Bengali Social Media Post Sentiment Analysis using Deep Learning and BERT Model

Samsul Islam
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
samsulratul98@gmail.com

Md. Jahidul Islam
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
jahid.aust39@gmail.com

Md. Mahadi Hasan
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
mahadiaustcse39@gmail.com

S. M. Shahnewaz Mahmud Ayon
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
ayonmahmud53@gmail.com

Ms. Syeda Shabnam Hasan
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
shabnam.cse@aust.edu

2022 IEEE Symposium on Industrial Electronics & Applications (ISIEA) | 978-1-6654-8012-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIEA54517.2022.9873680

***Abstract*— Social media platforms such as Facebook, Twitter, and others are becoming incredibly popular for expressing sentiments and thoughts. People use these platforms to express not only their happy moments, but also their feelings when they are depressed. Using sentiment analysis in natural language processing to analyse these social media posts, one's emotional state can be determined, such as happy, sad, or angry at a particular time. The majority of research in this topic is conducted in English, therefore sentiment analysis from Bengali is not very accurate. So, our goal is to work on this topic using Bengali datasets obtained from various social media posts to improve sentiment detection accuracy. This work can be used to help building a system in our country's mental health sector. In this research, we first gathered social media data. Then we used a number of feature selection and extraction techniques like Word2Vec, GloVe etc. and applied a number of deep learning model, such as RNN, LSTM, GRU etc. We have also applied hybrid and transformer-based BERT models like CNN-BiLSTM Bangla-BERT, mBERT etc and finally got the highest accuracy of 88.59% for the CNN-BiLSTM hybrid model using the GloVe feature vector.**

Keywords*— *Sentiment Analysis, Social Media, Deep Learning Algorithms, Hybrid model, BERT model, Transformer Based Model

I. INTRODUCTION

Sentiment refers to a person's feelings and ideas regarding something or someone that are held or expressed. It is an essential part of everyday human interactions. People often convey their emotions by various means such as facial expression, verbal discourse, written language, and so on. Because of accessible internet access, the habit of sharing personal thoughts and emotions on social media has grown significantly in recent years. People from all around the world can now communicate with one another and share their thoughts and feelings on any political or global topic.

Currently, more than 66.44 million Bangladeshis use various forms of social media [1]. Every year, this figure fluctuates substantially. As a result, these platforms have become a convenient way for people to contact others and share their feelings and thoughts. In fact, people nowadays share their emotions on social media more than they do in person. We can learn about a person's ups and downs in life by checking his or her social media posts on a regular basis. Suicidal attempts among teenagers, on the other hand, have grown recently. We can see how depressed he/she was when checking their social media posts after learning that he/she is no longer alive. All of these social activities have grown into a massive resource that may be analysed to determine human sentiment. Popular social media platforms such as Twitter and Facebook have created a good means for their users to communicate their opinions, thoughts, reviews, and feedback on something through text, which can be used to assess sentiment of a specific user from anywhere in the world. This can be extremely beneficial in any significant issue or crisis in their personal life or in other fields such as market analysis, product reviews, and so on. Furthermore, recent works on Bengali sentiment analysis have primarily focused on binary sentiment analysis, in which happy emotions are assigned as positive value and sad emotions are assigned as negative value. However, this does not reflect the entire picture because people do not simply share their happy or sad thoughts on social media. Sometimes they express their rage, and other times their posts are neutral. As a result, distinguishing each emotion in a text is important for understanding the human mind.

For our research purpose, we have collected social media posts from Facebook and Twitter. After applying different data preprocessing, feature selection and extraction techniques, we have applied a number of deep learning, hybrid and BERT models and found out the best model according to their sentiment detection accuracy.

II. RELATED WORKS

We classify sentiment analysis into three distinct categories: happy, sad, and angry. Our purpose is to look at how people feel on social media networks. Despite the fact that there are a number of research publications on sentiment analysis in Bengali, the accuracy needs to be improved. As a result, we have chosen to work in this field because there is a considerable opportunity to increase accuracy.

Ma, Long, and Yan Wang [2] showed a semantic graph with depression symptoms from Twitter dataset where they collected almost 120 thousand tweets of 140 characters from the entire Twitter community for 15 days using the "Depression" keyword and prepared about 64 thousand distinct tweets in the data preprocessing step. Neural Network Language Model (NNLM), Artificial Neural Network (ANN), and Rapid Automatic Keyword Extraction (RAKE) were used to develop their own hybrid method called Automatic Extract Keyword for Specific Terms

(AEKW). They first found out the co-occurrence words of "depression" and scored them using the RAKE algorithm to build a semantic graph with the centre vertex as "depression" and the other words' weights represented as the importance score between two vertices.

A number of machine learning models were applied [3] to analyse sentiment on Bengali news comments where they took data from Prothom-Alo's user comments to identify different emotions. Three classifiers such as Support Vector Machine (SVM), LSTM and CNN were used for the final approach and got 67.48% accuracy from SVM, 78% accuracy from LSTM and 63% accuracy from CNN.

N. R. Bhowmik [4] proposed an extended lexicon dictionary for sentiment analysis approach using supervised machine learning where they collected data from Bengali absa datasets [5]. There are two datasets based on two domains which were used to build up their extended sentimental dictionary. They have proposed their own model named BTSC which has a total of 30 steps and achieved the result of 82.21% accuracy in cricket dataset and 80.58% accuracy in restaurant dataset on BiGram feature matrix with the proposed method.

A sentiment analysis of Bengali language using Deep learning approaches has been proposed [6] on the dataset collected from three Facebook groups by using Facebook graph API where they have used two classifiers such as CNN and LSTM. Hybrid model CNN BiLSTM was also used for the final approach. They got 84.00% accuracy from SVM, 66.67% accuracy from LSTM and 90.49% accuracy from CNN BiLSTM.

S. Sharmin and D. Chakma [7] showed an approach to analyse sentiment of Bengali language using convolutional neural network where they have collected data from social media such as BBC Bangla and Prothom Alo. They have used three classifiers such as LR, CNN and LSTM. They got the highest 76.25% accuracy from CNN.

M. R. Karim [8] proposed to detect Explainable Hate Speech on Under-resourced Bengali Language extending dataset [9] with additional 3,000 labelled samples. They have used 4 types of BERT-variants such as Bangla BERT, mBERT cased, XLM-RoBERTa, mBERT uncased and ensemble all of them and got the highest accuracy of 88% from ensembling..

A sentiment analysis approach was proposed to classify emotion in a resource-constrained language using a transformer-based model [10] where they developed a corpus called BEmoC to classify emotion in Bengali text. They have used 3 types of BERT-variants such as Bangla BERT, mBERT, and XLM-RoBERTa and got the highest accuracy of 70.11% from XLM-RoBERTa.

M. A. Hasan and J. Tajrin [11] classified Sentiment in Bengali Textual Content exploring several publicly available dataset such as Sentiment Analysis in Indian Languages (SAIL) Dataset, ABSA Dataset, BengFastText Dataset, YouTube Comments Dataset, Social Media Posts (CogniSenti Dataset) and designed classifiers using both classical and deep learning algorithms. Finally, they compared the results of these dataset and got the highest accuracy of 72.9% for the YouTube Comments dataset from XLM-RoBERTa-large.

M. R. Karim [9] prepared three public datasets of hate speech, sentiment analysis, and document classification of Bengali texts, which are larger than currently available Bengali datasets by both quantity and subject coverage. They proposed an approach to classify benchmarks for under-resourced Bengali language based on multichannel Convolutional-LSTM Network and got the highest 74.6% accuracy from MConv-LSTM..

Transformer based deep neural network [12] was applied to classify Sentiment in Bengali Textual Content using four public datasets including YouTube comment datasets, News comment sentiment dataset, Authorship Attribution dataset and News Classification dataset. They got the highest accuracy of 93.8% for the Authorship attribution dataset from XLM-RoBERTa-large.

III. DATASET PREPARATION

We have collected about 4000 different types of social media posts from Facebook and Twitter. A number of data preprocessing techniques have to be used as this raw data contains many unnecessary texts and symbols. We've also gone through various datasets and gathered some related information from them. Though these datasets are not directly linked to the topic we are working on, they are connected in some way. We gathered approximately 6000 data from various sources and incorporated them into our dataset. To train our model, we collected 10,000 data in total. We have tried to avoid duplicate data. In our dataset, we have three sentiments that we tried to keep equal in number to make it well balanced.

We have also collected a benchmark dataset [6] for evaluating our model.

IV. METHODOLOGY

Fig. 1, shows the work direction of our model. We have focused on training different deep learning, a hybrid and few transformers based BERT models. A number of different hyperparameters were set to find better accuracy from the models.

CNN BiLSTM was used with the help of keras. For this firstly sequential layer, embedding layer and convolution layer 1d (ReLU activation function) were used stepwise. Maxpool 1d was also used to flattened the input size and then dropout layer. In order to optimize, the Adam optimizer is applied. The loss function is the cross entropy function, and the activation function is the sigmoid function. In purpose of fitted the training data batch size 32 and epoch 30 with validation was utilized.

BERT is a transformers based model. A transformer is a deep learning model that uses the self-attention mechanism to weight the importance of each element of the input data differently. Its primary applications are in natural language processing (NLP) and computer vision (CV). BERT is a pre-training language representation method which means a general -purpose "language understanding" model is trained on a large text corpus like Wikipedia and then this model is used downstream NLP tasks like question answer, sentiment analysis etc.

BERT is pre trained using this bidirectional capability on two NLP Tasks.

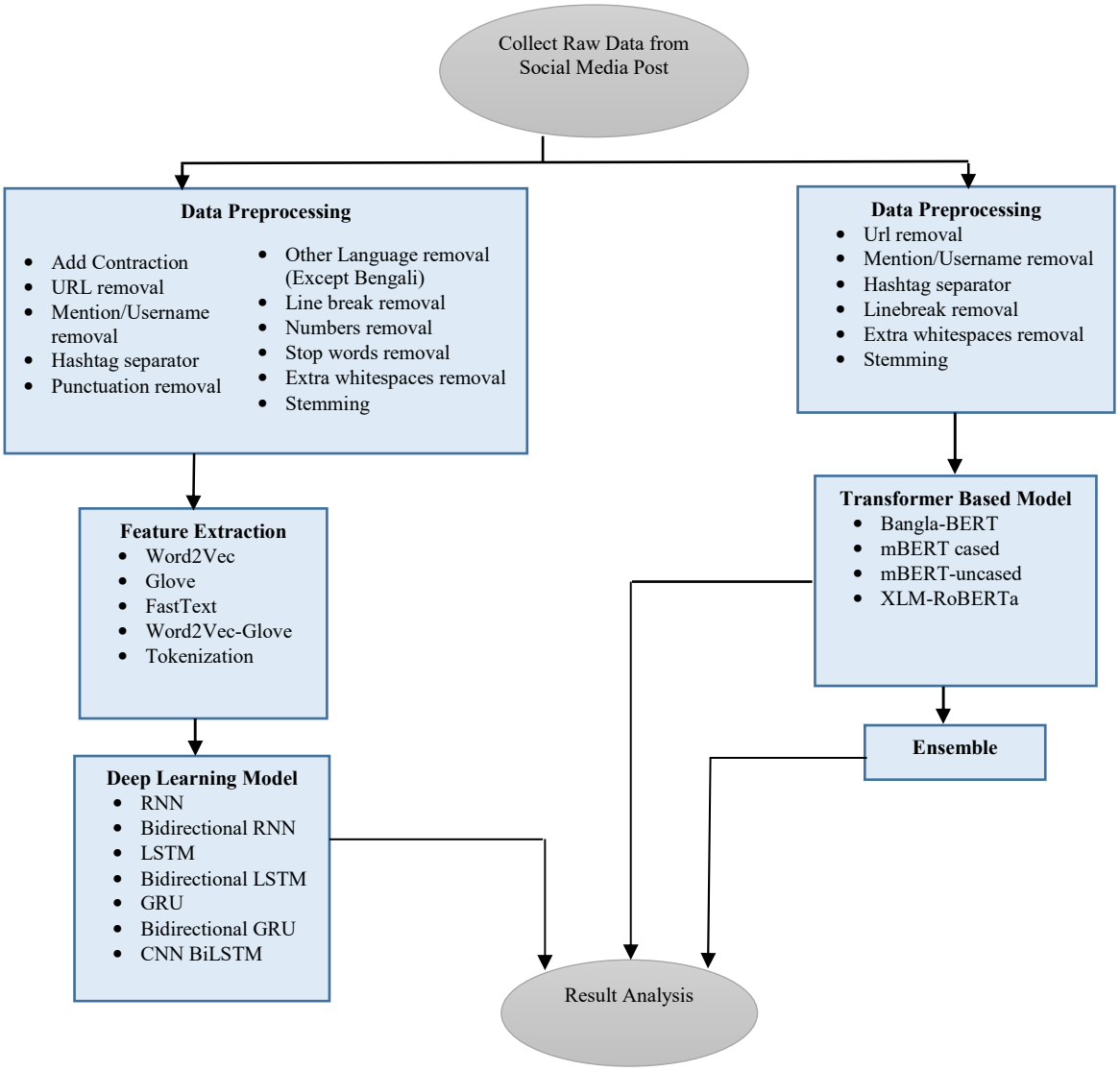

Fig. 1. Model Diagram

Masked Language Model (MLM) training hides a word in a sentence and lets the program predict the hidden (masked) word based on that word's context. The Next Sentence Prediction Training instructs the machine to determine if two given sentences are logically and sequentially related or have a random relationship.[13]

A. *Dataset Preprocessing*

As raw data was a little messy, it had to be cleaned and preprocessed first. Before extracting features, we preprocessed each post.

For deep learning models, we removed Stop words, added the complete form of contracted words, removed urls, Mention/Username, Hashtag, Punctuation, Other Language(Except Bengali), Line Break, Numbers, and Extra whitespaces from our dataset as part of the preprocessing. To discover the root word, we also used stemming.

For BERT models, major preprocessing tasks were not done except removed urls, Mention/Username, Hashtag, Line Break, Extra whitespaces, stemming for. preprocessing. Because research has shown that BERT-based models perform better classification accuracy on unclean texts.

B. *Neural word embeddings*

As raw data was a little messy, it had to be cleaned and preprocessed first. Before extracting features, we preprocessed each post.

After doing the preprocessing step, we extracted the different features using different techniques in our dataset. We used different embedding techniques to vectorize the data with 100D using the following techniques:

1) **GloVe:** It was used to find the frequent Bengali words that are generally used together and vector representation. We have used [14] pretrained model for GloVe.

2) **FastText:** FastText represents each word as n-gram of characters. FastText was used to break each word into different parts by using n-gram. This is very helpful to analyse sentiments for unknown and rare Bengali words. [14] We used pre trained model to generate vectors.

3) **Tokenizing:** Tokenizing is used for split words from a sentence and generating vector values. We used the tokenizer() method from keras.

4) **Word2Vec:** Word2Vec generates a vector value for each word of the sentence. Skip gram word2vec was used to predict the surrounding words from the central word. [14] We used pre trained model to generate vectors.

5) **Word2Vec-GloVe:** We took Word2Vec and GloVe vectorized values and calculated their average to create a new 100D vectorized value.

C. Training of DNN baseline models

We have used GloVe, FastText, Word2Vec and Word2Vec-GloVe as a feature vector to train our dataset with Deep Learning Models. After the feature extraction RNN, Bidirectional RNN, LSTM, Bidirectional LSTM, GRU, Bidirectional GRU were used to train the model. A hybrid model with CNN-BiLSTM was also used to find better results. Two sets of hyperparameters were taken to find the best result which are given in table I. The hidden neuron number should be between the input layer size and the output layer size. So the number of hidden neurons should be less than twice the size of the input layer. So we used hidden size = 70. We used batch size of 32 as it performs better. Additionally, training takes too long if the learning rate is low. We chose 40 epochs since accuracy stays constant after epoch 40.Training of transformer-based models

After the simple data preprocessing the given hyperparameter was used for transformer-based BERT models. We trained monolingual Bangla BERT-base mBERT(cased and uncased), XLM-RoBERTa small models, Bangla-BERT base [15] and performed the ensemble of these models with different hyperparameter combinations shows in table II.

TABLE I. HYPERPARAMETER SET FOR DNN BASELINE MODEL

	Setup 1	**Setup 2**
Input size	100	100
Hidden layer size	70	70
Number of layers	1	3
Number of classes	3	3
Number of epochs	40	40
Learning rate	0.001	0.001

TABLE II. HYPERPARAMETER COMBINATIONS FOR TRAINING BERT VARIANTS

Hyperparameter	**Bangla -BERT**	**mBERT cased**	**mBERT uncased**	**XLM-RoBERTa**
Learning-rate	3e-5	2e-5	5e-5	2e-5
Epochs	6	6	6	6
Max seq length	128	128	128	128
Dropout	0.3	0.3	0.3	0.3
Batch size	16	16	16	16

TABLE III. COMPARISON OF DIFFERENT BERT ALGORITHMS

Classifier	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Bangla-BERT	72.11	72.32	72.11	72.12
mBERT cased	72.7	72.08	72.7	72.59
mBERT uncased	72.41	72.69	72.41	72.68
XLM-RoBERTa	72.57	72.46	72.57	72.59
Ensemble	74.22	74.77	74.22	74.38

V. RESULT ANALYSIS

In this section, results are provided of all algorithms as well as the benchmark dataset performance. For deep learning, hybrid, and transformer-based models, the dataset is split into 80:20 train-test set ratios. Separating the train from the test dataset allows us to assess the performance of different splitting strategies. We compared the outcomes and examined how our defined setups affected the Deep learning algorithms. The best model was tested with a benchmark dataset and the results were compared for quality assurance.

For BERT algorithms, we have got highest accuracy for XLM-RoBERTa. As Ensemble merges the predictions from more than one model, so it gave the highest accuracy which can be seen in Table III.

Table IV shows the accuracy for setup 1 from table I. Here, Bidirectional RNN gave the highest accuracy for GloVe feature. By focusing on the relationships between word pairs rather than words alone, Glove gives word vectors a bit more clear meaning. Also BiRNN retains every piece of knowledge throughout time. Only the ability to remember past inputs makes it helpful for time series prediction.

Table V shows the accuracy for setup 2 from table II. Here, Bidirectional GRU gave the highest accuracy for FastText feature.

We have also trained CNN BiLSTM hybrid model which results are shown on Table VI. Here, we have got the highest accuracy for GloVe feature. GloVe is concerned with word co-occurrences throughout the entire corpus. And for that reason, it produces greater results. Additionally, it makes use of the word vectors to refer to sub-linear connections in vector space. It consequently outperforms Word2vec in the word analogy challenges. CNN BiLSTM learns both character-level and word-level characteristics in the original formulation used for named entity recognition. It solves the fixed sequence-to-sequence prediction problem. A new feature vector is extracted from the per-character feature vectors, such as character embedding and (optionally) character type, for each word by the model using a convolution and a max pooling layer.

As we have prepared our own dataset, we need to check the quality of our research comparing with other renowned papers. Since we got the highest accuracy of 88.59% for CNN BiLSTM model, we checked our model with a benchmark dataset to ensure quality. Table VII shows the result using benchmark dataset which is pretty similar to our own dataset result.

So the overall highest accuracy is 88.59% for CNN BiLSTM with GloVe and Word2Vec-GloVe feature vector.

TABLE IV. COMPARISON OF DIFFERENT DNN ALORITHMS FOR SETUP 1 WITH DIFFERENT FEATURE

Feature		RNN	Bi-RNN	LSTM	Bi-LSTM	GRU	Bi-GRU
GloVe	Accuracy	55.56	66.67	55.6	57.33	55.56	55.6
	Precision	55.62	66.73	55.65	57.22	55.62	55.65
	Recall	55.56	66.67	55.6	57.33	55.56	55.6
	F1-Score	55.78	66.71	55.75	57.48	55.78	55.75
Word2Vec	Accuracy	56.05	57.22	57.33	55.56	57.12	55.46
	Precision	56.81	57.34	57.22	55.62	57.22	55.53
	Recall	56.05	57.22	57.33	55.56	57.33	55.46
	F1-Score	56.70	57.40	57.48	55.78	57.22	55.69
Word2Vec-GloVe	Accuracy	59.15	57.92	59.39	60.32	58.41	59.2
	Precision	59.97	57.75	59.48	60.27	59.41	59.31
	Recall	59.15	57.92	59.39	60.32	58.41	59.2
	F1-Score	59.36	57.81	59.42	60.28	58.58	59.24
FastText	Accuracy	62.08	63.03	63.08	61.89	63.97	61.79
	Precision	61.84	63.36	62.98	62.2	63.73	62.11
	Recall	62.08	63.03	63.08	61.89	63.97	61.79
	F1-Score	61.73	63.15	63.01	61.91	63.63	61.81

TABLE V. COMPARISON OF DIFFERENT DNN ALGORITHMS FOR SETUP 2 WITH DIFFERENT FEATURE

Feature		RNN	Bi-RNN	LSTM	Bi-LSTM	GRU	Bi-GRU
GloVe	Accuracy	55.56	58.67	57.6	59.33	55.76	57.6
	Precision	55.62	64.73	57.65	59.22	55.82	57.65
	Recall	55.56	64.67	57.6	59.33	55.76	57.6
	F1-Score	55.78	64.71	57.75	59.48	55.98	57.75
Word2Vec	Accuracy	56.01	56.4	56.33	53.64	59.12	56.46
	Precision	56.85	57.22	56.22	53.62	59.22	55.53
	Recall	56.01	56.4	56.33	53.64	59.33	56.46
	F1-Score	56.16	57.22	56.48	53.78	59.22	56.69
Word2Vec-GloVe	Accuracy	56.01	56.21	59.19	60.52	60.41	61.2
	Precision	56.85	56.75	59.28	60.47	61.41	61.31
	Recall	56.01	56.21	59.19	60.52	60.41	61.2
	F1-Score	57.12	57.32	59.62	60.48	61.58	61.24
FastText	Accuracy	58.96	60.2	61.89	61.69	61.74	62.83
	Precision	59.57	61.18	62.91	62.16	62.17	64.32
	Recall	58.96	60.2	61.89	61.69	61.74	62.83
	F1-Score	59.12	59.92	61.95	61.48	61.81	63.10

TABLE VI. RESULT OF CNN BILSTM ALGORITHM

GloVe	88.59
Word2Vec	84.25
Word2Vec-GloVe	88.59
FastText	86.25

TABLE VII. RESULT OF BENCHMARK DATASET

Accuracy	85.09
Precision	85.13
Recall	85.09
F1-Score	85.06

VI. CONCLUSION

We have used a number of Deep Learning, Transformers based BERT models and a hybrid model on different hyperparameters to find the best result. Among them, we got the highest accuracy of 88.59% for the CNN-BiLSTM hybrid model using the GloVe and Word2Vec-GloVe feature vector. We have taken an approach to detect sentiment from Bengali text data. The result we have got is better than some of the papers we have included in the literature review section. Our next plan is to enrich the dataset to get better accuracy on advanced models as these are data hungry models. We'll also try to develop an extension and deploy it soon so that everyone can use the extension.

REFERENCES

[1] DIGITAL 2020: BANGLADESH, https://datareportal.com/reports/digital-2020-bangaldesh. Last accessed 20 Jun 2021.

[2] L. Ma and Y. Wang Constructing a semantic graph with depression symptoms extraction from twitter In 16th IEEE International Conference on Computational Intelligence in Bioinformatics and Computational Biology, pp. 1–5, IEEE Siena Tuscany, Italy (2019).

[3] M. A.-U.-Z. Ashik, S. Shovon, and S. Haque, "Data set for sentiment analysis on bengali news comments and its baseline evaluation," in 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1–5, IEEE, 2019.

[4] N. R. Bhowmik, M. Arifuzzaman, M. R. H. Mondal, and M. Islam, "Bangla text sentiment analysis using supervised machine learning with extended lexicon dictionary," Natural Language Processing Research, vol. 1, no. 3-4, pp. 34–45, 2021.

[5] Bangla ABSA Datasets, https://github.com/AtikRahman/Bangla_ABSA_Datasets. Last accessed 25 Jun 2021.

[6] M. Hoq, P. Haque, and M. N. Uddin, "Sentiment analysis of bangla language using deep learning approaches," in International Conference on Computing Science, Communication and Security, pp. 140–151, Springer, 2021.

[7] S. Sharmin and D. Chakma, "Attention-based convolutional neural network for bangla sentiment analysis," AI & SOCIETY, vol. 36, no. 1, pp. 381–396, 2021.

[8] M. R. Karim, S. K. Dey, T. Islam, S. Sarker, M. H. Menon, K. Hossain, M. A. Hossain, and S. Decker, "Deephateexplainer: Explainable hate speech detection in under-resourced bengali language," in 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10, IEEE, 2021.

[9] M. R. Karim, B. R. Chakravarthi, J. P. McCrae, and M. Cochez, "Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network," in 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp. 390–399, IEEE, 2020.

[10] A. Das, O. Sharif, M. M. Hoque, and I. H. Sarker, "Emotion classification in a resource constrained language using transformer-based approach," arXiv preprint arXiv:2104.08613, 2021.

[11] M. A. Hasan, J. Tajrin, S. A. Chowdhury, and F. Alam, "Sentiment classification in bangla textual content: A comparative study," in 2020 23rd International Conference on Computer and Information Technology (ICCIT), pp. 1–6, IEEE, 2020.

[12] T. Alam, A. Khan, and F. Alam, "Bangla text classification using transformers," arXiv preprint arXiv:2011.04446, 2020.

[13] Bert explained: State of the art language model for nlp, https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270. Accessed: 2021-12-23.

[14] Bengali Natural Language Processing(BNLP), https://github.com/sagorbrur/bnlp. Last accessed 20 Jun 2021.

[15] Bangla BERT Base, https://huggingface.co/sagorsarker/bangla-bert-base. Last accessed 20 Jun 2021.