Classification of Bangla Songs using NLP and Transformer Models

Khalid Shifullah , MD. Sabbir Hossain , MD. Humaion Kabir Mahedi , and Annajiat Alim Rasel

Department of Computer Science and Engineering
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{*khalid.shifullah, md.sabbir.hossain1 humaion.kabir.mehedi*}@g.bracu.ac.bd
{*annajiat*}@gmail.com

***Abstract*—In this world which evolves every day based on the inclusion of numerous data, finding ways to make it easier for the new generation to avail themselves with knowledge about their historical glory on basis of individual cultural heritage is of utmost necessity. Talking of Bengali music, it is rich in diversity, and in course of time, many of its forebearers may even make mistakes in passing on the authentic idea about its genre on basis of the origin of a song. This paper discusses the issue of effectively utilizing song lyrics while indexing databases containing information regarding the genre of that particular song in question. Automatic song classification experiments have been demonstrated in this study by taking an initial dataset of 4015 lyrics which have a possible categorization of more than 21 genres along with the uncategorized ones under the unidentified section. Two transformer models: Bangla-BERT and XLM-RoBERTa have been applied in the process.**

***Index Terms*—transformers, XLM-RoBERTa-base, Bangla-BERT-base, text classification**

I. Introduction

Bengali is the fifth most common native language and the seventh most common language overall in the world according to the number of speakers. In terms of usage, Bengali is the seventh most common Indo-European language. Of the 22 scheduled languages of India, it is the second most extensively spoken after Hindi. It is the official, national, and most widely spoken language of Bangladesh. Hence it is evident that a wide number of population may prefer songs in their native language or only would listen to and would want to know more about songs written in a language they use at the very least. Technological advancements has made people of this era incline more towards online searching for information to a query than books or practical knowledge regarding the matter. Smartphone in hands of people are considered the first hand information source than any other thing. People already use it to listen to songs through many renowned music apps such as Spotify, Apple music, Pandora, SoundCloud Go, Google Play Music etc. but hardly any of those may provide with the exact idea about the genre of any literary song, most specifically the Bengali ones due to lack of resources. Early Bengali music was influenced by Sanskrit chants and developed under the influence of Vaishnav poetry. When the musical heritage was developed throughout the middle Ages under the patronage of Sultans and Nawabs as well as the strong landowners and baro bhuiyans, there was a fusion of Hindu and Islamic styles. Then with course of time came songs of different categories among which few worth mentioning are Baul, Nazrul Geeti, Robindro Sangeet, Bhaoaiya, Bhatiali, Lalon Geeti, Songs of Hason Raja, Patriotic songs, Ghazal and so on [3]. The very recent genre of songs are classified into 'Adhunik' or modern songs. Access to knowing these categories will empower any individual in terms of his very own cultural inheritance. Classification of Bangla songs using Bangla lyrics in order to predict the genre can lead to create more efficient recommender system. For this study, two different transformer model will be used, one is XLM-RoBERTa-base [4] and Bangla-BERT-base [8]. Efficiency and effectiveness of the transformer models for this task will be analyzed.

II. Related Works

There is not much existing work on classifying the category of Bengali songs using their lyrics. Although, several works on classifying Bangla texts over several domains using transformer models. One particular study on Bangla text classification [7], which worked on detecting coarse-grained technical domains like computer science, physics, chemistry, etc. has been done using Bangla-Bert-base of a given text. The study has used a dataset of 53,574 texts with domains of Biochemistry, Communication Technology, Computer Science, Management, and physics. After data cleaning, two different BERT transformer encoders, BERT-base and BERT-large have been used. The study compared their work of the pre-trained Bangla language model Bangla-BERT-base with Bengali Electra, Indic Transformers Bangla BERT, and Multilingual BERT [5]. The study produced an accuracy of 84 percent in classifying coarse-grained technical domains.
A different study on Bangla text classification [2], has used two transformer classes like XLM-Roberta, and multilingual BERT. Six different datasets of four different domains like YouTube comments, comments on news articles, news classification, and authorship attribution have been used. AOA (Adam optimization algorithm) and Cross entropy loss criterion algorithm were used to train the models. The model training parameters were a learning rate of 1e-5 with 10 epochs.

The study shows its maximum accuracy of 93 percent in the news classification domain. As for the YouTube comment datasets where sentiment analysis and emotion detection was performed, support vector machine and Naïve Bayes classifiers were used with IF-IDF embedding technique and n-gram tokens. Datasets have been split into 80 percent training set and 20 percent testing set.

III. Dataset

For this study, the dataset has been collected from the Kaggle website. The dataset contains three headers which are Title, Category, and Lyrics and a total of 4105 song lyrics from 21 different categories. Figure-1 shows the skewness of the data with considering the categories or labels.

TABLE I
Dataset Skewness

Category	Number of Lyrics
Modern Music	1042
Rabindra Sangeet	777
Film	761
Band	710
Nazrul Sangeet	203
Miscellaneous	127
Patriotic Song	91
Baul Song	60
Life Song	58
Devotional Song	58
Pop	57
Lalon	45
Folk	43
Drama	18
Islamic	13
Rap	11
Hason Raza	9
Uncategorized	8
Rhyming Song	7
Bhatiali	5
Bhavaya Song	2

From Fig. 1, it is visible that only six categories have more than 100 data, and rest 15 categories have less than 100 data. Merely, 4 categories contain more than 500 data. It is safe to say that, the dataset is heavily skewed. Furthermore, lyrics contain stopwords and special characters which will be excluded in the preprocessing section. For this study, two of the three headers will be used which are category and lyrics.

IV. Methodology

For this study, we will be using two hugging face [1] transformer model which are XLM-RoBERTa base and Bangla-Bert-base. These named hugging face transformers are used to classification and detection purposes. Moreover, accuracy and effectiveness of these transformers will we analyzed. Fig. 2 shows the workflow diagram of this study.

A. Preprocessing

- Data Balancing: The dataset which contains same number of data for each label or category is known as a balanced dataset. In order to perform model with higher accuracy and balanced rate of detection, the dataset should be balanced. Techniques like oversampling, under sampling, class weight are used to perform balancing. As for this study, we will perform undersampling. Dataset used is heavily skewed and particularly small. For undersampling, we will be removing categories that contains less than 10 data for effectiveness of the classifier.
- Duplicates Removal: For the integrity of the classifier all the duplicates data will be moved. If duplicate data exits on dataset is possible that the fitted model will be biased.
- Special Characters Removal: For data cleaning it is important to remove special characters as they do not carry any sentiments or useful for the classification model. Many word embedding techniques may not support these special characters. Hence, special characters from the dataset will be removed.
- Special Characters Removal: For data cleaning it is important to remove special characters as they do not carry any sentiments or useful for the classification model. Many word embedding techniques may not support these special characters. Hence, special characters from the dataset will be removed.
- Stopwords Removal: Frequently used word of any language are called stops words and usually they do not carry any particular sentiments. Most of the times it is efficient to remove the stopwords from the dataset. As for this study we will be removing a list of 398 stopwords of Bengali language [6].

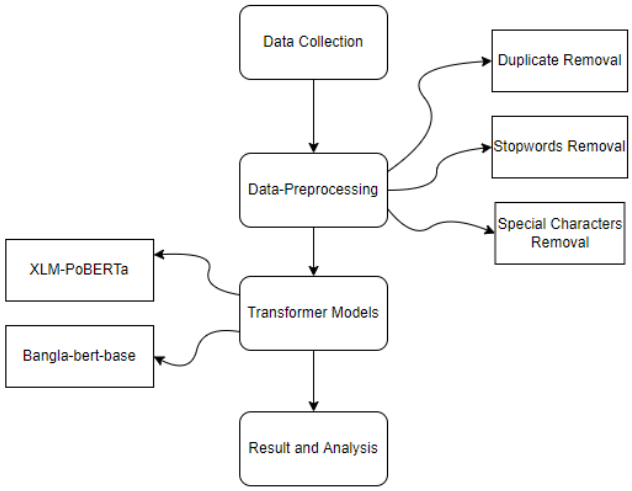

Fig. 1. Work-Flow Diagram

B. Transformer Models

a) XLM-RoBERTa-Base: XLM-RoBERTa is a multilingual version of the transformer that originated from RoBERTa. 100 languages from 2.5TB of filtered CommonCrawl records are used as its pre-training material. Transformer model RoBERTa is a pre-trained model trained through a vast corpus applying an unsupervised scheme. That is, it has been trained

in a process that involves texts only and excluding human labels or inputs to create inputs and outputs from the texts. Because of this, a lot of publicly available data can be utilized. It was pre-trained using the Masked language modeling (MLM) aim, to be more precise. When given a sentence, the model randomly selects almost fifteen percent input to be masked, after which it must predict the words that have been masked. This contrasts with conventional RNNs (recurrent neural networks), that sequentially perceive the words, autoregressive models such as GPT, which mask the next characters internally. Thus it is made possible to learn a two-way statement representation by the model. A standard classifier can be trained using the features generated by the XLM-RoBERTa model through the use of labeled sentences, for instance. In such a manner, the model acquires an internal representation of 100 languages from among which, features helpful for subsequent tasks can be extracted. The masked language modeling raw model is primarily designed to be refined on a downstream task. For tasks like sequence classification, character classification, or question answering where the complete sentence is used to reach a decision, this model is primarily meant to be improved.

b) Bangla-Bert-base: Further pre-trained transformer model for the Bengali language is the Bangla BERT Base which was trained using only the raw texts. OSCAR1 derived Bengali CommonCrawl corpus as well as Bengali Wikipedia Dump 2 dataset were used to pre-train this BERT-based model [6]. Human labels or inputs were not used to create inputs or outputs from the texts. Using mask language modeling (MLM) this model was made to undergo pre-training to achieve more precision.

C. Training parameters

Both of the transformers learning rate is set to 2e-5 and 10 epochs. The dataset has been split in 8:2 ratio. Where 80 percent were used for training set and 20 was used to testing set.

V. Result and Analysis

Both transformers XLM-RoBERTa-Base and Bangla-Bert-base displayed decent accuracy. XLM-RoBERTa-base produced 73.03 percent accuracy with an f1-score of 71 percent. However, Bangla-Bert-base produced 77.56 percent accuracy with an f1-score of 77.56 percent. F1 score, accuracy along with other performance matrices of these two transformers age shown in table 2.

TABLE II
Performance Score

Transformer	Accuracy	F1-score	Precision	Recall
XLM-RoBERTa	73.03	71	73	70
Bangla-BERT-Base	77.46	77	78	77

Moreover, Fig 3, 4 shows the confusion matrix of XLM-RoBERTa-base and Bangla-BERT-base. Each row and column of the matrix implies each category or label. Matrix values with equal row and column shows the number of

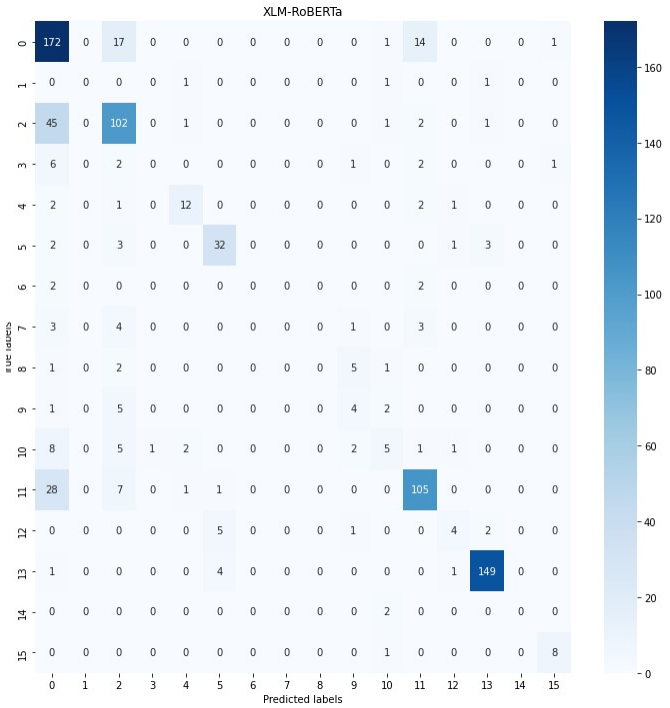

Fig. 2. Confusion Matrix XLM-RoBERTa

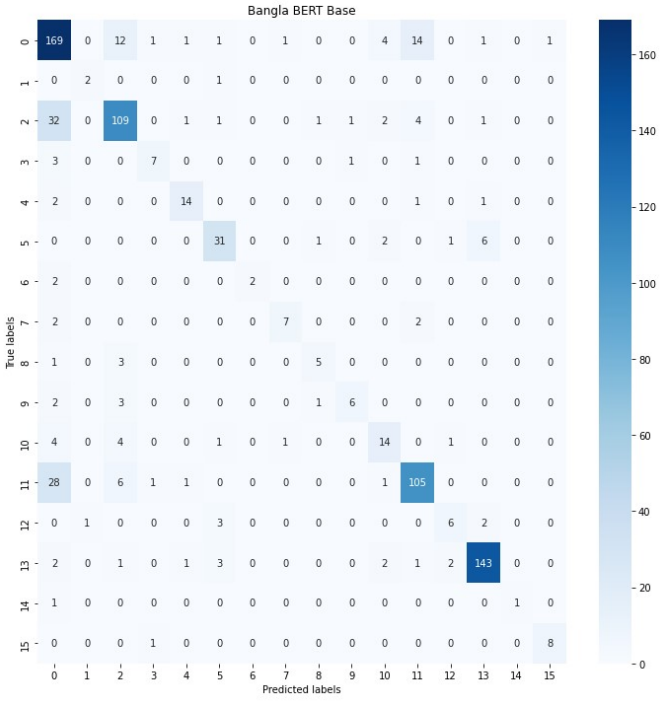

Fig. 3. Confusion Matrix Bangla-BERT

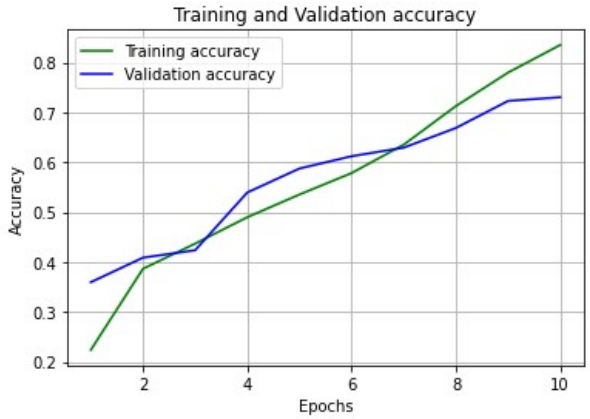

Fig. 4. Accuracy vs Epochs (XLM-RoBERTa)

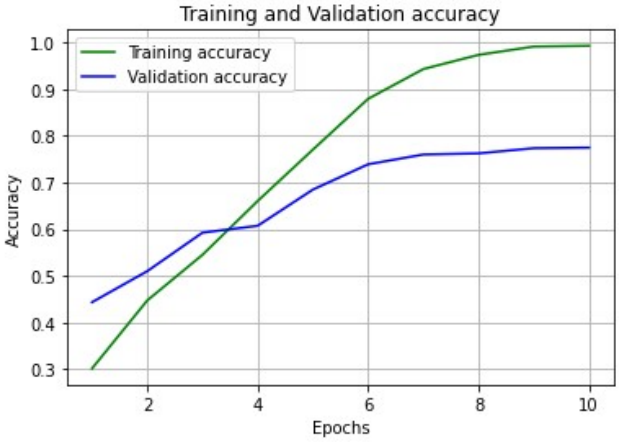

Fig. 5. Accuracy vs Epochs (Bangla-BERT)

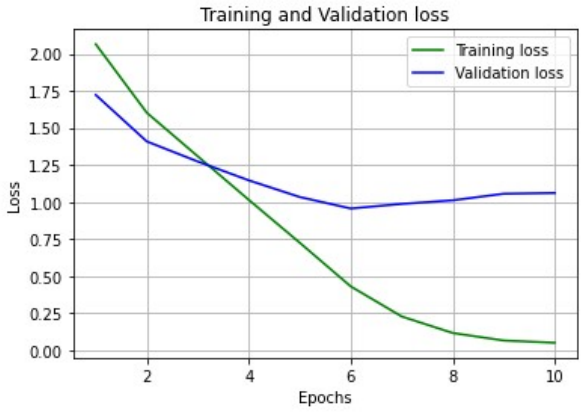

Fig. 6. Loss vs Epochs ((Bangla-BERT)

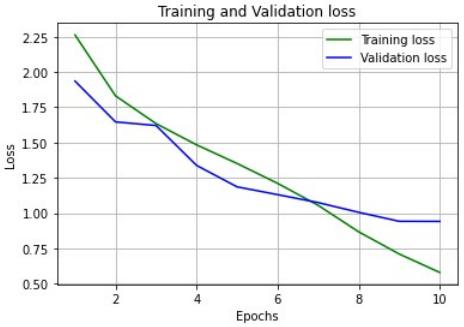

Fig. 7. Loss vs Epochs ((XLM-RoBERTa))

correctly predicted data of particular category. It shows that, categories with large number of data shows higher accuracy. Plot between Accuracy vs Epochs shown in figure 5,8 for XLM-RoBERTa and figure 6, 7 for Bangla-BERT-Base. For a balanced dataset and transformer lines training accuracy and validation accuracy should be as close as possible. As shown in the figures, these lines deviates from each other. For XLM-RoBERTa deviation starts from epoch 7 and for Bangla-BERT the number is epoch 4. The database has been used was small and not balanced. Moreover, it contained 16 labels or categories after data cleaning which means very few training set for each category. With bigger and more balanced dataset will make this model more efficient and effective.

VI. Conclusion

This study is to showcase the procedure of automated generation of genre of songs just by taking a single lyric as input. There aren't many studies that categorize Bengali songs currently. However, there have been numerous attempts to categorize Bangla texts across various domains using transformer models. Both transformer used in this study showed almost similar accuracy and performance score. However, Bangla-Bert-base were relatively more successful than XLM-RoBERTa-base. With a balanced and larger dataset these transformers will be able to provide greater performance. Information and technology being an open space for exploration, it makes this study open as well for further approach in research for same cause or more.

References

[1] Hugging face, the ai community building the future.
[2] Tanvirul Alam, Akib Khan, and Firoj Alam. Bangla text classification using transformers. *arXiv preprint arXiv:2011.04446*, 2020.
[3] Abhijit Bhowmik and AZM Chowdhury. Genre of bangla music: A machine classification learning approach. *AIUB Journal of Science and Engineering (AJSE)*, 18:66–72, 08 2019.
[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
[6] Gene Diaz. Stopwords bengali (bn), 2016.

[7] Koyel Ghosh and Apurbalal Senapati. Technical domain classification of bangla text using bert. *Biochemistry (bioche)*, 2:741, 2021.
[8] Sagor Sarker. Banglabert: Bengali mask language model for bengali language understading, 2020.