validation.

We observe similar performances of models with different text inputs. The translated text (English) with XLNet performs slightly better than the original text with XLM-R for annotated valence, contrary to our expectations of observing a dip in performance due to information loss during translation. This may mean that models are well-pre-trained in the English language. The transliterated text performs inferior when compared to others, implying that the XLNet model is not largely pre-trained on transliterated text. Also, there may be a loss of information when transliterating the text.

4.8 XLM-R on other datasets

We further validate the XLM-R model on two existing lyrics datasets of BolLy and Sentiraama.

4.8.1 BolLy

This dataset [9] comprises 1055 Bollywood songs from 1970-2017. Each song in the dataset has been manually annotated by three annotators for valence. The annotations were done solely based on lyrics, and we were not allowed to listen to songs. The three annotators are native Hindi speakers in the age group of 20-24. Out of 1055 songs, 712 were annotated as positive, and the remaining 343 were annotated as negative.

4.8.2 Sentiraama Lyrics

This dataset [25] comprises 339 Tollywood songs. They were manually annotated by two annotators for valence. The annotators are native Telugu speakers who were given the situational

Table 4.5 Results of classification by Valence on Sentiraama and BolLy.

	Method	Accuracy	Precision	Recall	F1-score
Hindi	SVM [9]	75.49%	82%	63%	63%
	XLM-R	80.18%	77.85%	75.89%	76.39%
Telugu	SVM [25]	-	85%	84%	84.25%
	XLM-R	87.26%	86.62%	86.23%	86.29%

context of the movie along with the lyrics for annotation. Out of 339 songs, 230 are positive, and 109 are negative.

4.8.3 Architecture

We propose the use of pre-trained multilingual transformer model XLMRoBERTa (XLM-R) [17]. It is a multilingual version of RoBERTa [38], which is pre-trained on 2.5TB of filtered common crawl data containing 100 languages, including Hindi and Telugu. XLM-R is very competitive with strong monolingual models on natural language understanding tasks. For implementation, we use the rich pre-trained (xlmroberta-base) model. We fine-tune the large model with AdamW optimizer [39] with a learning rate of $2e^{-6}$. We use Cross-Entropy Loss for calculating loss. We use the default max length of lyrics as 512 and batch size 8.

4.8.4 Results

We maintain the same train-test split as done by the original authors of the dataset. We compare our XLM-R with their models on BolLy and Sentiraama datasets. We observe an improved performance of the context-aware XLM-R model compared to previous methods. We use the same metrics of accuracy, precision, recall and F1-score as above. (Refer table 4.5)

4.9 Perceptual Validation

To analyze the misclassified lyrics of the model, we perform perceptual validation of the model's performance using two new annotators. The two new annotators annotated the misclassified lyrics of the model. The annotators were asked to rate the song lyrics as positive or negative based on the emotion perceived from solely the lyrics. About 50% of the misclassified lyrics were agreed upon by the new annotations.

This is because emotion perception is highly subjective and depends on individuals. There are lyrics which contain positive and negative aspects, which creates ambiguity for annotators. Familiarity with the song's audio may also lead to a bias in annotations. The familiar audio

features may confound in determining the emotion conveyed by lyrics. So, having a limited set of annotators results in the model learning annotations agreed upon by the annotators, which may not generalize for all. Having more annotators can help improve the quality and reliability of the dataset. Typically musical excerpts are annotated for perceived emotion by 15-20 annotators to ensure only those excerpts with the high agreement are included in a dataset. This is one limitation of the current dataset the excerpts have been annotated by a limited number of annotators.

4.10 Conclusion

In this study, for the first time, we created a dataset of Telugu song lyrics manually annotated for both valence and arousal. In addition, we also provide Spotify IDs. We also show the better performance of the context-aware XLM-R model on three datasets compared to context-free models. The differences between annotated VA ratings and Spotify VA values highlight the need to build cultural-specific models for better song recommendations, especially since there has been a staggering increase of Indian users on Spotify [1]. We also observe the model trained on arousal human annotations' ability to classify Spotify's arousal values.

Chapter 5

Preferences for Music with and Music without lyrics

5.1 Introduction

There are many songs which do not have lyrics to convey emotions. However, such songs which do not contain lyrics are also listened to by millions of people around the world. Lyrics are crucial in contributing to musical enjoyment for many, while for some, it is all about the way the music sounds. This difference can be one reason why some individuals have a liking for music without lyrics, that is, instrumental music. We investigate the preferences of instrumental songs by individuals on online music streaming platforms.

5.2 Related Work

Studies have shown that music without lyrics can elicit strong emotional responses [12, 7]. Only a few studies [42, 8] have looked into the associations between the preference of instrumental music. [8] looked at associations between personality and preference for instrumental music as they occur naturally on online streaming platforms based only on a 3-month listening history but for a shorter listening period of 3 months. The study showed that users with high Openness and low Extraversion traits have a positive correlation with instrumentalness. However, there is a need to look at long-term music consumption since it may be less affected by seasonal variations in addition to personality manifesting in long-term musical preferences. [42, 8] use instrumentalness value from Spotify API, which predicts the likelihood of absence of vocals and treats songs with > 0.5(arbitrary threshold) as instrumental tracks. But, the instrumentalness value is not representative of the number of words in the lyrics. Some songs may contain lyrics(around 100 words) but still score > 0.5 on instrumentalness. The words in the lyrics may be the reason people like the song. So there is a need to identify instrumental songs based on the number of words in lyrics. In this study, we aim to link an individual's personality traits to the proportion of instrumental tracks(identified based on the number of words in lyrics) in their long-term listening history and evaluate using two datasets to assess the reliability of results.

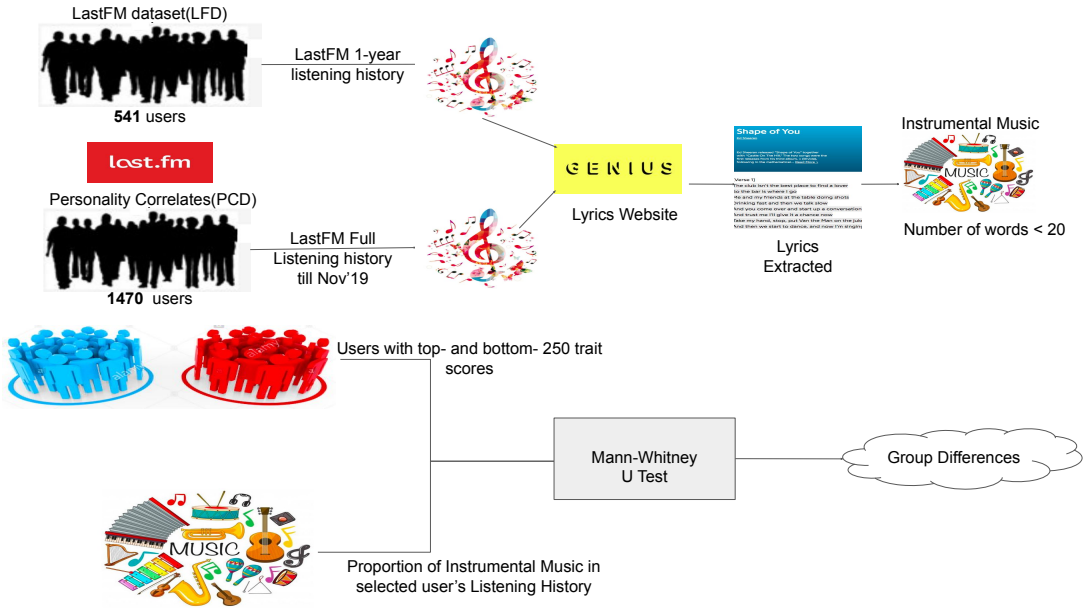

Figure 5.1 Methodology

5.3 Datasets

The datasets from [62] (LFD) and [42](PCD) were used for the analysis. The LFD dataset contains 541 users (Mean Age = 25.4, SD = 7.3) with an average of 5312 unique songs and 16033 play counts per user. The PCD dataset (444 males, 82 females and 15 others.) contains 1470 users with an average of 23600 play counts and 4267 unique songs per user. Both datasets have the Big-Five Inventory(BFI) scores of each user. Song lyrics were extracted using the Genius website, which further helped us identify instrumental songs by examining the song structure of lyrics.

LFD dataset contains 541 users' 1-year listening history on Last.fm. [62] PCD dataset contains 1470 users' listening history until November 28, 2019, on Last.fm. [42]
We only considered those users for whom lyrics for more than 70% of their listening history could be extracted, which totalled 1120 for PCD and 399 for LFD. Both datasets have the Big-Five Inventory(BFI) scores of each user.

5.4 Methodology

For each user, we calculated the percentage of instrumental music in their listening history. Genius website was used to identify the instrumental music tracks for each user based on the number of words in the lyrics. For each trait, participants with top- and bottom-250 scores were grouped to be representative of high and low trait values. Group differences in the percentage

of instrumental music were evaluated using Mann-Whitney U tests. We further performed bootstrapping to account for Type I error and ensure that the observed differences are not due to chance. We performed Mann-Whitney U tests by considering the top 500 tracks based on play count and also all tracks in the users' listening history.

Refer figure 5.1.

Table 5.1 Results of Mann-Whitney U Tests

	LFD		**PCD**
Personality Trait	**Cronbach Alpha**	**MWU Test**	**MWU Test**
Openness	0.67	Instrumental ($p = 0.095$)	Instrumental ($p < .001$)
Conscientiousness	0.69	-	-
Extraversion	0.85	-	Lyrics ($p < .05$)
Agreeableness	0.79	-	-
Neuroticisim	0.71	Lyrics ($p < .05$)	Lyrics ($p \leq .05$)

5.5 Results and Discussion

More than 95% of the listening history had songs with lyrics. We only considered those users (n = 1120 PCD, n = 399 LFD) for whom lyrics for more than 70% of the songs could be extracted for the Genius approach. For the PCD dataset, Cronbach alphas were not provided by the authors. The Mann-Whitney U test results can be found in table 5.1 In this study, we found associations between personality traits and preferences for instrumental music. Individuals scoring high on trait Openness are associated with an affinity for instrumental music, which reflects their preferences for complex musical genres which typically have fewer lyrics (ex: jazz, classical) on online streaming platforms [42, 8]. Neuroticism associated with an affinity towards music with lyrics is a novel result. This result possibly explains their affinity towards sad music [62, 26], which typically possesses linguistic cues to enhance sadness-specific emotions [12]. Extraverts demonstrated a greater predilection for music with lyrics in PCD, as observed by a previous study on online consumption [8], albeit of upbeat and energetic genres. The limitation

of this study is the threshold(< 20 words) used for identifying instrumental songs. This study can be extended by using a number of words and instrumentalness value in identifying the instrumental track.

Chapter 6

Conclusion and Future Work

Music has been an integral part of human culture for centuries, with lyrics being a crucial component of songs, conveying a wide range of emotions, thoughts and messages. In the first study, we proposed the use of transformer-based models using XLNet architecture to identify emotional connotations of music based on lyrics in the English language, outperforming existing methods for multiple datasets. Our multi-task helped in faster convergence and reducing overfitting. A robust methodology was also used to enhance web crawlers' accuracy in extracting lyrics.

However, there is a lack of datasets for Indian language songs containing both manual valence and arousal ratings of lyrics. A new manually annotated dataset of Telugu songs' lyrics collected from Spotify was presented, with valence and arousal annotated on a discrete scale. The fine-tuned XLM-R model performed better than the TF-IDF with SVM for identifying valence, arousal and respective emotion quadrant from lyrics. We observed the difference in the quadrant classification of Spotify-retrieved and annotated values. One possible explanation for low agreement between Spotify and human annotations for Telugu songs is the difference in the mapping of musical features, which could be attributed to cultural differences.

We also validated the context-aware XLM-R on two Indian language lyrics valence datasets and observed its superior performance over context-free models. In addition, a perceptual validation study was conducted on misclassified lyrics and found about 50% of the misclassified lyrics were agreed upon by the new annotations. This implies the high subjectivity of emotion perception and the need for at least 15-20 people to annotate an excerpt.

Furthermore, a study was conducted to understand the individual differences between the preferences of music with and without lyrics, with findings suggesting that individuals scoring with the dominant Openness trait preferred music without lyrics, reflecting their preferences for complex genres, while individuals with dominant Neuroticism trait showed affinity towards music with lyrics.

This study can improve applications such as playlist generation of music with similar emotions. Also, music recommendation systems can derive from incorporating emotional connotations of lyrics with acoustic content-based and collaborative filtering approaches.

6.1 Limitations

- While the transformer-based approach for emotion recognition using lyrics outperformed existing methods, it may not be suitable for all types of music or languages. The effectiveness of this approach may vary depending on the dataset and the specific characteristics of the music being analyzed.
- The manually annotated dataset of Telugu song lyrics is relatively small and may not be representative of all Telugu music.
- The human annotations and perceptual validation study was conducted on small sample size and may not generalize to larger populations.

6.2 Future Work

- Lyric emotion datasets can be created for other Indian and low-resource language songs using open-sourced annotation methods.
- Complementary information can be extracted from music, improving the performance of emotion recognition models.
- Multi-lingual lyric emotion recognition model can help in having a model for identifying lyrics in multiple languages.
- The transformer models trained on the datasets can act as a baseline to develop personalized recommendation systems.

Related Publications

1. Yudhik Agrawal, R Guru Ravi Shanker, and Vinoo Alluri. "Transformer-based approach towards music emotion recognition from lyrics." European Conference on Information Retrieval. Springer, Cham, 2021.

2. R Guru Ravi Shanker, and Vinoo Alluri. "Music Emotion Recognition via Deep Learning and comparison with human perception."9th Annual Conference of Cognitive Sciences. 2022.

3. R Guru Ravi Shanker, Yudhik Agrawal and Vinoo Alluri. "Personality Traits and preference for Instrumental Music on music streaming platforms." 16th International Conference on Music Perception and Cognition. 2021.

4. R Guru Ravi Shanker, B Manikanta Gupta and Vinoo Alluri. "Associating emotion perception with Spotify-labeled emotions in Telugu songs" 17th International Conference on Music Perception and Cognition. 2023.

5. R Guru Ravi Shanker, B Manikanta Gupta, BV Koushik and Vinoo Alluri. "Tollywood Emotions: Annotation of Valence-Arousal in Telugu Song Lyrics",
arXiv, https://arxiv.org/pdf/2303.09364.pdf

Bibliography

[1] *India among Spotify's top engagement markets: Gustav Gyllenhammar, VP.* https://economictimes.indiatimes.com/industry/media/entertainment/india-among-spotifys-top-engagement\-markets-gustav-gyllenhammarvp/articleshow/89878621.cms. 28

[2] J. Abdillah, I. Asror, Y. F. A. Wibowo, et al. Emotion classification of song lyrics using bidirectional lstm method with glove word representation weighting. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 4(4):723–729, 2020. 12, 19

[3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. J. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*, pages 30–38, 2011. 19

[4] Y. Agrawal, R. G. R. Shanker, and V. Alluri. Transformer-based approach towards music emotion recognition from lyrics. In *European Conference on Information Retrieval*, pages 167–175. Springer, 2021. 19

[5] N. Ahmadpour. Occ model: application and comparison to the dimensional model of emotion. In *KEER2014. Proceedings of the 5th Kanesi Engineering and Emotion Research; International Conference; Linköping; Sweden; June 11-13*, number 100, pages 607–617. Linköping University Electronic Press, 2014. 10

[6] F. Aisopos, G. Papadakis, and T. Varvarigou. Sentiment analysis of social media content using n-gram graphs. In *Proceedings of the 3rd ACM SIGMM international workshop on Social media*, pages 9–14, 2011. 6

[7] S. O. Ali and Z. F. Peynircioğlu. Songs and emotions: are lyrics and melodies equal partners? *Psychology of music*, 34(4):511–534, 2006. 29

[8] I. Anderson, S. Gil, C. Gibson, S. Wolf, W. Shapiro, O. Semerci, and D. M. Greenberg. "just the way you are": Linking music listening on spotify and personality. *Social Psychological and Personality Science*, 12(4):561–572, 2021. 29, 31

[9] G. D. Apoorva and R. Mamidi. Bolly: Annotation of sentiment polarity in bollywood lyrics dataset. In *International conference of the pacific association for computational linguistics*, pages 41–50. Springer, 2017. 4, 9, 18, 19, 26, 27

[10] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. arxiv 2016. *arXiv preprint arXiv:1607.04606*, 2016. 6

[11] M. M. Bradley and P. J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology …, 1999. 5, 7

[12] E. Brattico, V. Alluri, B. Bogert, T. Jacobsen, N. Vartiainen, S. K. Nieminen, and M. Tervaniemi. A functional mri study of happy and sad emotions in music with and without lyrics. *Frontiers in psychology*, 2:308, 2011. 29, 31

[13] G. C. Bruner. Music, mood, and marketing. *Journal of marketing*, 54(4):94–104, 1990. 9

[14] E. Çano and M. Morisio. Moodylyrics: A sentiment annotated lyrics dataset. In *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, pages 118–124, 2017. 12, 13, 18

[15] P. Chen, L. Zhao, Z. Xin, Y. Qiang, M. Zhang, and T. Li. A scheme of midi music emotion classification based on fuzzy theme extraction and neural network. In *2016 12th International Conference on Computational Intelligence and Security (CIS)*, pages 323–326. IEEE, 2016. 9

[16] M. Cliche. Bb_twtr at semeval-2017 task 4: Twitter sentiment analysis with cnns and lstms. *arXiv preprint arXiv:1704.06125*, 2017. 4, 7

[17] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019. 18, 27

[18] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 7, 18

[19] M. D. Devi and N. Saharia. Exploiting topic modelling to classify sentiment from lyrics. In *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*, pages 411–423. Springer, 2020. 18, 19

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 7, 12, 14, 22

[21] P. Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993. 10

[22] D. M. El-Din. Enhancement bag-of-words model for solving the challenges of sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 7(1), 2016. 19

[23] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer, 2006. 5

[24] M. Fell, Y. Nechaev, E. Cabrio, and F. Gandon. Lyrics segmentation: Textual macrostructure detection using convolutions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2044–2054, 2018. 12

[25] R. R. R. Gangula and R. Mamidi. Resource creation towards automated sentiment analysis in telugu (a low resource language) and integrating multiple domain sources to enhance sentiment prediction.

In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018. 4, 9, 18, 19, 26, 27

[26] S. Garrido and E. Schubert. Music and people with tendencies to depression. *Music Perception: An Interdisciplinary Journal*, 32(4):313–321, 2015. 31

[27] Q. Han, J. Guo, and H. Schuetze. Codex: Combining an svm classifier and character n-gram language models for sentiment analysis on twitter text. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 520–524, 2013. 6

[28] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 6

[29] X. Hu and J. S. Downie. When lyrics outperform audio for music mood classification: A feature analysis. In *ISMIR*, pages 619–624, 2010. 12

[30] Y. Hu, X. Chen, and D. Yang. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *ISMIR*, 2009. 12

[31] F. Kleedorfer, P. Knees, and T. Pohle. Oh oh oh whoah! towards automatic topic detection in song lyrics. In *Ismir*, pages 287–292, 2008. 12

[32] A. L. Knutson. Japanese opinion surveys: the special need and the special difficulties. *Public Opinion Quarterly*, 9(3):313–319, 1945. 4

[33] K. Krippendorff. Computing krippendorff's alpha-reliability. 2011. 21

[34] C. Laurier, M. Sordo, J. Serra, and P. Herrera. Music mood representations from social tags. In *ISMIR*, pages 381–386, 2009. 9

[35] H. Lee, F. Hoeger, M. Schoenwiesner, M. Park, and N. Jacoby. Cross-cultural mood perception in pop songs and its alignment with mood detection algorithms. *arXiv preprint arXiv:2108.00768*, 2021. 18, 19

[36] K. Liew, Y. Uchida, and I. de Almeida. Cultural differences in music features across taiwanese, japanese and american markets. *PeerJ Computer Science*, 7:e642, 2021. 18

[37] X. Liu, Q. Chen, X. Wu, Y. Liu, and Y. Liu. Cnn based music emotion classification. *arXiv preprint arXiv:1704.05665*, 2017. 9

[38] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 7, 22, 27

[39] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 15, 22, 27

[40] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational*

Linguistics: Human Language Technologies, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. 19

[41] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva. Emotionally-relevant features for classification and regression of music lyrics. *IEEE Transactions on Affective Computing*, 9(2):240–254, 2016. 14, 18, 19, 20, 21

[42] A. B. Melchiorre and M. Schedl. Personality correlates of music audio preferences for modelling music listeners. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 313–317, 2020. 29, 30, 31

[43] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 6

[44] D. Nath, A. Roy, S. K. Shaw, A. Ghorai, and S. Phani. Textual lyrics based emotion analysis of bengali songs. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 39–44. IEEE, 2020. 9, 18, 19

[45] B. Ohana and B. Tierney. Sentiment classification of reviews using sentiwordnet. In *9th. it & t conference*, volume 13, pages 18–30, 2009. 19

[46] X. Ou and H. Li. Ynu@ dravidian-codemix-fire2020: Xlm-roberta for multi-language sentiment analysis. In *FIRE (Working Notes)*, pages 560–565, 2020. 22, 24

[47] P. Pandey and S. Govilkar. A framework for sentiment analysis in hindi using hswn. *International Journal of Computer Applications*, 119:23–26, 2015. 9

[48] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002. 4

[49] B. G. Patra, D. Das, and S. Bandyopadhyay. Mood classification of hindi songs based on lyrics. In *Proceedings of the 12th international conference on natural language processing*, pages 261–267, 2015. 18

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. 22

[51] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 6

[52] R. Plutchik. Plutchik's wheel of emotions. *Accessed (Dec 2, 2019) at: https://www. researchgate. net/publication/234005320_Discovering_Basic_Emotion_Sets_via_Se mantic_Clustering_on_a_TwitterCorpus/figures*, 1980. 10

[53] S. Qaiser and R. Ali. Text mining: use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1):25–29, 2018. 5

[54] K. Rakshitha, H. M. Ramalingam, M. Pavithra, H. D. Advi, and M. Hegde. Sentimental analysis of indian regional languages on social media. *Global Transitions Proceedings*, 2(2):414–420, 2021. 9

[55] G. R. R. Reddy and R. Mamidi. Addition of code mixed features to enhance the sentiment prediction of song lyrics. *arXiv preprint arXiv:1806.03821*, 2018. 18, 19, 20

[56] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 2004. 18

[57] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. x, 10, 11

[58] S. Saarikallio, V. Alluri, J. Maksimainen, and P. Toiviainen. Emotions of music listening in finland and in india: comparison of an individualistic and a collectivistic culture. *Psychology of Music*, 49(4):989–1005, 2021. 19

[59] P. Sharma and T.-S. Moh. Prediction of indian election using sentiment analysis on hindi twitter. In *2016 IEEE international conference on big data (big data)*, pages 1966–1971. IEEE, 2016. 4, 9

[60] K. Shrivastava and S. Kumar. A sentiment analysis system for the hindi language by integrating gated recurrent unit with genetic algorithm. *Int. Arab J. Inf. Technol.*, 17(6):954–964, 2020. 19

[61] A. Surana, Y. Goyal, and V. Alluri. Static and dynamic measures of active music listening as indicators of depression risk. *arXiv preprint arXiv:2009.13685*, 2020. 19

[62] A. Surana, Y. Goyal, M. Shrivastava, S. Saarikallio, and V. Alluri. Tag2risk: Harnessing social music tags for characterizing depression risk. *arXiv preprint arXiv:2007.13159*, 2020. 30, 31

[63] A. Verma, S. Walbe, I. Wani, R. Wankhede, R. Thakare, and S. Patankar. Sentiment analysis using transformer based pre-trained models for the hindi language. In *2022 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1–6. IEEE, 2022. 9

[64] M. Wankhade, A. C. S. Rao, and C. Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022. 4

[65] Y. Xia, L. Wang, and K.-F. Wong. Sentiment vector space model for lyric-based song sentiment classification. *International Journal of Computer Processing Of Languages*, 21(04):309–330, 2008. 12

[66] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019. 7, 12, 14

[67] Y. Zhang, R. Jin, and Z.-H. Zhou. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1(1):43–52, 2010. 5