International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Jan-24)

http://dx.doi.org/10.12785/ijcds/150128

Verse-Based Emotion Analysis of Bengali Music from Lyrics Using Machine Learning and Neural Network Classifiers

Maraz Mia[1], Pulock Das[2] and Ahsan Habib[3*]

[1,2,3]*Institute of Information and Communication Technology, Shahjalal University of Science and Technology, Sylhet, Bangladesh*
**: Corresponding Author*

Received 4 May. 2023, Revised 2 Jan. 2024, Accepted 6 Jan. 2024, Published 15 Jan. 2024

Abstract: Throughout the decades, music has been involved with human emotions and inner imaginations. Along with the overgrowing demands for the overhead reduction of human-aided processes, the urge for automation in the linguistic domain has become a preponderance. Research related to Natural Language Processing (NLP) has been flourished well enough in the case of English and other contemporary modern languages. Although Bengali is an enriched multipurpose language with so wide varieties and precious dialects, very few researches have been conducted on emotion recognition in Bengali music. Besides, the Bengali music archive is blessed with pioneering works of some world-class writers, and in addition to that, modern music enthusiasts' versatile music predilection has made the task of automation in sentiment analysis quite interesting and challenging. In this study, we introduce a verse-based emotion analysis system for Bengali songs that is able to identify certain emotions from textual data. To appraise the relevant results of our resorted approach, we used several Machine Learning (ML) and Neural Network (NN) models. Eventually, we were able to achieve the best accuracy of 80% and 65% using the Bidirectional Encoder Representations from Transformers (BERT) model for both three and two emotion classes.

Keywords: Natural Language Processing, Music Information Retrieval (MIR), Sentiment Analysis, Multi-Class Classification, Machine Learning, Neural Network

1. INTRODUCTION

Music emotion detection from textual data is a process of identifying the sentiments or mood represented by the lyrics of a musical work. It is useful in various applications, such as music recommendation systems, personalized playlists, and mood-based music selection. With the advent of more sophisticated and advanced machine learning and neural network techniques, automation for the music emotion detection process from textual data has become one of the prominent aspects of Natural Language Processing (NLP) related tasks. Our research work lies in the domain of Music Information Retrieval (MIR) and its application disperses in various aspects of life. The purpose of MIR is to develop computational methods to automatically analyze, organize, and retrieve music data. MIR involves the application of signal processing, machine learning, and information retrieval techniques to extract useful information from audio signals and music-related data, such as metadata, lyrics, and user-generated contents [1, 2, 3].

We can define sentiment as one's particular view or thought on a subject matter based upon a feeling that derives from the inherent topics of the related subject. For several decades emotion analysis, which is principally concerned with sentiment or opinion conveyed by humans, has been acting on the progressive trend. The emotion extraction process is conducted very systematically and contingent on the resources that are written using natural language [4]. Music is the most involved matter in human life and more common than we estimated previously, it works as mental catharsis for many individuals. In the context of music literacy which includes poetry and lyrical works, Bengali is blessed with so many of the literary wonders of world-class songwriters, yet the digitization of this vast musical encyclopedia is still a long way to go. Day by day, the number of Bengali music enthusiasts has been growing rapidly because the contents of Bengali songs have attracted more people, defying the national boundary. Additionally, the rapid growth in the online community and services such as YouTube, Amazon, Spotify, etc., have turned the data conferring and availability process so smooth and easy that people from different groups, ethnicity, or culture have access to the same materials outlined on the underscored media.

Amongst the many other features of a song such as vocal, instrumental tune, genre, and rhythm, lyrics tend to be the most pivotal element as they affect human's feelings

E-mail address: maraj262000@gmail.com, pulockkamol50@gmail.com, ahabib-iict@sust.edu

and have a direct connection with the listener and the contents of music usually reflects a certain perspective of life as well as the personality of a person because music has a huge impact on an individual's life which let the person listen to music that suits his perceptiveness at best [5]. In terms of regular usage, music emotion detection from textual data using machine learning and neural network techniques has the potential to improve the user experience in various music-related applications and to enable new forms of musical expression and creativity.

In this research, at first, we created a dataset of verses from the selected Bengali song lyrics annotated with three different emotion labels. Then we applied several ML and NN models to our dataset in a supervised multi-class classification way and tried to achieve the best verse-wise emotion-predicting model. After getting the best classifier, we created a UI (user interface) integrated with the model using HTML (Hypertext Markup Language) and the help of Flask which is a micro Web framework written in Python. In summary, our contribution to the paper includes:

- Building a dataset that contains verse-wise emotions for Bengali songs,
- Introducing multi-class classification approach in Bengali song's sentiment analysis,
- Applying and evaluating different ML and NN models on our dataset,
- Integrating a UI in a local environment to visually illustrate the classification scheme.

The rest of the paper is segmented as follows. Section 2 discusses the related works. Section 3 provides a detailed description of the resorted approaches. Section 4 discusses the functionality of the models. Section 5 presents the performance evaluation and detailed elaboration of the proposed classifiers. Finally, Section 6 concludes the paper.

2. LITERATURE REVIEW

One of the most significant Indo-Iranian languages, Bengali (or Bangla, as locals call it), is the sixth most prevalent language in the world and is spoken by more than 265 million people. Moreover, it is the mother tongue of the people of Bangladesh and 2nd language of India. However, being this phenomenal language, the digitized resource in the case of the Bengali language is quite diminishing compared to English and other well-digitally documented languages. Also, the Bengali language has so many complicated grammatical structures and diverse linguistic usage which make the underscored tasks more challenging. Nevertheless, quite a number of research works have been continued on sentiment analysis from social media comments, speech recognition, MIR, text classification, etc.

In the case of sentiments, emotions can be defined as thoughts on a particular subject by reasoning internal feelings. At least seven so-called primary emotions are represented by cultural facial expressions: sadness, joy, fear, wrath, surprise, disgust, and contempt, provided in the study of Robinson and Hatten [6]. Music can indeed evoke real emotions rather than merely fantasies, but it's also true that some emotional feelings are more likely to do this than others. Sometimes seemingly happy music can be turned into a lugubrious one if some changes in vocal tone occur. So, along with lyrical components, a vocalist's effort on a song is also important to express the real emotion of the song. However, they also showed that the major role of conveying a sentimental issue is played by the lyrical part of the music. Thus, to fully extract the real meaning of a song, we have to deep dive into the lyricist's writing style and the real nature of the lyrics themselves.

Machine learning techniques have already been used to complete a significant amount of work in the field of sentiment analysis. Analysis of the emotional content or sentiment of songs is a topic that has not been extensively studied for low-resource languages. Although being a relatively old subject of study, the majority of the work has only been done for languages with abundant resources. Chen and Tang [7] suggested techniques that are required to categorize music by moods even from music files uploaded to social networks and achieved an F1-score of 88%. Napier and Shamir conducted a study regarding a historical background for English songs along with modern hit songs which reported a steady decline in the interest in the happy vibe in the lyrical contents and more inclination toward sad and negative tones [8]. The publication of Patra et al. [9] dealt with emotion detection for Hindi songs, resulting in a maximum F1-score of 0.68 for a polarity-based approach and 0.38 for multi-class classification. Elfaik and Nfaoui used the Deep Learning (DL) model in various Arabic datasets for binary sentiment classification in their research [10] and achieved the best F1-Score of 95.90%. On the contrary, very few works have been done in Bengali in the field of music sentiment analysis. But, most of the work mainly dealt in a polarity-wise strategy, more specifically the songs were identified as either positive or negative (songs with emotions such as happiness, motivation, excitement, calm, and astonished were considered as positive and anger, sad, anxious, gloomy, tensed, and rebellious were considered negative) where the accuracy is considerably higher because of binary classification and in consideration for variety in songs collection [11, 12, 13]. Unfortunately, we haven't found any profound work regarding multi-class emotion analysis for Bengali music yet so far. Apart from this, some other studies have been done on the related subject and peripheral to our study with a multi-class classification method. Marouf and Hossian [14] came up with a binary classification system that identifies songwriters based on lyrical content with an accuracy of 86.29%. A study on emotion detection from YouTube comments achieved an accuracy of 65.97% and 54.24% accuracy in three and five-label sentiment respectively [15]. Additionally, a more closely related research compared with ours' was able to gain an accuracy of 69.36% for multiple music genres classification from lyrics [16].

While documenting our study, we have introduced several ingenious ideas and strategies most of them missing in the previous works which are most closely related to our study. First of all, approximately in all of the previous works, the whole lyrics of a particular song were taken into account in case of emotion detection. However, it's more likely that the emotions of verses tend to be different within the song's respective contexts. We carefully analyzed the lyrical patterns and contexts of Bengali songs' verses and found that even in a love-related song, there are many occurrences of verses expressing sad feelings. Secondly, We also created a new verse-based textual dataset for Bengali songs with appropriate emotion tags. To create this, we used an existing collection of Bengali song lyrics and then applied for a verse separator program. Finally, we accumulated each verse generated from the separator program and saved those into a secure file for later annotation. Thirdly, the multi-class classification process has been implemented in many NLP-related studies however, all of the previous work related to sentiment analysis of Bengali songs mainly focused on binary classification. Here, we introduced three emotion classes and predicted them accordingly.

3. METHODOLOGY

In this section, the steps of our proposed approach will be discussed. In the first section, the resorted dataset and an overview of our dataset are given. Then, our novel proposal of verse-based segmentation is discussed in the second section with details. Further, a summary of multi-class classification for Bengali music is provided in the third section. In the fourth section, we will discuss the emotions to be detected that we choose for our study. After that in section five, the dataset cleanup process is given. In section six we will go through the feature extraction process for our models. Finally, we will provide a brief overview of our selected ML and NN models in section seven.

A. Dataset Preparation

Dataset annotation is an important step in natural language processing (NLP) to train machine learning models. As Bengali is a low-resource language, the task was quite ticklish because we had to build our custom dataset with relevant emotion tags. It was also more difficult than we imagined because the literacy works of Rabindra Nath Tagore, Kazi Nazrul Islam, Dwijendralal Roy, and many other writers had a wide range of literary styles and even the contents of the songs are so versatile that tagging one particular verse with the proper label was very vacillating. The main obstacle to the Bengali language for computer processing is the lack of a standardized, annotated training dataset. However, we found a well-documented collection of Bangla song lyrics and later we created our own dataset out of it by applying a set of rules to identify verses from song lyrics.

TABLE I. Overview of Datasets

Properties	BanglaMusicStylo	Our Dataset
Format	DOCX	CSV
#Songs	2781	2152
#Unique Verses	9516	6500
#Words	149483	138766
#Unique Words	28749	25320
#Characters	650909	600853

We used the collection of data ”BanglaMusicStylo” prepared by Hossian et al. [17] for our experiment. Songs by numerous vocalists from various genres made up this data set. Table I contains information on the dataset. The writers of this collection compiled 204 lyricists' songs. The lyrics to more than 10 songs by 38 different lyricists are included. This has distinct folders for the song lyrics of various authors. The same folder contains various tracks by the same author. Various song lyrics are kept in text Microsoft.docx files using the Bengali font ”Siyam Rupali.” Songs from many sub-genres of Bangladeshi music, including religious songs, ethnic songs, and pop songs, are in this dataset. Deviating a little bit from the seven basic emotions [6], here we proposed a total of three emotions: love, sadness, and idealism (life-oriented). Because, according to the study of Swaminathan and Schellenberg [18], modern music listeners have a predilection toward this kind of feelings more frequently. Some of the verses had different meanings of emotion outside our selected emotion criteria, in that situation we ignored the verse. We then finally stored our data in comma-separated values (CSV) formatted file. Table I depicts an insight into our dataset and the dataset that we used to build it along with their context.

B. Verse-Based Approach

Songs consist of many different components: lyrics, verses, refrain/chorus, and meter. We can compare verses with the stanza of poems which are a group of rhyming lines that represent a portion of a poem and are usually built up into a whole poem. So, verses are different portions of a song, but combined together, they represent the entire song. The chorus of a song is the repeated part and also it is a verse itself, usually used in the beginning and ending of a particular song. In the musical sense, the meter is the measurement of continuously occurring beats or bars. Songs can have several verses or just a single verse and choruses. Approximately all of the previous research considered the whole song to be singularly sentimental and classified the song based on the polarity of negative or positive vibes. Here, we propose a verse-based analysis system because in a particular song verses may convey several different sentiments [19, 20]. To achieve the propounded idea, we had to resort to a verse extraction method, suitable only for the Bengali music writing style. We need to separate verses as each verse of a particular song can convey different meanings thus making a song attributed with more than one emotion rather than only one. The verse separation method is a way to locate and distinguish between the verses in song lyrics. Either manually or automatically, this can be accomplished by scrutinizing the song and noting the beginning and stop positions of the verses. We chose the computerized option. Here, the dataset we followed has two different styles of writing. We developed a program to separate verses from the principle dataset. Hence, we had to use two different kinds of logic for the verse extraction methods that are described below.

1st Method: It's the most frequent style adopted in the dataset by the lyricists. According to this style, verses are separated by a new line. It may contain the separator character (in most cases). But, the main splitter here is the extra new line after every verse.

2nd Method: In this structure, the ending of a verse can be determined by a resting character named ”Devanagari Double Danda” frequently used in Bengali literature. The character comprises '॥'. Though it's not a universal rule of defining the end of the verse, in this dataset, this pattern of writing also exists.

We then used a simple logic-based verse separator algorithm and applied it to the BanglaMusicStylo dataset to generate our expected dataset. First, we read the individual songs of every lyricist. A particular song has either of the writing styles stated above so, we needed to identify which category style fits the underscored song. Then we separate the verses by the defined rules and finally write the newly found verses into our main dataset file.

C. Multi-Class Emotion Classification

The idea is not new but the application of the idea to the Bengali music emotion classification is not perused well enough in previous works. As mentioned earlier in this paper, most of the preceding works on Bengali songs have been conducted using the classification task based on binary classification methods [21]. So, comparatively, the task for this was easier. But, if the number of classes is augmented, the complexity of the class identification also increases, and the accuracy measurements decrease. In this research, we introduce 3 different emotion classes with distinct meanings and feelings in each class.

D. Dataset Labeling With Emotion Tags

Data labeling is the method of attributing tags or labels to corresponding data such as pictures, videos, text, and audio. These tags serve as a representation of the class of items to which the data belongs and aid machine learning models in learning to recognize that specific class of objects when they appear in data without a tag. We primarily chose 3 emotions for our study thus our dataset has to be tagged with 3 different numerical labels. Also, the annotators of the dataset had to be some kind of savant or enthusiastic in music-related corpus and we made sure of that. Further, we validated our dataset annotation by inter-annotator opinion [11] and we got a 95% similarity between them. We follow several schemes while annotating our dataset with respected emotion labels and here is given below how we did it.

Love: When a song depicts more than one sentiment if the dominating connotation is related to love and it has an overall sentiment of affection, amorous, relational, or amiable mood, or if something is mentioned with lovely vibes, then the verse is referred to as a love emotion.

Sad: When a song depicts more than one emotion, but the prevailing meaning is sad and if the verse as a whole has a gloomy, depressing, or unhappy tone, or if something is mentioned with a sad vibe, then the verse is said to have a sad emotion.

Idealistic: When we found emotions that express idealistic views or life orientation or quotidian stuff related to human life, supernatural thoughts like life and death, and religious belief, then we annotated the verse to be an idealistic verse.

Dataset frequency distribution describes how the data is split up among the many values or categories of a specific variable or feature in the dataset. In data analysis, understanding a dataset's distribution is crucial since it can shed light on the nature of the data and aid in choosing the best analytic techniques. In Table II, information regarding our dataset's class frequency and some important word features are given. There are two well-known and the most popular lyricists in Bengali song culture. They are none other than Kazi Nazrul Islam and Rabindranath Tagore. The majority of our dataset contents are comprised of the poetry works of these two writers. Fig. 1 shows us that Kazi Nazrul Islam has more sad song-verse percentage than Rabindranath Tagore's sad songs. Here rebellious, motivation to strive for type features of Nazrul's song played a good role. Wherever Rabindranath Tagore's song count in these three classes seems to be balanced.

E. Data Cleanup

Data cleanup includes the process of locating and fixing or deleting flaws, inconsistencies, and inaccuracies in a dataset. For analysis and other uses, data cleaning aims to make sure the data is correct, full, and consistent. To process our data to be ready for future analysis, first, we created a labeled dataset with a proper emotion tag. Then before feeding the data to our model, we had to remove unwanted characters (punctuation) and words (more accurately the stopwords: Stopwords are typical terms that are often used in a language but lack any real significance or value in the context of information retrieval and natural language processing activities. For the sake of enhancing the effectiveness and precision of text processing algorithms, these terms are typically filtered out or eliminated from the text corpus). Fig. 2 explicates our approach more explicitly.

TABLE II. Emotion Classes

Class	Count	Some important Word Features from scikit-learns' Chi-Square Test [22]
Love	2544	প্রেমের,প্রেম,ভালোবাসি,তোমাকে,ভালবাসা,তোমায়,হৃদয়ে,ভালবাসি,প্রেমে,বন্ধু,মনের
Sad	1868	ব্যথা,কেঁদে,বেদনা,ব্যাথা,স্মৃতি,দুঃখ,বিরহ,বেদনার,বিদায়,কান্না,অশ্রু,নিঠুর,জ্বালা,পাষাণে,শোকে
Idealistic	2088	ছুটে,ঘাটে,টাকা,পাগলা,বাবা,মাঝি,মালা,মাঠে,শহুরে,মায়ের,গাঙ্গের,উড়ে,ব্যাথা,শহরে,মানুষ

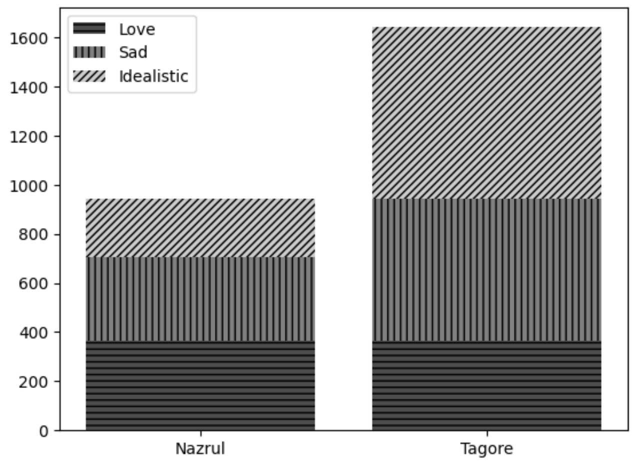


Figure 1. Data Majority Distribution

Also, we have implemented the rule-based Bengali stemmer [23] and its primary function is to standardize words and condense them into a common form, which can help to increase the precision and effectiveness of our tasks.

F. Features Extraction

Predicting a text's sentiment is the aim of sentiment analysis and feature extraction is one of the most important parts of it. There are many feature extraction models available for natural language processing (NLP) like Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), FastText, Word to Vector, Global Vectors for Word Representation (GloVe), etc. GloVE and TF-IDF feature extraction techniques are used in this research to extract important features for the models.

GloVE Specifically, the 'glove.6B.100d' vector indicates GloVe embeddings from the "glove.6B" dataset, which contains 6 billion tokens, and each word is represented by a 100-dimensional vector [24]. Using GloVe to transform the words in a text into numerical vectors is a well-known technique for feature extraction in sentiment analysis. We used a custom Bengali 'bn_glove.39M.100d' vector [25]. With GloVe, we may represent a text as a matrix of word vectors, with each row being a word in the text, and therefore capture the semantic links between words. A machine learning model can then be used to predict the sentiment of the text using the generated vector representation.

TF-IDF stands for Term Frequency-Inverse Document Frequency, which is a numerical statistic used to evaluate the importance of a word in a document. The frequency of a word within a document is measured by TF, whereas the significance of a term across all documents in a corpus is measured by IDF. We get a weighted score that represents the importance of a term in a particular document by multiplying these two values together. The idea behind TF-IDF is that if a word appears frequently in a document, but rarely in the rest of the corpus, it is likely to be a keyword for that document. In contrast, if a word appears frequently in many documents in the corpus, it is less likely to be a good indicator of the content of a specific document.

$$\text{TF} = \frac{\text{Term } (i) \text{ frequency in documents}}{\text{Total number of words in documents}}$$

$$\text{IDF}(i,j) = \log \frac{\text{Total documents}}{\text{Documents with Term } (i)}$$

$$\text{TF-IDF}(i,j) = \text{TF}(i,j) \times \text{IDF}(i,j)$$

G. Model Selection

Several models have been used successfully for training sentiment analysis models. In this study, we used several ML and NN models based on their efficiency and performance on NLP-related tasks. Our resorted ML classifiers are Naive Bayes (NB), Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), and Extreme Gradient Boosting (XGBoost) and NN models are Bidirectional Long Short-Term Memory (BLSTM), BERT and Multi-Layer Perceptron (MLP). We used our mentioned model for both three and two class prediction tasks.

H. User Interface

After evaluating all the models, the best and the most relevant models were picked and saved in a Pickle file. Later, a web-based UI was created using HTML with a user input text area for song lyrics, a prediction submit button, a reset button, and a result text area. With the help of Python's Flask API (Application Programming Interface) and the previously saved model, a verse-wise prediction for an entire song from the input text-area field

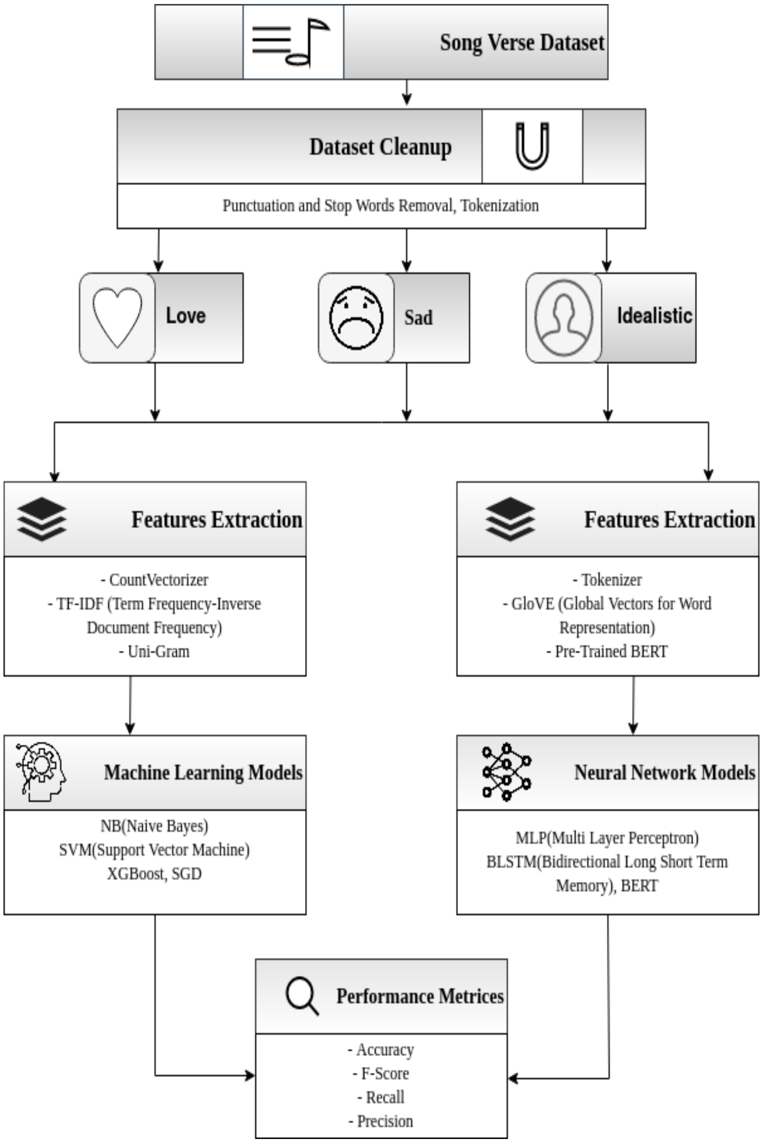


Figure 2. Emotion Detection Process Diagram

is generated and shown in the result text-area field along with the emotions and identified verses in the local browser environment (Fig.6).

4. EXPERIMENT WITH MODELS

We used in total four Machine Learning and three Neural Network models for classification purposes. In this section, we will discuss all of the mentioned models' configurations and other features and parameter setups. We needed a validation portion from the dataset because we used several different models and we had to choose the best amongst them by fine-tuning the hyper-parameters in the training session with minimum training loss. In general, we split our dataset into three sets of training, testing, and validation data with the proportion of 68%, 20%, and 12% respectively. In every classifier, we followed the same distribution pattern as illustrated in Fig. 3. We also tried different training, validation, and test data proportions manually. However, in most of the cases especially in our main classifier BERT, the resorted ratio of training, testing, and validation dataset worked more effectively. We conducted a three-class and binary classification process with each model to predict the polarity-based (love or sad) and multi-class emotion prediction.

A. ML Models

For all of the ML models, we used the TF-IDF feature matrix along with 10-fold cross-validation to train the models. We also used the classifiers that are available in scikit-learn [22] with some custom parameters.

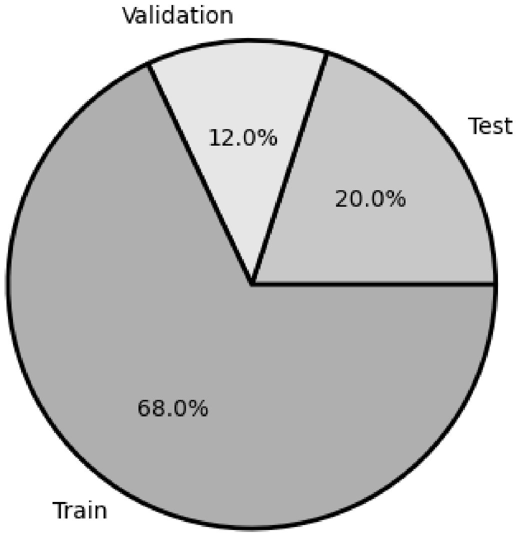


Figure 3. Training, Test and Validation Data Distribution

- The naive Bayes classifier is a probabilistic machine learning algorithm and it is based on the Bayes theorem which assumes that the features are conditionally independent of each other with the given classes. We used the Multinomial classifier with default parameters. The equation of calculation for Bayes Theorem is as follows:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Where the probability $P(A|B)$ is called the posterior, P(A) is called the prior probability, $P(B|A)$ is the likelihood probability of B conditioned on A, and P(B) is the marginal probability of event B.

- Support Vector Machine is a supervised learning algorithm and it works by finding the hyperplane that maximally separates the classes in a given dataset. Here, we used the SVC(Support Vector Classification) classifier with a linear kernel.

- Stochastic Gradient Descent is more of an optimization technique applied to Linear classifiers (SVM, logistic regression, etc.) with SGD training. As our dataset is imbalanced, we used the custom class weight, and the formula is given below.

$$\text{Weight}(i) = \frac{\text{Total Data}}{\text{Number of Classes} \times \text{Class } (i) \text{ Count}}$$

- An efficient and effective implementation of the gradient boosting technique based on decision trees is offered by the open-source package known as Extreme Gradient Boosting (XGBoost). We used the default parameters except for the loss function, for binary classification we used ”binary:logistic” and for three classes ”multi:softprob” was used.

B. Neural Network Models

Neural Network models are considered to be performing better than ML models also neural networks provide powerful new tools for modeling language and have been used both to improve the state-of-the-art models in many tasks and to tackle new problems that were not easy in the past [26, 27]. For our three resorted NN models, we had to set different configurations for each classifier.

- The MLP is made up of many layers of nodes, each of which is linked to every node in the layer adjacent to it. The nodes in each layer apply a non-linear activation function to the weighted sum of their inputs, which produces an output that is propagated to the next layer. We used the MLPClassifier from scikit-learn which includes 2662 TF-IDF input features, two hidden layers of 1024 and 256 neurons respectively with relu (rectified linear activation unit) activation function and an output layer of 3 neurons for three classes and 2 neurons for binary classification with log_softmax activation function. Further, we used the memory-efficient feature extraction approach proposed by Mahmud et al. [28].

- The BLSTM model is a form of Recurrent Neural Network (RNN) that comprises two long short-term memory (LSTM) layers, one of which processes the input sequence forward and the other in reverse. A memory cell and three gates—an input gate, an output gate, and a forget gate—that regulate the movement of data through the cell make up each LSTM layer. The output of each LSTM layer is the concatenation of its forward and backward outputs, which are passed onto the next layer or the output layer. Then, using the built-in Keras Embedding layer, we defined the embedding layer and tokenized the clean data. The word embedding itself is learned via the embedding matrix, from which the word embedding layer extracts the words' embedding vectors.

- BERT, short for Bidirectional Encoder Representations from Transformers is a Neural Network framework for natural language processing. Training a transformer model from scratch usually takes a long time. Similarly, a BERT model can also be more time-consuming. However, the purpose of BERT is to create one model that can be reused for many different tasks. For our task, we used the pre-trained Bengali BERT model implemented by Sarker [29] which is upgraded and primarily based on the suggested model of Bhattacharjee et al. [30].

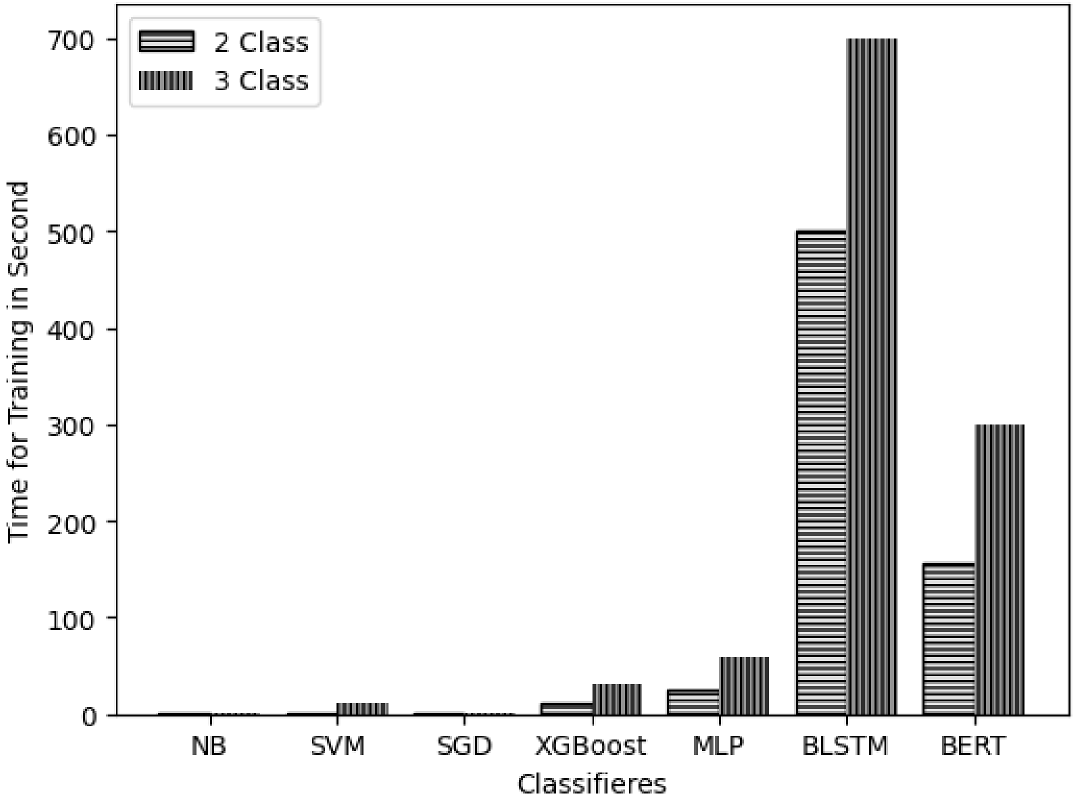


Figure 4. Models' Execution Time in Training

TABLE III. Validation Performance

Models	3-Class Accuracy	2-Class Accuracy
NB + TF-IDF	0.57	0.77
SVM + TF-IDF	0.64	0.77
SGD + TF-IDF	0.64	0.77
XGBOOST+TF-IDF	0.59	0.72
MLP [28]	0.60	0.75
BLSTM + GloVE	0.59	0.73
BERT [29]	0.93	0.95

5. RESULTS AND DISCUSSION

In this study, we proposed a new approach to music emotion classification using several ML and NN models along with different feature extraction methods. We evaluated our model on the newly created dataset of verses from various Bengali songs, including modern, folk, and rock. Metrics such as precision, recall, F1-score, and accuracy based on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values computed from the confusion matrix were used to assess the performance of our models. The precision, recall, and F1-score of our model were also higher than those of the other models, indicating better performance in identifying the correct sentiments for each song verse.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

We used cross-validation results to see if there was any over-fitting of the data for our best functional models. The cross-validation findings, as shown in Table III, strongly support the models' performance with basic classifiers. We also evaluated the training time needed for every classifier in both classification strategies and as usual the NN models more specifically the BLSTM model took the most time to be completed compared with ML models as depicted in Fig. 4.

TABLE IV. Test Performance

Models	Precision		Recall		F1-Score		Accuracy	
	3-Class	2-Class	3-Class	2-Class	3-Class	2-Class	3-Class	2-Class
NB + TF-IDF	0.59	0.78	0.58	0.78	0.58	0.78	0.58	0.78
SVM + TF-IDF	0.63	0.78	0.62	0.78	0.62	0.78	0.62	0.78
SGD + TF-IDF	0.64	0.78	0.64	0.78	0.63	0.78	0.64	0.78
XGBOOST+TF-IDF	0.61	0.73	0.59	0.73	0.59	0.72	0.59	0.73
MLP [28]	0.59	0.74	0.57	0.73	0.58	0.73	0.58	0.73
BLSTM+GloVE	0.58	0.72	0.58	0.72	0.58	0.71	0.58	0.72
BERT [29]	0.65	0.80	0.65	0.80	0.65	0.80	0.65	0.80

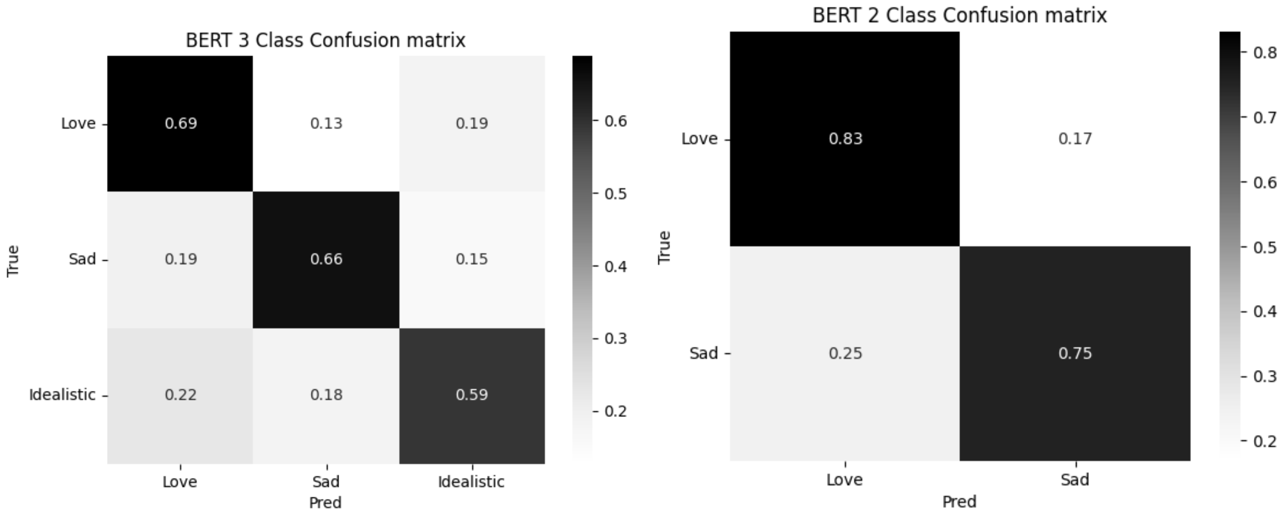
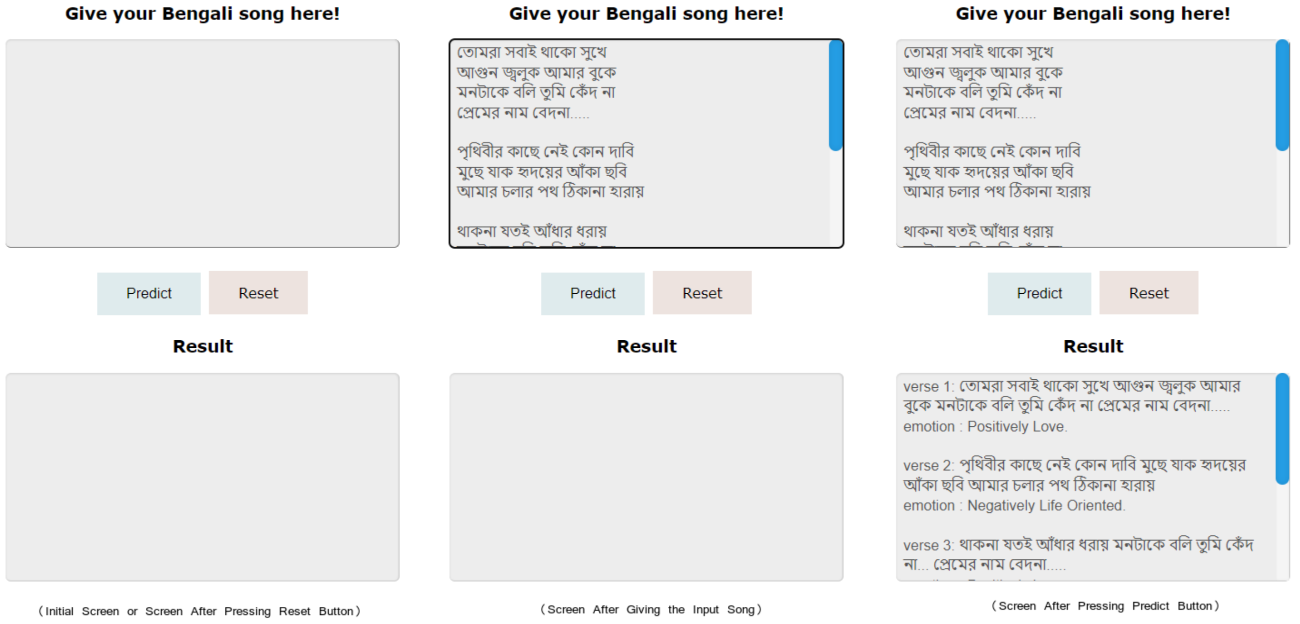


Figure 5. Confusion Matrix HeatMap for both Cases

The overall performances for test data are given in Table IV. The performance data shows a steady consistency in each measurement metric which indicates the evaluated results are well coherent with each other. As we expected, the NN model performed better than the ML models although the performances of MLP and BLSTM models were relatively poor compared with other classifiers. The BERT model achieved the best accuracy of 65% in three classes and 80% in polarity classification. Amongst the ML model, the performance of the SGD classifier was apparent. It gained a close accuracy of 64% as BERT in 3-class classification. More significantly the performances of the ML classifiers NB, SVM, and SGD in binary classification were also really close to our best-performed model BERT, and in every case, the accuracy measurements were 78% respectively. However, for 3-class classification XGBOOST and NB model didn't perform well and their accuracies were about 58% and 59%.

The diagonal elements in the confusion matrix represent the number of occurrences for which the predicted label is equal to the true label, whereas off-diagonal elements are those for which the classifier is mislabeled. The greater the diagonal values of the confusion matrix, the more right predictions there are. The confusion matrix shown in Fig. 5 depicts that in the case of binary classification, our models predicted the Love and Sad emotions very well because of the simple expression and meaning of these particular categories. However, when it comes to 3-class classification the Sad and Idealistic emotions were not particularly recognized well enough because, in most of the cases in the sad context of the songs in our dataset, many idealistic references were used to illustrate the meaning of the songs by the songwriters. After all, we got low accuracy measures for both classification schemes because of the noisy dataset. Besides, there were many inter-class entanglements in our dataset, mostly occurring in the Kazi Nazrul Islam's and Rabindranath Tagore's parts. Also, the Idealistic emotion is a vast collection of choice of words and sentiments and that's where our models performed moderately. In short, the limitations of our work are as follow:

- Noisy dataset,
- Limited emotion classes,
- Moderate accuracy in ternary classification.


Give your Bengali song here!
Predict
Reset
Result
Give your Bengali song here!
Predict
Reset
Result
Give your Bengali song here!
Predict
Reset
Result

(Initial Screen or Screen After Pressing Reset Button)

তোমরা সবাই থাকো সুখে
আগুন জ্বলুক আমার বুকে
মনটাকে বলি তুমি কেঁদ না
প্রেমের নাম বেদনা.....

পৃথিবীর কাছে নেই কোন দাবি
মুছে যাক হৃদয়ের আঁকা ছবি
আমার চলার পথ ঠিকানা হারায়

থাকনা যতই আঁধার ধরায়

(Screen After Giving the Input Song)

তোমরা সবাই থাকো সুখে
আগুন জ্বলুক আমার বুকে
মনটাকে বলি তুমি কেঁদ না
প্রেমের নাম বেদনা.....

পৃথিবীর কাছে নেই কোন দাবি
মুছে যাক হৃদয়ের আঁকা ছবি
আমার চলার পথ ঠিকানা হারায়

থাকনা যতই আঁধার ধরায়

verse 1: তোমরা সবাই থাকো সুখে আগুন জ্বলুক আমার বুকে মনটাকে বলি তুমি কেঁদ না প্রেমের নাম বেদনা.....
emotion : Positively Love.

verse 2: পৃথিবীর কাছে নেই কোন দাবি মুছে যাক হৃদয়ের আঁকা ছবি আমার চলার পথ ঠিকানা হারায়
emotion : Negatively Life Oriented.

verse 3: থাকনা যতই আঁধার ধরায় মনটাকে বলি তুমি কেঁদ না... প্রেমের নাম বেদনা.....

(Screen After Pressing Predict Button)

Figure 6. UI Illustration for Prediction

For the UI illustration part, we used the previously saved Pickle file (created from the SGD model as it was easy to integrate and its performance was very close to the performance of BERT) and utilized that for the prediction part. Based on the user input given in the input text-area field on the web page, when a user clicks the Predict button, a result will be generated from the Flask API and it'll be shown in the Result text-area field along with the identified verses. As we introduced two classification modes (binary and ternary), six possible outcomes could be generated which are Negatively Sad, Negatively Loved, Negatively Idealistic (Life Oriented), Positively Sad, Positively Loved, and Positively Idealistic as depicted in Fig 6.

6. CONCLUSION

We have operated our research in an advocative way and we needed to create a whole new dataset that has some precise and sophisticated emotion labels on it. Although we achieved an accuracy of 65% in multi-class classification and 80% in binary emotion classification, and as the study is relatively new on the subject matter, our proposed approach was completely empirical and there are numerous points where we will upgrade in the upcoming future. In this study, we have attempted to determine whether or not it is possible to determine a song's impressions only from its lyrics. As Bengali has a perplexing writing and colloquial structure, the task of emotion segmentation was an intensive one. Yet, we have analyzed the lyrical pattern of several pioneering works of legendary poets and lyricists and classified their emotions accordingly. Our Future experiments will involve running the same test on a larger variety of textual features. Also, we will increase the emotion classes for a better human-like experience and apply it in real-life digital systems. We intend to create a hybrid mood classification system based on audio and lyric data later on. Using multi-level categorization, we also want to increase the lyrics mood classification system's realistic features. Moreover, we will try to implement explainable AI (Artificial Intelligence) to provide trustworthiness for automatic prediction.

References

[1] G. Shibata, R. Nishikimi, E. Nakamura, and K. Yoshii, "Statistical Music Structure Analysis Based on a Homogeneity-, Repetitiveness-, and Regularity- Aware Hierarchical Hidden Semi-Markov Model," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*. Delft, The Netherlands: ISMIR, Nov. 2019, pp. 268–275.

[2] A. Maezawa, "Music boundary detection based on a hybrid deep model of novelty, homogeneity, repetition and duration," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 206–210.

[3] G. Sargent, F. Bimbot, and E. Vincent, "Estimating the structural segmentation of popular music pieces under regularity constraints," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 344–358, 2017.

[4] B. Liu, "Sentiment analysis and opinion mining," in *Synthesis Lectures on Human Language Technologies*, vol. 5, 05 2012.

[5] A. J. Lonsdale and A. C. North, "Why do we listen to music? a uses and gratifications analysis," *British Journal of Psychology*, vol. 102, no. 1, pp. 108–134, 2011. [Online]. Available: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1348/000712610X506831

[6] J. Robinson and R. Hatten, “Emotions in music,” *Music Theory Spectrum*, vol. 34, pp. 71–106, 10 2012.

[7] X. Chen and T. Y. Tang, “Combining content and sentiment analysis on lyrics for a lightweight emotion-aware chinese song recommendation system,” in *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, ser. ICMLC 2018. New York, NY, USA: Association for Computing Machinery, 2018, p. 85–89.

[8] K. Napier and L. Shamir, “Quantitative Sentiment Analysis of Lyrics in Popular Music,” *Journal of Popular Music Studies*, vol. 30, no. 4, pp. 161–176, 12 2018. [Online]. Available: https://doi.org/10.1525/jpms.2018.300411

[9] 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*. Trivandrum, India: NLP Association of India, Dec. 2015, pp. 261–267.

[10] H. Elfaik and E. H. Nfaoui, “Deep bidirectional lstm network learning-based sentiment analysis for arabic text,” *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 395–412, 2021. [Online]. Available: https://doi.org/10.1515/jisys-2020-0021

[11] 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)*, 2020, pp. 39–44.

[12] D. Nath and S. Phani, “Mood analysis of bengali songs using deep neural networks,” in *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*, M. S. Kaiser, J. Xie, and V. S. Rathore, Eds. Singapore: Springer Nature Singapore, 2021, pp. 1103–1113.

[13] N. P. Urmi, N. U. Ahmed, M. H. R. Sifat, S. Islam, and A. S. M. M. Jameel, “Banglamusicmood: A music mood classifier from bangla music lyrics,” in *International Conference on Mobile Computing and Sustainable Informatics*, J. S. Raj, Ed. Cham: Springer International Publishing, 2021, pp. 673–681.

[14] A. A. Marouf and R. Hossian, “Lyricist identification using stylometric features utilizing banglamusicstylo dataset,” in *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2019, pp. 1–4.

[15] N. Irtiza Tripto and M. Eunus Ali, “Detecting multilabel sentiment and emotions from bangla youtube comments,” in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018, pp. 1–6.

[16] S. Ahmed, M. H. K. Mehedi, M. A. Rahman, and J. B. Sayed, “Bangla music lyrics classification,” in *Proceedings of the 2022 8th International Conference on Computer Technology Applications*, ser. ICCTA ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 142–147.

[17] R. Hossain and A. Al Marouf, “Banglamusicstylo: A stylometric dataset of bangla music lyrics,” in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018, pp. 1–5.

[18] S. Swaminathan and E. G. Schellenberg, “Current emotion research in music psychology,” *Emotion Review*, vol. 7, no. 2, pp. 189–197, 2015. [Online]. Available: https://doi.org/10.1177/1754073914558282

[19] K. Watanabe and M. Goto, “A chorus-section detection method for lyrics text,” in *Proceedings of the 21st International Society for Music Information Retrieval Conference*. Montreal, Canada: ISMIR, Oct. 2020, pp. 351–359.

[20] K. Trohidis, G. Tsoumakas, G. M. Kalliris, and I. P. Vlahavas, “Multi-label classification of music into emotions,” in *International Society for Music Information Retrieval Conference*, 2008.

[21] R. Kumari and S. K. Srivastava, “Machine learning: A review on binary classification,” *International Journal of Computer Applications*, vol. 160, no. 7, pp. 11–15, Feb 2017. [Online]. Available: http://www.ijcaonline.org/archives/volume160/number7/27084-2017913083

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] M. R. Mahmud, M. Afrin, M. A. Razzaque, E. Miller, and J. Iwashige, “A rule based bengali stemmer,” in *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2014, pp. 2750–2756.

[24] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543.

[25] S. Sarker, “BNLP: natural language processing toolkit for bengali language,” *CoRR*, vol. abs/2102.00405, 2021. [Online]. Available: https://arxiv.org/abs/2102.00405

[26] K. S. Babulal and A. K. Das, “Deep learning-based object detection: An investigation,” in *Futuristic Trends in Networks and Computing Technologies*, P. K. Singh, S. T. Wierzchoń, J. K. Chhabra, and S. Tanwar, Eds. Singapore: Springer Nature Singapore, 2022, pp. 697–711.

[27] P. Kumar, K. S. Babulal, D. Mahto, and Z. Khurshid, “Analyzing deep neural network algorithms for recognition of emotions using textual data,” in *Key Digital Trends Shaping the Future of Information and Management Science*, L. Garg, D. S. Sisodia, N. Kesswani, J. G. Vella, I. Brigui, S. Misra, and D. Singh, Eds. Cham: Springer International Publishing, 2023, pp. 60–70.

[28] Q. I. Mahmud, N. I. Chowdhury, and M. Masum, “A multi layer perceptron along with memory efficient feature extraction approach for bengali document categorization,” *Journal of Computer Science*, vol. 16, no. 3, pp. 378–390, Mar 2020. [Online]. Available: https://thescipub.com/abstract/jcssp.2020.378.390

[29] S. Sarker, “Banglabert: Bengali mask language model for bengali language understading,” 2020. [Online]. Available: https://github.com/sagorbrur/bangla-bert

[30] A. Bhattacharjee, T. Hasan, K. Samin, M. S. Rahman, A. Iqbal, and R. Shahriyar, “Banglabert: Combating embedding barrier for low-resource language understanding,” *CoRR*, vol. abs/2101.00204, 2021. [Online]. Available: https://arxiv.org/abs/2101.00204

Maraz Mia graduated from Shahjalal University of Science and Technology in 2023 with B.Sc.(Eng.) in Software Engineering major. His research interest comprises the study of Machine Learning (ML), Neural Networks (NN), Natural Language Processing (NLP), and the implementation of Artificial Intelligence (AI) in Cyber Security.

Pulock Das is currently working as a Software Engineer at TallyKhata. He graduated from Shahjalal University of Science and Technology in 2023 with B.Sc.(Eng.) in Software Engineering major. He is aspiring to become a proficient researcher and practitioner in areas based on Machine Learning (ML), Deep Learning (DL), and Artificial Intelligence.

Ahsan Habib is actively working as an Associate Professor at the Institute of Information and Communication Technology of Shahjalal University of Science and Technology. Before joining SUST, he also worked at Metropolitan University and North East University Bangladesh during 2008-2018. He received his B.Sc.(Engg.) degree in Computer Science and Engineering from Shahjalal University of Science and Technology (SUST), Sylhet, Bangladesh, in 2004. In 2012, he completed his MSc. (Engg.) degree in Information and Communication Technology from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, and received his Ph.D. degree in Computer Science and Engineering from Shahjalal University of Science and Technology, Sylhet, Bangladesh, in 2019. His research interests include Data Compression, Natural Language Processing (NLP), Neural Networks (NN), Machine Learning (ML), and Cloud Computing.