National College of Ireland

Sentiment Analysis of Hindi Song Lyrics using a BiLSTM Model with BERT Embeddings

MSc Research Project
Data Analytics

Jay Milind Kulkarni
Student ID: x21173176

School of Computing
National College of Ireland

Supervisor: Mr. Abdul Shahid

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Jay Milind Kulkarni
Student ID:	x21173176
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Mr. Abdul Shahid
Submission Due Date:	14/08/2023
Project Title:	Sentiment Analysis of Hindi Song Lyrics using a BiLSTM Model with BERT Embeddings
Word Count:	6711
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Jay Milind Kulkarni
Date:	17th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sentiment Analysis of Hindi Song Lyrics using a BiLSTM Model with BERT Embeddings

Jay Milind Kulkarni
x21173176

Abstract

Songs, Poems, and Music has played an important role in expressing human emotions over centuries. Over a period of time, with the development of humans and with the introduction of movies, and albums there has been an increase in the popularity of songs. With the boon of digitalization, it has now become easy to access songs all over the world. There has been an increase in research that has been carried out to identify the sentiments of the songs. However, this has been carried out mostly for languages that have a sufficient amount of digitally available resources such as English, German, Chinese or Spanish, and a few others. However, there are still other languages such as Hindi, Marathi, Latin American languages, and many more where very little research has been carried out. This research is carried out for Hindi song lyrics data which have sentiment labels as Party, Sad, and Romantic. The model that was implemented was the BiLSTM model for classification with input as text data which was converted as BERT embeddings using the BERT model. However, there was a class imbalance in the dataset with "Romantic" being the majority class, and "Party" and "Sad" as the minority classes, and an attempt to resolve this was done by introducing class weights and K-Fold Cross-validation. Three models were implemented to classify the emotions of songs and out of which the best model obtained an accuracy of 63%.

1 Introduction

Emotions are an important part of a human being, as they are the means by which a human feels happiness, sorrow, romance, anger, and others. The study of emotions with the help of machines can be termed Sentiment Analysis. Mercha and Benbrahim (2023) defined sentiment analysis as a computational study of emotions, sentiments, or a person's opinions in order to understand a human's personality, reaction to changes in the environment, or attitude towards various things or situations. Research in sentiment analysis has been mainly language centric and as a result of this there has been ample research done for languages such as English, Chinese, German, and a few others because there is a sufficient amount of data resources available for these languages in digital form. Hence these are highly-resourced languages. However, for other languages such as Asian or Indian, and Latin American there has been very little amount of research possible due to a lack of data resources. But in recent years the research in these low-resourced languages has gained a lot of attention. This research is performed on the Hindi Language which is a low-resourced language. The study of sentiment Analysis is being carried out

in various domains and one such domain is music. Chaudhary et al. (2019) explained in their study that music is also being used for healing humans such as Vedic chants in the Hindu religion. Psychologists use music therapy for treating patients with psychological problems.

Music is a form of art that includes harmonies, melodies, or rhythms. It is a combination of various sounds and compositions which in turn are responsible for evoking emotions or creating mesmerizing experiences. Music comprises instruments, vocals, or voice notes which include musical tones, lyrics, and singing. Lyrics can be considered a form of text which depicts the emotions of the music. It represents semantics and can be used for the study of sentiment analysis to identify emotions portrayed by lyricists. There has been an increase in the popularity of online music platforms such as Spotify, Lync or Amazon Music, and others, which in turn has made it to access songs from all over the world. The playlists created on these platforms are manually created depending on the emotions to give a better user experience, rather than actual identification of the user's mood during the time of using the platform. Hence, the semantic study of songs can be useful in order to create a system that suggests songs based on the user's mood or emotions.

The sentiment analysis of music has been carried out across multiple languages. Mercha and Benbrahim (2023) in their study discussed such methodologies which were carried out either by taking advantage of sentimental resources available in the rich-resource language and using it for sentiment analysis of low-resourced languages or building independent language oriented with no translation models to extract relevant characteristics from the text for performing sentiment analysis. Sentiment analysis of the semantic data is a complex and context-oriented task. Hence, it makes a challenge to predict relevantly according to the word or sentence. Thus mapping the words to their contextual sentences or phrases makes it complex. In another study, Apoorva and Mamidi (2018) performed sentiment polarity on the Bollywood song lyrics dataset where the researchers implemented three machine learning models Multinomial Naive Bayes, Bernoulli Naive Bayes, and SVM with an accuracy of 69.61%, 71.57%, and 75.49%. From both this research, it can be observed that the complexity of semantic features makes it a challenge for emotion detection, and also implementing a basic machine learning model creates a base for the classification but this can be improved with a deep learning model in order to extract the semantic features from text and identify emotions. In this research, the approach used for sentiment classification is a combination of a pre-trained model and a standalone model. This approach takes advantage of the Bidirectional encoder representations from the Transformers (BERT) multilingual model which is trained in multiple languages for creating word embeddings which are then fed to a Bidirectional Long Short-Term Memory. This research employs a unique architecture inspired by Gou et al. (2023) rather than using techniques such as Bag of Words, N-grams, or Term Frequency- Inverse Document Frequency (TF-IDF) and Machine learning models. The Hindi song lyrics dataset used for the research is available on Kaggle and contains 793 songs along with their artist names and sentiment label which are party, romantic, and sad.

1.1 Research Question

How well can a BERT-BiLSTM model identify the sentiments of Hindi song lyrics?

1.2 Research Objective

1. Analyse Hindi song lyrics data and perform stopwords removal to understand the emotions through the semantic phrases
2. Employing BERT multilingual model for the generation of word embeddings of Hindi text data.
3. Addition of class weights to solve class imbalance problem and hyperparameter tuning to improve model performance.
4. Evaluation of model using k-fold cross-validation.

1.3 Document Structure

The remainder of this research paper is as follows Section 2 is related work in this domain and the steps followed to reach a conclusion for applying the appropriate methodology will be discussed in Section 3. Section 4 is the architecture of the solution and any requirements are specified here. Section 5 explains in detail the implementation of the solution. After implementation, section 6 discusses the evaluation and results of the model. Finally, the last section 7 concludes with the results and objectives achieved and discusses the potential future work that can be carried out.

2 Related Work

In this section methodologies that were utilized in the past for sentiment analysis of high-resourced and low-resourced languages are described in depth. Sentiment Analysis of Songs or Hindi Text has recently gained a lot of attention; experiments conducted by a few researchers are discussed here. There are two subsections where the first one is research that was carried out in Indian languages such as Hindi, and Bengali text and song data are discussed to understand the methodologies used, the results obtained in the classification of sentiments for the hindi language, and their limitations and in the second one is the research carried out in different languages and domains, to understand the architectures implemented and if they can be applied on Hindi text data.

2.1 Sentiment Analysis of Indian languages

Shelke and Deshmukh (2020) in this paper performed a comparative study of sentiment analysis of various Indian languages such as Hindi, Bengali, Tamil, Malayalam, Telugu, and Konkani. The various techniques used are Lexicon, Machine learning, and dictionary. Further in the study, a list of lexicon resources available for Indian languages is shared and challenges faced in creating them are mentioned. In this study, a wide variety of languages and techniques are being implemented for sentiment analysis and also a list is provided for resources which is helpful for other research purposes. However, as this study is related to multiple low-resourced languages, a transfer learning model could have been implemented here for exploiting its knowledge of multi-lingual languages for improving the model performance. Chaudhary et al. (2019) discussed the four genres in hindi music and they are Classical, Folk, Ghazal, and Sufi. The music signals from these genres are further divided into positive arousal, negative arousal, positive valence,

and negative valence. MIR toolbox was for the extraction of spatial features. Machine learning models such as K-nearest neighbor, Naive Bayes, and support vector machine (SVM) were implemented and SVM has the best results for classification as compared to the other two models. The evaluation was done on the basis of accuracy, precision, and recall. Although the results obtained were great in this research researchers could have used deep learning models and evaluated them for this dataset.

Dhar et al. (2022) worked on the classification of emotions of Hindi song lyrical data. In this study, the classification is done on the basis of the Navrasa of Indian classical music. Feature extraction is done using TF-IDF and doc2vec and then they are applied to machine learning models such as support Vector Machines, Logistic Regression, Multinomial Naive Bayes, and KNN. SVM has obtained the best results as compared to others with an accuracy of about 66.7%. The feature extraction techniques used here did not focus on understanding the meaning of the text. The researchers could have used other feature extraction techniques and deep learning models to improve the model performances. Kumar and Albuquerque (2021) in this research evaluated the model performances of cross-lingual contextual word embeddings and zero short transfer learning which employs the knowledge from highly resourceful English language to low-resourced Hindi language. XLM-RoBERTa a cross-lingual classification model is trained using an English language dataset and zero-short transfer learning is used to evaluate two sentence-level datasets. The results obtained are significant. Hyper-parameter tuning could be tried to improve the accuracy and model performance.

Suresh Kumar and Rajan (2023) proposed and implemented a multi-modal which incorporates both acoustic and textual features to improve the accuracy. This is a transformed model and evaluates the performance using a Bi-directional GRU system with and without an attention layer. However, the transformer model yielded better results with an accuracy of 77.94% as compared to the GRU-based multimodal model with a single attention layer. The researchers have implemented a model which has unique architecture and provides better results. Bafna and Saini (2020) in this study intended to have a system that would employ supervised machine learning on a Hindi corpus for emotion classification and prediction of verses. Here, the Hindi poem dataset has been used by eager machine-learning algorithms. Stopwords were removed and then TF-IDF was applied. Random forest, Naive Bayes and SVM were implemented out of which SVM has achieved the best classification accuracy. In this research, there is scope for improvement by implementing other feature extraction techniques and deep learning models, and for evaluation, K-Fold cross Validation can be used.

Patra et al. (2018) introduced a multi-modal system that classifies emotions of songs based on audio and lyrical features for Hindi and Western songs. The researchers analyzed that an ample amount of research has been carried out on Western songs hence a mood taxonomy is proposed which is used for the annotation of both Hindi and Western songs. Further, LibSVM and Feedforward neural networks were implemented for the classification of audio, lyrics, and a combination of both features. F-measures obtained for the feed-forward neural network are around 0.751 and 0.835 for Hindi and Western songs respectively. A unique approach was implemented in this study for understanding the audio and lyric features of Hindi songs.

Sumith et al. (2022) has used binary classifiers in their study as they have two classes of emotions and they are happy and sad. These classifiers were applied to Hindi and English song lyrics datasets. Initially, Hindi text was translated into English and an English Dataset was used for comparison. Two techniques of feature extract were used

and they were Bag of Words and TF-IDF. The models that were used are SVM, Random forest, Naive Bayes, and AdaBoost and stacking of all these classifiers was done to create a single model. During stacking, only the SVM model was replaced with SVC and finally, K-fold validation was used. Although a unique approach has been implemented in this study, there is a scope for improvement as from the results it is clear that class prediction is not that effective and individual accuracy of classes is not satisfactory.

Apoorva and Mamidi (2018) performed sentiment classification of Bollywood songs from the year 1970 and these were manually annotated by annotators into Negative and Positive sentiments. The dataset was created by extracting data from websites and a lot of cleaning processes had to be done. Three models Multinomial Naive Bayes, Bernoulli Naive Bayes, and SVM. The individual accuracy obtained for classes is satisfactory. The proposed and implemented method in the research has provided good results and as an enhancement, a deep learning model can be applied to this dataset to understand its impact on model performance. Velankar et al. (2021) in their study worked on a system that classifies multi-class moods of Hindi songs in Devanagari text. The system uses contextual mood classification which is a knowledge base and it is kept updated incrementally with new data. The contextual knowledge graph with the mood classes and important terms with it represents a graph of the songs dataset used. The accuracy obtained from this method is around 64% which is a satisfactory result. The authors have also provided how this method can be utilized for applications such as summarization, retrieving context, indexing, and classification of context.

Mukherjee (2019) analyzed and tried to solve a new problem in this field and that is the classification of code-mixed text data, in which a language other than English is written in English. Hence, this study focuses on Hindi-English code-mixed data for the classification with a model divided into two parts, the first one learns the word-level features given from input text data, and the other part uses trained word embeddings as input and learns word-level features from them. The accuracy achieved by them is satisfactory as compared to the base model. This is a new problem in this field and the researchers have tried to predict labels with a suitable model architecture. Ganguly et al. (2021) performed the classification of Bengali songs using a multi-modal classifier. There are four classes of moods from Bengali songs that are predicted on two datasets one is the audio and the other is the lyrical data of songs. For the lyrical data TF-IDF and n-grams were used for feature extraction and then this data was fed to five different models which are Naive Bayes classifier, SVM, Single layer perceptron, multi-layer perceptron, and RNN. For further audio data feature extraction, the Librosa library is used and models such as RNN and LSTM were applied. The performance of SVM and single-layer perceptron was better. In this data the audio and lyrical features have been used separately, a combined study of both the features on this model could have been carried out to understand the impact of the combination of both features of Bengali songs on model performance.

This Mia et al. (2023) is another research that is being carried out on Bengali song lyric data for the classification of two or three emotion classes using machine learning and Neural networks. TF-IDF technique is used for feature extraction and implemented by Naive Bayes Classifier, Support Vector Machine, Stochastic Gradient Descent, and XGboost, for evaluation 10 Fold-cross validation is used. The models that understand neural networks are BiLSTM and BERT. The BERT model has proved to give better results as compared to other models. This is another research where the deep learning model has proved to be effective in classification as compared to machine learning models.

Nath and Phani (2021) performed a binary class classification of the Bengali song text using machine learning and deep learning. Techniques used for feature extraction are Bag of Words, Term Frequency, TF-IDF, and Word2vec. The classifier models are naive Bayes (NB), K-nearest neighbors (KNN), decision tree (DT), random forest (RF), support vector machine linear kernel (SVM), C-support vector classification with rbf kernel (SVC(rbf)), and SVM polynomial kernel (PSVM), logistic regression (LR). Deep learning models used are CNN, RNN + LSTM, and CNN + LSTM. Performance was evaluated on the basis of precision, recall, and f1-score, and the CNN model with Bag of Word features produced better results than all the other models in the study.

2.2 Sentiment Analysis in different domains and languages

Ferdosian et al. (2021) has carried out a study and implemented a sentiment analysis system for a client who is one of the global and leading in the customer management industry. The client company utilizes text analytics to understand the need of customers and provide them with a better customer experience. For this, the company uses customer feedback data for analysis and since they have customers globally they face a language barrier in understanding languages other than English. There was a need to develop a system that works on multiple languages. But they understood that it would take a lot of investment to train a model with multiple languages and especially low-resourced languages. Hence, the transfer learning model was utilized in order to classify customer experiences. The transfer learning model is trained on an initial domain to gain knowledge and transfer it to another domain. For this research knowledge of the English language model is transferred to German based model. Three models mBERT, XLM-R, and LSTM were used. The results of the pre-trained model were the best as compared to LSTM. However, the researchers could have used k-fold cross-validation for evaluating the performance of the LSTM model as it was being trained only on the German dataset used in this research and similarly, the transfer learning models were trained on large corpora of the German language, and the dataset used in this study.

Mercha and Benbrahim (2023) conducted this survey for multilingual data which is being generated due to the growth of social media, the web, and other online platforms. This study gives an overview of the methods used for performing sentiment analysis in a diverse list of languages. The research was conducted for the implementation of both machine learning and deep learning model to utilize and classify sentiments of the textual data. The study revealed that deep learning models performed better as compared to machine learning models. Here various approaches were discussed the first one was a lexicon-based approach, the second was machine learning and the third was deep learning. The lexicon-based approach exploits the ability of words that express sentiments for predicting the labels of the textual content present in the document. The researchers have conducted a good survey of techniques that can be exploited for sentiment classification and discussed the problems that can be addressed during this process and they have also achieved better results for deep learning models.

Gou et al. (2023) built a sentiment classification system of a dialogue text with model architecture using BERT embeddings for word and sentence level vectors, then the word level vectors are combined with BiLSTM to capture the semantics, and then word level and sentence vectors are connected and are fed to a linear layer for classification of emotions from dialogues. This research was carried out to improve the experience of human-machine interaction. This can be experienced over various domains and their applications

such as chatbots, customer complaints, or service centers that employes chatbots for interaction with customers. It is essential to understand the emotion of the customers or users and then generate a response to resolve their issues or provide them with solutions. In the BERT embedding processor the sentence vectors are special classification tokens (CLS). The features generated from the sentences represent the semantics of the whole sentence. The approach used by researchers is slightly different than previous ones as it exploits the advantages of the BERT model which is pre-trained on a huge corpus of the Hindi text data rather than using TF-IDF, N-grams, or Word2vec. The results obtained are better than other approaches.

Zhou et al. (2022) in their study implemented a sentiment analysis model which is different from current models which deviate from capturing semantic information from text. They introduced a hybrid model which combines doc2vec with a deep learning model and an attention mechanism. Due to the implementation of doc2vec, there has been an improvement in the overall extraction of semantic information from the paragraph vector obtained from the doc2vec model thus by reducing the loss of information. IMDB dataset and a DailyDialog dataset were used to evaluate the model performance. The model results indicate an improved model performance. This is another hybrid model which has obtained better results and employs knowledge from a pre-trained model in this case it doc2vec model on large data and applies it as an input to a deep-learning model.

Pyrovolakis et al. (2022) worked on multi-modal song mood detection which worked on combinations such as audio and lyrics data, lyrics data, and audio data separately to analyze the impact of text and audio features on the model performance. For the lyrical data embeddings were used such as Word2vec, Glove, TF-IDF, and BERT embeddings for the semantic analysis which were then used by LSTM and BERT models. For audio data, only data augmentation was performed and this was used by the CNN model for classification. For the final model of classification using both audio and lyrical features, the BERT and CNN model are combined using a technique called late fusion. This was used so that knowledge from both models can be used to achieve better results. After obtaining the results the researchers proved that audio and lyrics both features combined improve model performance as compared to individual features. This study has led to findings that can be exploited for the classification however, it is necessary to evaluate if this works across all languages or just on high-resourced languages.

Jiang et al. (2021) With the help of bag of words data representation was changed to numerical format and fed to an RNN model. The performance of the model was satisfactory and the data from different languages used are Arabic, Danish, Japanese, Spanish, and Turkish. Although the researchers have obtained satisfactory results there is a possibility of improving the performance by using Word2vec or Doc2vec instead of Bag of Words. Subramanian et al. (2022) worked on only an audio dataset of music for the classification of emotions. In this study, the performance of neural networks was compared with machine learning models. For this, a BiLSTM model was implemented where the audio data was electronically modified for feature extraction. The results obtained for this model are good even though there are multiple classes and these are happy, sad, fear, neutral, angry, and disgust. The researchers are successful in the classification of emotions but they have used audio features only. The lyrical data should also have combined with audio data to analyze the impact of both data on model performance.

Joshi et al. (2021) has created a system that uses text data for classification using deep learning models and as an enhancement CNN model is used to detect facial expressions

of the user. In this study, various combinations of deep learning models are used and they are Long Short-Term Memory (LSTM), Convolution Neural network (CNN), CNN-LSTM, and LSTM-CNN for the detection of emotions happy, sad, love, and angry. There was a lot of pre-processing of data done to remove punctuations, stopwords, symbols, HTML, etc. and then the data was converted into feature vectors using TF-IDF. The results obtained were significant, however instead of TF-IDF researchers could have used Word2vec, Glove, or BERT embeddings to improve the model performances as these pre-trained models on multi-lingual data.

2.3 Literature Summary

The literature survey carried out in 2 depicts that enough research has not been done on Hindi song lyrics. However, ample research across all other languages and domains has been performed and still, it is a popular field for analysis as there are limitations that prevent from achieving appropriate results. The method in this paper i.e. to use the BERT model for generating BERT Embeddings for feature extraction which then can be applied as input to a BiLSTM model is a unique architecture implemented on Hindi song lyrics data. This research focuses on capturing the semantic information efficiently and classifying the sentiments based on this information rather than typical feature extraction techniques such as TF-IDF or Bag of Words which fail to capture the contextual meaning of the data.

Paper	Model	Class	Accuracy
Genre Based Classification of Hindi Music	SVM	4	75.83 %
	NB		66.7 %
	KNN		64.17 %
Emotion recognition from lyrical text of Hindi songs	SVM	9	66.7 %
On Exhaustive Evaluation of Eager Machine Learning Algorithms for Classification of Hindi Verses	SVM	4	47 %
	Decision Tree		66 %
	Neural Network		50.85 %
	NB		55 %
Sentiment Classification of English and Hindi Music Lyrics Using Supervised Machine Learning Algorithms	SVM	2	67.15 %
	NB		65.69 %
	RFC		67.15 %
	ADA		66.42 %
BolLy: Annotation of Sentiment Polarity in Bollywood Lyrics Dataset	Multinomial NB	2	69.61 %
	Bernoulli NB		71.57 %
	SVM		75.49 %
Integrating BERT Embeddings and BiLSTM for Emotion Analysis of Dialogue	BERT-BiLSTM-CNN	7	85.3 %
	BERT-BiGRU-CNN		85.15 %
	BERT-BiLSTM		85.44 %

Table 1: Accuracy Results for Various Models

3 Methodology

The sentiment classification of Hindi Song lyrics data will be done by applying a deep learning model. A BERT-BiLSTM model will be applied to the labeled dataset. The methodology followed is based on CRISP-DM which is further discussed below in detail:

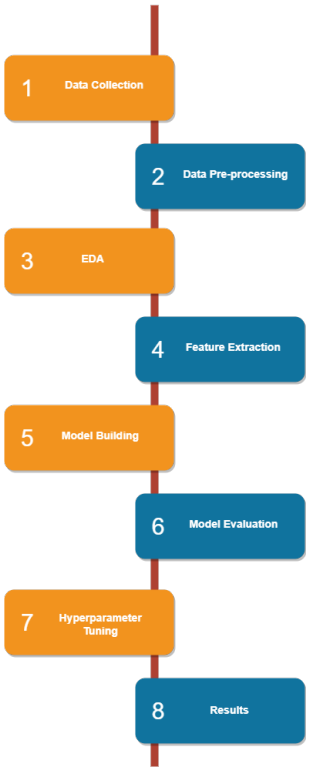


Figure 1: Steps for Implementation of BERT-BiLSTM Model

3.1 Data Collection

The data used in this research was used by [1] in their research and is publicly available on Kaggle. [2]. The dataset consists of Hindi lyrics text from 793 songs along with their artist names and type or sentiment labels as party, romantic, and sad. The data is in the form of a CSV file. Further steps would be to understand the data and it can be converted to a Python data frame for performing further operation. The following image is a snapshot of the Hindi song Devanagari text:

[1] https://github.com/rajveekadchha/mood_classification_of_hindi_songs/blob/master/NLP_Report.pdf

[2] https://www.kaggle.com/datasets/arjunramoji/hindi-songs-lyrics-with-artists

	index	Song name	type	artist	lyrics
0	1	नीले नीले अम्बर पर	romantic	किशोर कुमार	नीले नीले अम्बर पर चाँद जब आये प्यार बरसाए हमक...
1	2	अक्कड़ बक्कड़	party	बादशाह	अक्कड़ बक्कड़ बॉम्बे बो 80, 90 पुरे 100 रात के...
2	3	अखियाँ	sad	पोपोन	ओ थक गेया अँखियाँ ओ जग्ग दियाँ अखियाँ माहियाँ ...
3	4	अंग से अंग लगाना	romantic	अलका याग्निक, बिनोद राठौड़, सुदेश भोसले	अरे जो जी में आए.. अरे जो जी में आए.. तुम आज ...
4	5	अगर ज़िन्दगी हो	romantic	आशा भोसले	अगर ज़िन्दगी हो तो तेरे संग हो अगर ज़िन्दगी हो...
...	...	...	...	...	...
788	789	हैप्पी हैप्पी	party	बादशाह	विंटर का महीना उस पर तुझ जैसी हसीना बोलो फिर ...
789	790	हो गया है तुझको	romantic	लता मंगेशकर,उदित नारायण	आई अब की साल दिवाली मुंह पर अपने खून मले आई अ...
790	791	होंठों से छू लो तुम	romantic	जगजीत सिंह	होंठों से छू लो तुम मेरा गीत अमर कर दो होंठों ...
791	792	होली के दिन	party	किशोर कुमार	चलो सहेली.. चलो रे साथी.. चलो सहेली.. चलो रे स...
792	793	होशवालों को खबर क्या	sad	जगजीत सिंह	हम्म.. आ.. हा हा हा.. होशवालों को खबर क्या बेख...

Figure 2: Hindi Songs Lyrics Dataset Snapshot

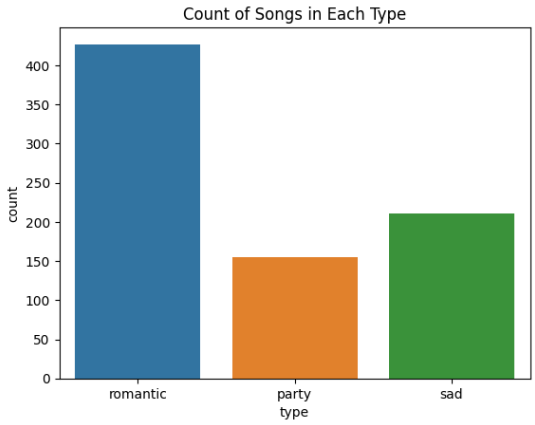
3.2 Data Preprocessing

After analyzing the data, it was then processed to make it more informative and understandable for further operations. The following are the steps:

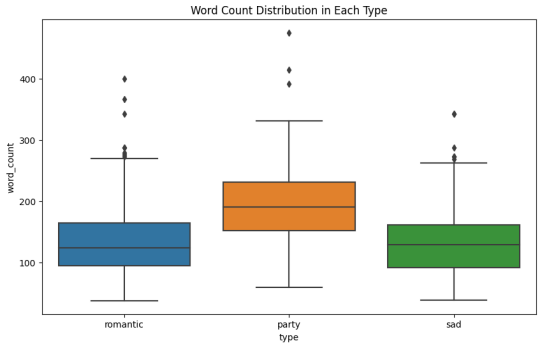
1. Removal of unwanted columns
2. Check for any missing values
3. Check for any duplicate rows
4. Stopwords Removal
5. Check for imbalance in the data
6. Logic for forming a list of most common words for each class i.e. romantic, party and sad.

3.3 Exploratory Data Analysis(EDA)

EDA is performed for having a pictorial view of the data before deciding on any testing hypothesis or assumptions. The visualizations help in making decisions on how to handle data for obtaining desired results. In this case, visualizations were created for plotting the count of each class for analyzing the imbalance in the data. Further, word clouds of the most common words were created for each class to analyze text data and understand the emotional words in them.



(a) Count of Songs in Each Type



(b) Word Count Distribution in Each Type

Figure 3: Class and Word Counts

(a) Word Cloud for Class Sad

(b) Word Cloud for Class Party

(c) Word Cloud for Class Romantic

Figure 4: Word Clouds for Different Classes

3.4 Feature Extraction

Feature Extraction is an important stage as the data is in the form of Hindi text where the script is in Devanagari and a model cannot understand the language. Traditional word embedding techniques such as Glove and Word2vec treat words as separate units and create static representations of words. However, in this case, song lyrics data is being used so rather than words capturing context is important for the identification of sentiments. The BERT embeddings represent contextualized words that capture the meaning of the words depending on the surrounding words in the sentences. Hence, BERT embeddings are used here to convert lyrical data into dense vectors. These embeddings are then incorporated by the BERT-BiLSTM model which will use this contextual information for sentiment classification.

3.5 Model Building

The model architecture of this research of applying BERT embeddings to a BiLSTM model is inspired by the architecture implemented by Gou et al. (2023). However, the

BERT model and BiLSTM layer architecture built in this research are different than Gou et al. (2023) as it is a custom Pytorch model. After a thorough literature review, it was observed that in the field of sentiment classification of languages various machine learning models and a few deep learning models were implemented. A few models in Hindi text classification incorporated transfer learning which included translating Hindi text to English text. However, the aim of this research was to work on the Hindi Devanagari script for sentiment classification and the architecture implemented by Gou et al. (2023) focused on the extraction of contextual information from the sentences and then applying it to a BiLSTM model and this type of architecture was never applied to Hindi song lyric multi-class data.

3.6 Model Evaluation and Hyperparameter Tuning

Evaluation of the model is being carried out on the basis of the following parameters:

1. Confusion Matrix: It is a table that is used to determine the performance of a classification model.
2. Precision: It is used to define the performance of a model depending on the number of positive predictions made.
3. Recall: It is the number of samples a model identifies precisely which is belonging to the target class.
4. Accuracy: It is calculated as correct predictions divided by total predictions.
5. F1-Score: It is a combination of precision and recall by taking a harmonic mean.
6. K-fold cross Validation: It is a resampling method used for evaluating the model. The value of K is equal to the number of subsets the data can be divided into. The steps involved are as follows:
    - Randomly splitting of data into K subsets
    - Every subset that is created is treated as a validation set and use the remaining subsets for training. Then evaluate it on the validation set for calculating prediction error.
    - Repeat the steps for K times until it is trained and tested on all created subsets.
    - Finally to generate the overall prediction error of the model take an average of all the prediction errors in each case.

Further for hyperparameter tuning batch size, epochs, hidden text size, dropout rate, and class weights were tuned for obtaining better model performance.

4 Design Specification

The below Figure 5 is the design for the system and the model architecture is shown on the left side of the chart.

Once the lyrical data is analyzed and preprocessed it is passed through the BERT model for creating BERT embeddings and then these embeddings are fed to the BiLSTM model which uses contextual features obtained from BERT for classification.

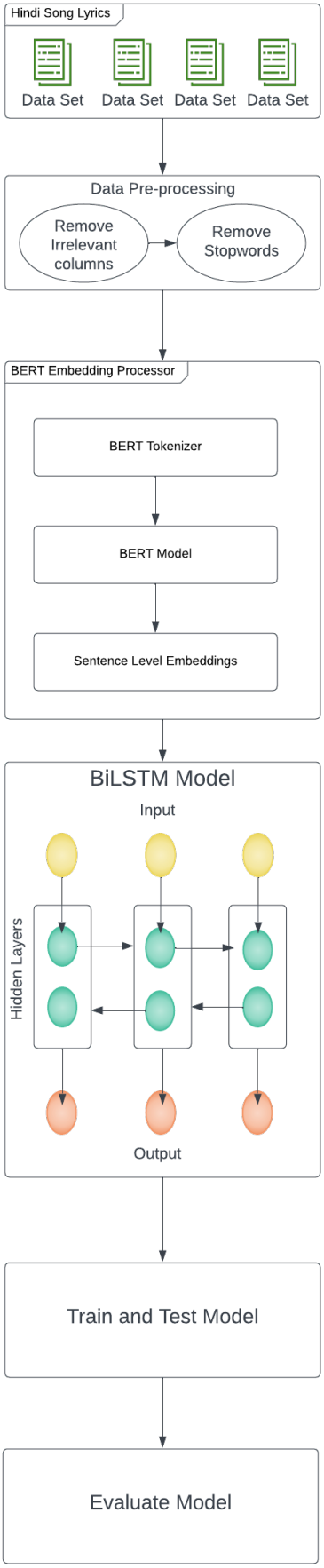


Figure 5: System Design of BERT-BiLSTM Model

4.1 BERT Embedding Processor

A pre-trained BERT tokenizer is used for tokenizing the data. A sequence of tokens is generated for lyrics data. This tokenized data is then fed to a pre-trained BERT multilingual uncased model which generates word embeddings that contextualize words meaning by considering the words around it in a sentence. These embeddings are high-dimensional sequence vectors that correspond to each token from the input sequence. Thus encoding semantic information and contextual features from the lyrics data. These generated BERT embeddings are the model's input features that empower the model to capture textual relationships which can be missed by traditional word embeddings. As these BERT embeddings pass through various layers of the model they provide contextual information which increases the model's ability to understand and classify sentiments of the song's lyrics.

4.2 BERT-BiLSTM Model

The BERT Embedding data flows into the BiLSTM layer which is responsible for identifying sequential dependencies that are present in the lyrics data. This architecture is designed in such a way that it captures information from both past and future contexts from each word in the lyrics, to make sure the model captures the relationship between the words. The dropout layer is added after the BiLSTM layer in order to prevent overfitting as it randomly deactivates some neurons during the training period in order to create a robust learning environment. The output from the dropout layer is fed to a linear layer which accounts for mapping the LSTM output to class predictions. The number of classes present for the classification corresponds to the number of output units used for this layer, in this case, there are 3 classes the total number of output units is 3.

5 Implementation

5.1 Environment Setup

The research is developed using Python 3.10.12 in a Google Colab Pro Notebook environment (RAM: 32GB, GPU: T4 + High RAM, CPU: Intel Xeon). Hindi song lyrics data was available on Kaggle and downloaded. Then this data was uploaded on Google Drive and from there, it was loaded into a python data frame using the "Pandas" library. "Numpy" for computational operations on arrays or matrices. "sklearn (scikit-learn)" is used for splitting data, label encoding, and metrics calculations. "torch" is a PyTorch library that is generally used for building and training neural networks. "transformers" is a library provided by Hugging Face for loading pre-trained BERT tokenizer and model. "matplotlib.pyplot" is used for creating visualizations and to plot graphs of model performances and data trends. "adverstool" library is to get a list of stopwords for hindi language in Devanagari font in order to process the data.

5.2 Model Building

After loading the data into a Python data frame, the index column, Song name, and artist name were dropped, and the position of lyrics and type columns were inter-changed. Further three models were built with slight changes in the architecture and parameters.

One of the models that was built was a BERT-BiLSTM model with input as BERT embeddings, however, this data was not processed i.e. the stopwords were not removed from the lyrics data to understand its impact on the model performance. Then, for the rest of the two models, BERT embeddings were created after removing the stopwords to improve the model performance.

5.2.1 BERT-BilSTM Model With Stopwords

For this model after the obtaining final data frame with columns "lyrics" and "type" the step of processing the data i.e. logic for removing stopwords was skipped and BERT embeddings were created and the type column with classes party, romantic, and sad were label encoded with 0,1,2 respectively. Then the data is divided into train and test with 80% and 20%. Further, the embeddings were converted to Pytorch tensors. The loss criterion used is CrossEntropy as this is muti-class data and an optimizer as ADAM. The batch size was kept at 10, hidden size of LSTM as 128. learning rate as 0.01 and epochs at 50. The model was trained on these parameters and tested for classification sentiments. This model was evaluated on the basis of confusion matrix, precision, recall, and AUC values. Further, to improve model performance hyperparameter tuning was done on this by changing the learning rate, batch size, hidden size of LSTM, and epochs. However, this model with the above-mentioned parameters provided satisfactory results.

5.2.2 BERT-BilSTM Model Without Stopwords

The model architecture is the same as the 5.2.1 model with minor changes. The first change is that after the creation of the final data, stopwords are removed from the data and then BERT Embeddings are created. Stopwords are a set of function words that appear frequently in a corpus and don't add any meaning to textual content. The next step is to split the data into train and test. Since the data available is less and there should be an ample amount of test data the split is set to 70% and 30% for this model. As there was a class imbalance observed while doing EDA 3a. Class imbalance occurs when one class has a lesser number of samples and another has more number of samples. In this case, the party and sad class have less number of samples as compared to the romantic class. In order to tackle this problem class weights were introduced for adding more weight to minority classes. Hence, to avoid overfitting the model a dropout layer is added to the model along with weight decay, and a learning rate scheduler is also added to have stable learning for the model to fine-tune its weights. The parameter values are set as batch size = 8, the hidden size of LSTM is 200, weights 1.9,1,2.6, the learning rate is 0.0001, and weight decay is 0.001, epochs is 80. This model was also evaluated on the basis of confusion matrix, precision, recall, and AUC values. After performing hyperparameter tuning and comparing the results of models, this model has performed well.

5.2.3 BERT-BilSTM Model Without Stopwords and K-Fold Cross Validation

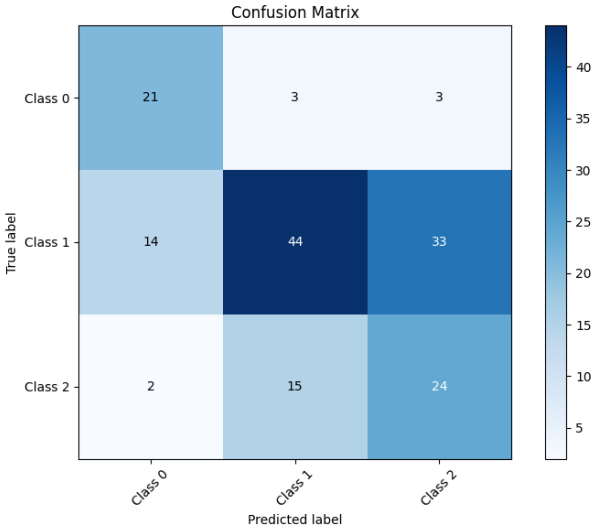
The third model architecture is similar to 5.2.2 but the learning rate scheduler and weight decay are removed and the stratified k-fold technique is used to ensure proportional distribution of classes during each fold. Thus reducing the risk of introducing bias towards majority classes. The parameter values that are used are epochs 40, weights 1.9, 1.8, 15,

and a learning rate of 0.001, the batch size is 4, and the value of k folds is 10. This model was also evaluated on the basis of confusion matrix, precision, recall, and AUC values, and the model output obtained is better than the rest of the models.

6 Evaluation

6.1 BERT-BilSTM Model With Stopwords

This model achieved an accuracy of 55.97% and an AUC value of 0.77. From the confusion matrix, it can be observed that the predicted value count of classes 0 and 1 is good but for class 2 it is satisfactory. However, the Precision, Recall, and F1-Score values of each class are not that good. Hence this model is an average model and hence, other models were implemented to get better results.



(a) Confusion Matrix

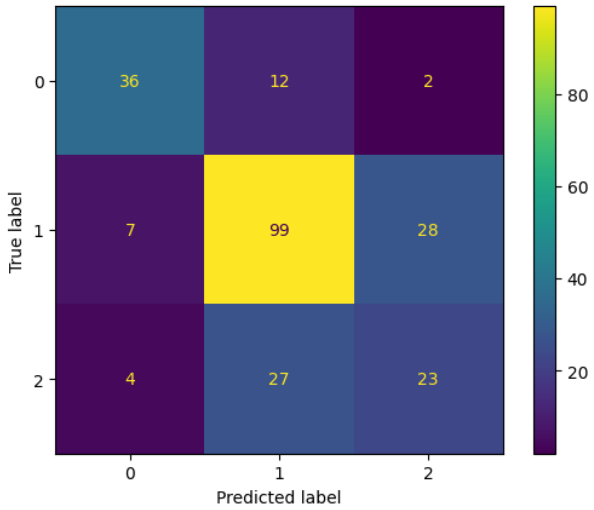
Classes	Precision	Recall	F1-Score
Class 0	0.567	0.778	0.656
Class 1	0.709	0.483	0.575
Class 2	0.4	0.585	0.475

(b) Precision, Recall, and F1-Score for Different Classes

Figure 6: Results for BERT-BilSTM Model With Stopwords

6.2 BERT-BilSTM Model Without Stopwords

The accuracy obtained by the model is 66.39% and an AUC value of 0.76 and from its confusion matrix, it can be analyzed that for Class 0 and 1 values are predicted accurately but for Class 2 it fails to predict correctly in some cases. The Precision, Recall, and F1-Score for all the classes are improved as compared to the previous model.



(a) Confusion Matrix

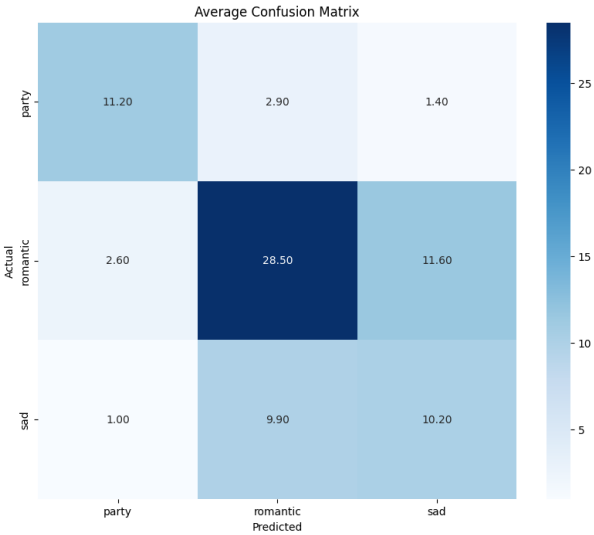
Classes	Precision	Recall	F1-Score
Class 0	0.765	0.72	0.741
Class 1	0.717	0.738	0.719
Class 2	0.433	0.425	0.428

(b) Precision, Recall, and F1-Score for Different Classes

Figure 7: Results for BERT-BilSTM Model Without Stopwords

6.3 BERT-BilSTM Model Without Stopwords and K-Fold Cross Validation

This model has achieved less accuracy than the model in 6.2 and i.e. 63% and an AUC value is 0.78. However, the precision-recall and F1-score values are better and the predicted value count for all the classes is satisfactory. Even though the accuracy is less, the results obtained from this model are better than 6.2. This model is the best model out of all the models for sentiment classification of Hindi Songs.



(a) Confusion Matrix

Classes	Precision	Recall	F1-Score
Class 0	0.77	0.72	0.74
Class 1	0.70	0.67	0.68
Class 2	0.44	0.48	0.44

(b) Precision, Recall, and F1-Score for Different Classes

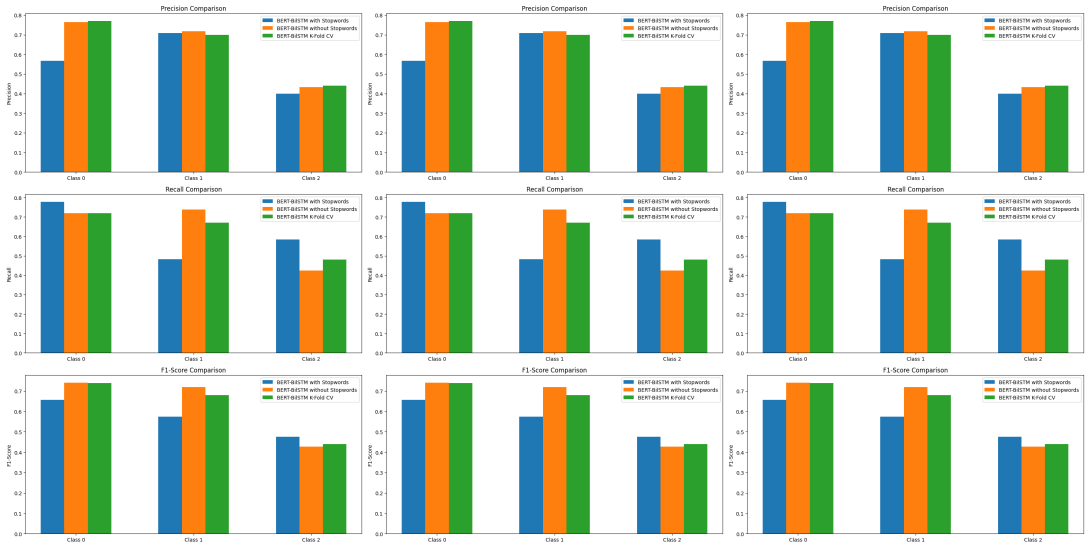
Figure 8: Results for BERT-BilSTM Model Without Stopwords and K-Fold CV

6.4 Discussion

The model described and evaluated in 5.2.3 and 6.3 respectively has proved to be better than the other two models as in 5.2.2 only class weights were introduced to resolve the class imbalance problem. Johnson and Khoshgoftaar (2019) carried out a survey on the class imbalance problems which are faced due availability of lesser data. Further, their research discussed three solutions to solve the imbalance problem, the first one was data-level techniques, algorithm-level techniques, and hybrid techniques. Data-level techniques include all the sampling techniques which modify the data to level the class counts. Another approach is algorithm-level one where the class imbalance or cost schema is applied rather than modifying the data. Hence, to maintain the contextual meaning of the data and not to modify the data with synthetic data, in this research class weights were introduced to solve the class imbalance problem and improve the model performance. This can be observed from the improved accuracy and other parameters from 6.1 and 6.2. This is one of the limitations of this research which gives scope for improvement and another limitation is the Hindi Devanagri script data which has less data available online for song lyrics. Hence, due to these limitations, the model accuracy is not as good as the baseline research conducted by Gou et al. (2023) but the precision, recall, and F1-Score are better than the baseline research also if compared with previous research work from table 1, it can be observed that an SVM model provides higher accuracy than the model accuracy of this research. However, the feature extraction techniques used for the SVM model were TF-IDF and Bag of Words which are less efficient in capturing semantic information of text data and a pre-trained BERT model is likely to capture semantic data more efficiently, one which is implemented in this research, for Hindi song lyrics sentiment classification, this is a unique approach and provides satisfactory results.

Table 2: Performance Comparison of BERT-BilSTM Models

Model	Performance Metrics				
	Class	Precision	Recall	F1-Score	Overall Accuracy
BERT-BilSTM with Stopwords	0	0.567	0.778	0.656	
	1	0.709	0.483	0.575	55.97%
	2	0.4	0.585	0.475	
BERT-BilSTM without Stopwords	0	0.765	0.72	0.741	
	1	0.717	0.738	0.719	66.39%
	2	0.433	0.425	0.428	
BERT-BilSTM without Stopwords K-Fold	0	0.77	0.72	0.74	
	1	0.70	0.67	0.68	63%
	2	0.44	0.48	0.44	



(a) Precision Comparison of Each Model (b) Recall Comparison of Each Model (c) F1-Score Comparison of Each Model

Figure 9: Metrics Comparison of Each Model

7 Conclusion and Future Work

The goal of this research was to perform sentiment classification of the Hindi text data from their semantic and contextual features. From the discussion in section 6.4 it can be observed that there are a few limitations of this research and hence, class weights were introduced and k-fold cross-validation was performed to solve the class imbalance problem.

Further, this model's performance can be improved by resolving these limitations. The first step would be to gather more lyrical data on the Hindi Devanagari text and have an equal number of counts for each sentiment class. Introduction of music data and checking the impact of the combination of both data on model performance. The addition of more sentiment classes of songs can provide a challenge for modifying model architecture.

Acknowledgment

I would like to thank Prof. Abdul Shahid my mentor for his support and guidance throughout the research and I would also want to thank all the NCI professors who provided support, guidance, and information throughout my course. In the end, I would like to express my gratitude to everyone who supported me through my whole academic journey.

References

Apoorva, G. D. and Mamidi, R. (2018). Bolly: Annotation of sentiment polarity in bollywood lyrics dataset, *Computational Linguistics: 15th International Conference of the Pacific Association for Computational Linguistics, PACLING 2017, Yangon, Myanmar, August 16–18, 2017, Revised Selected Papers 15*, Springer, pp. 41–50.

Bafna, P. B. and Saini, J. R. (2020). On exhaustive evaluation of eager machine learning algorithms for classification of hindi verses, *International Journal of Advanced Computer Science and Applications* **11**(2).

Chaudhary, D., Singh, N. P. and Singh, S. (2019). Genre based classification of hindi music, *Innovations in Bio-Inspired Computing and Applications: Proceedings of the 9th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2018) held in Kochi, India during December 17-19, 2018 9*, Springer, pp. 73–82.

Dhar, S., Gour, V. and Paul, A. (2022). Emotion recognition from lyrical text of hindi songs, *Innovations in Systems and Software Engineering* pp. 1–9.

Ferdosian, P., Grace, S., Manikandan, V., Moles, L., Datta, D. and Brown, D. (2021). Improving the efficiency and effectiveness of multilingual classification methods for sentiment analysis, *2021 Systems and Information Engineering Design Symposium (SIEDS)*, IEEE, pp. 1–4.

Ganguly, S., Das, D., Modak, A. and Chakraborty, S. (2021). Multimodal sentiment analysis of rabindra sangeet through machine learning techniques, *Advances in Speech and Music Technology: Proceedings of FRSM 2020*, Springer, pp. 223–234.

Gou, Z., Li, Y. et al. (2023). Integrating bert embeddings and bilstm for emotion analysis of dialogue, *Computational Intelligence and Neuroscience* **2023**.

Jiang, P., Chen, L. and Wang, M.-F. (2021). Transfer learning based recurrent neural network algorithm for linguistic analysis, *Transactions on Asian and Low-Resource Language Information Processing* **20**(3): 1–16.

Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance, *Journal of Big Data* **6**(1): 1–54.

Joshi, S., Jain, T. and Nair, N. (2021). Emotion based music recommendation system using lstm-cnn architecture, *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, pp. 01–06.

Kumar, A. and Albuquerque, V. H. C. (2021). Sentiment analysis using xlm-r transformer and zero-shot transfer learning on resource-poor indian language, *Transactions on Asian and Low-Resource Language Information Processing* **20**(5): 1–13.

Mercha, E. M. and Benbrahim, H. (2023). Machine learning and deep learning for sentiment analysis across languages: A survey, *Neurocomputing* **531**: 195–216.

Mia, M., Das, P. and Habib, A. (2023). Verse-based emotion analysis of bengali music from lyrics using machine learning and neural network classifiers, *International Journal of Computing and Digital Systems* **13**(1): 1–10.

Mukherjee, S. (2019). Deep learning technique for sentiment analysis of hindi-english code-mixed text using late fusion of character and word features, *2019 IEEE 16th India Council International Conference (INDICON)*, IEEE, pp. 1–4.

Nath, D. and Phani, S. (2021). Mood analysis of bengali songs using deep neural networks, *Information and Communication Technology for Competitive Strategies (ICTCS 2020) Intelligent Strategies for ICT*, Springer, pp. 1103–1113.

Patra, B. G., Das, D. and Bandyopadhyay, S. (2018). Multimodal mood classification of hindi and western songs, *Journal of Intelligent Information Systems* **51**: 579–596.

Pyrovolakis, K., Tzouveli, P. and Stamou, G. (2022). Multi-modal song mood detection with deep learning, *Sensors* **22**(3): 1065.

Shelke, M. B. and Deshmukh, S. N. (2020). Recent advances in sentiment analysis of indian languages, *International Journal of Future Generation Communication and Networking* **13**(4): 1656–1675.

Subramanian, R. R., Ram, K. A., Sai, D. L., Reddy, K. V., Chowdary, K. A. and Reddy, K. D. D. (2022). Deep learning aided emotion recognition from music, *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, IEEE, pp. 712–716.

Sumith, N., Wagle, S., Ghosh, P. and Kishore, K. (2022). Sentiment classification of english and hindi music lyrics using supervised machine learning algorithms, *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, IEEE, pp. 1–6.

Suresh Kumar, S. A. and Rajan, R. (2023). Transformer-based automatic music mood classification using multi-modal framework, *Journal of Computer Science & Technology* **23**.

Velankar, M., Kotian, R. and Kulkarni, P. (2021). Contextual mood analysis with knowledge graph representation for hindi song lyrics in devanagari script, *arXiv preprint arXiv:2108.06947* .

Zhou, Y., Zhang, Q., Wang, D., Gu, X. et al. (2022). Text sentiment analysis based on a new hybrid network model, *Computational Intelligence and Neuroscience* **2022**.