2023-10

Exploring the Efficacy of BERT in Bengali NLP: A Study on Sentiment Analysis and Aspect Detection

Hossain, Md. Junayed

Independent University, Bangladesh (IUB)

Draft Version

Exploring the Efficacy of BERT in Bengali NLP: A Study on Sentiment Analysis and Aspect Detection

Md. Junayed Hossain
Department of CSE
Independent University, Bangladesh
Dhaka, Bangladesh
junayed.ndc16@gmail.com

Sheikh Md. Abdullah
Department of CSE
Independent University, Bangladesh
Dhaka, Bangladesh
s.m.abdullah013@gmail.com

Mohammad Barkatullah
Department of CSC
Independent University, Bangladesh
Dhaka, Bangladesh
barkatopu1234@gmail.com

Md Fahad Monir
Department of CSE
Independent University Bangladesh
Dhaka, Bangladesh
fahad.monir@iub.edu.bd

***Abstract*—The development of accurate sentiment analysis and aspect detection for the Bengali language is crucial due to the rise of Bengali language usage in digital media. Sentiment analysis and aspect detection are essential tasks in Natural Language Processing (NLP) as they allow us to extract meaningful information from textual data. In this study, we explore the performance of advanced NLP techniques in Bengali text classification tasks, specifically sentiment analysis and aspect detection. To achieve this, we compare the performance of the Bidirectional Encoder Representations from Transformers (BERT) model with Bi-LSTM, LSTM, and GRU models. We collect two Bengali datasets and preprocess them to be compatible with the input format required by BERT. The model is then trained and tested on the preprocessed data. Our results show that the BERT model outperforms the traditional Bi-LSTM, LSTM, and GRU models with a 92.5% accuracy in sentiment classification and 90.4% accuracy in aspect detection. The precision, recall, and F1-score values further support the superior performance of BERT. Our study highlights the effectiveness of using advanced NLP techniques such as BERT in text classification tasks for the Bengali language. This opens up new avenues for future work in the field of Bengali NLP, specifically in the area of sentiment analysis and aspect detection.**

***Index Terms*—BERT, sentiment analysis, aspect detection**

I. Introduction

The field of Natural Language Processing (NLP) [1] has seen significant advancements in recent years, with text classification and sentiment analysis being some of its foundational tasks with practical applications [2]. The proliferation of digital content in different languages, including Bengali, has made it imperative to develop accurate models for text classification and sentiment analysis in multiple languages. Bengali, being the sixth most widely spoken language in the world, particularly in Bangladesh and India, has a large population of speakers, but it has received limited attention in NLP research. To address this gap, there is a need for robust models for sentiment analysis and text classification in Bengali. The purpose of this study is to assess the effectiveness of various deep learning and Bidirectional Encoder Representations from Transformers (BERT) [3] models for text classification on a variety of Bengali datasets. To train the model, we need a total of two data sets. Both data sets were taken from Kaggle and are titled "Bengali Ekman's Six Basic Emotions Corpus" [4] and "SentNoB" [5]. We evaluate how well the models categorize texts into various groups. Our objective is to give a thorough assessment of deep learning models for Bengali sentence classification in two ways such as sentiment analysis and aspect classification and also highlight their potential for additional study and advancement in Bengali NLP.

This study aimed to explore the effectiveness of the BERT [3] model and other deep learning models such as Gated Recurrent Unit (GRU) [6], Bi-directional Long Short Term Memory (Bi-LSTM) [7], and Long Short Term Memory (LSTM) [8] for the Bengali text classification. The study specifically focused on the two subtasks of aspect detection and sentiment analysis. To accomplish this, we collected and preprocessed two different Bengali dataset, which was then fed into the various models for two different subtasks. The results were analyzed using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The results showed that all of the models performed well in the subtasks, with BERT and GRU achieving the highest overall scores. However, the study highlights the need for further research to improve the performance of these models and to develop more advanced models for Bengali text classification.

The structure of this study consists of five sections. In section one we already described the introduction part. Section two provides a comprehensive overview of related work in the field. The proposed methodology is outlined in detail in section three. The results and analysis of various methods are presented in section four. Finally, the study is concluded in section five with a summary of the findings and recommendations for future work.

II. Related Work

Recently people's interest in sentimental analysis has grown highly as it is very effective, as well as helpful to understand the feelings and thoughts of human beings. There have been more than 7000 research conducted on sentimental analysis according to [9]. As researcher's interest in sentimental analysis is increasing day by day, a lot of research has been done and is still going on in the Bengali context. In [10], Deep learning models have been used to detect emotions and recognize sentiments from various videos on YouTube, where comments were written in Bengali. To categorize the sentences, three-class, and five-class sentiment labels have been used. The three-class sentiment label provided an accuracy of 65.97% and five-class sentiment label gave accuracy of 54.24%. In [11], research has been conducted to investigate outcomes of various deep learning architectures for semantic analysis of movie reviews from Stanford Sentiment Treebank (SST) dataset. Naive Bayes, Recurrent Neural Network, Recursive Neural Network, Convolutional Neural Network, Convolutional Neural Network + word2vec have been used but Convolutional Neural Network + word2vec provided the best test accuracy (46.4%). In [12], BERT and ELECTRA models have been used on 3 different Bengali datasets for Bengali text document classification. The performance of those models was evaluated based on accuracy, precision, f1-score and recall. The outcome shows that ELECTRA model got higher accuracy and f1-score compared to BERT model while classifying Bengali documents. In [13], in order to create classifiers from multiple labeled datasets which are available on the internet, research has been done utilizing both traditional techniques like SVM and Random Forest and a variety of deep learning algorithms. The models have been assessed based on their effectiveness and the complexity of the available time resources. The findings indicate that transformed-based models performed better than other models. In [14], an analysis of textual data from several social media sites in Bangla has been performed the ed to detect and categorize sentiment and emotion with the use of traditional machine learning and neural network techniques. In this work, a comparison of several significant informative feature extraction was carried out as well. The results show that in case of Sentiment analysis, CNN classifier gave better output compared to other models. It provides an accuracy of 83% and F1-Score (0.822147) in case of analyzing Sentiment. On the other hand, Bi-LSTM gave better results in case of analyzing emotion as its accuracy is 65% and it also provided the highest F1-Score (0.618396).

As observed from this related work, there is a knowledge gap in understanding the performance of advanced NLP models on different datasets for distinct sub-tasks. Additionally, it remains unclear how the addition of traditional deep learning models may impact the overall performance. Thus, this study aims to address these gaps by conducting a comprehensive analysis of sentiment analysis and aspect classification in the Bengali language.

III. Describing Dataset

We first review the available datasets we used for this research in this section. After that, we conducted the data analysis to determine how unique these lexical contents are concerning the six different sentiments and the positive and negative sentiment classes.

A. Dataset description

We have collected two different open-source data:

- **SentNoB: A Dataset for Analysing Sentiment [15] :** The dataset used in this study contains a collection of social media user comments related to news articles and videos from various fields, including politics, education, and agriculture. The comments are labeled with one of three polarity categories: positive, negative, or neutral. These polarity labels provide insight into the sentiment expressed by the users towards the topic of the news article or video. The dataset is diverse and covers a broad range of topics, allowing for a comprehensive analysis of sentiment and polarity across different domains. The inclusion of social media comments is also significant, as it provides a real-world snapshot of public opinion on current events and issues. In the below Table : I, we can view the detailed segmentations of class distribution of examples in each class.

Table I
Class details for SentNoB: A Dataset for Analysing Sentiment.

Class Level	Total Instancesthe
Positive	6410
Negative	5709
Neutral	3609
Total	**15728**

- **Bengali Ekman's Six Basic Emotions Corpus [16] :**Based on Ekman's six fundamental emotions, this dataset has 36,000 Bangla data points. This balanced sample has 6000 data points for each of the six emotions. In the below Table : II, we can view the detailed segmentations of class distribution of examples in each class.

Table II
Class details for Bengali Ekman's Six Basic Emotions Corpus.

Class Level	Total Instances
Joy	6000
Sadness	6000
Anger	6000
Fear	6000
Surprise	6000
Disgust	6000
Total	**36000**

IV. Methodology

Here the task of classifying Bengali sentences can be divided into two parts: for dataset: [15], is used for determining the sentiment, and dataset: [16] is used for identifying the aspect

of the sentence. For both of the tasks, the Bidirectional Encoder Representations from Transformers (BERT) [3] model is used for its remarkable qualities and effectiveness. We used tokenization and special token addition to preprocess the dataset because BERT needs the data in a specific format [17]. The test dataset was then used to evaluate the BERT model after it had been given the training data. To cross-validate the output of the BERT model, additional Deep Learning [18] models such as Gated Recurrent Unit (GRU) [6], Bi-directional Long Short Term Memory (Bi-LSTM) [7], and Long Short Term Memory (LSTM) [8] were utilized, trained, and tested on our dataset. Several performance measures are used to assess the performance of our model. Here in Figure : 1, shows the proposed methodology of this research.

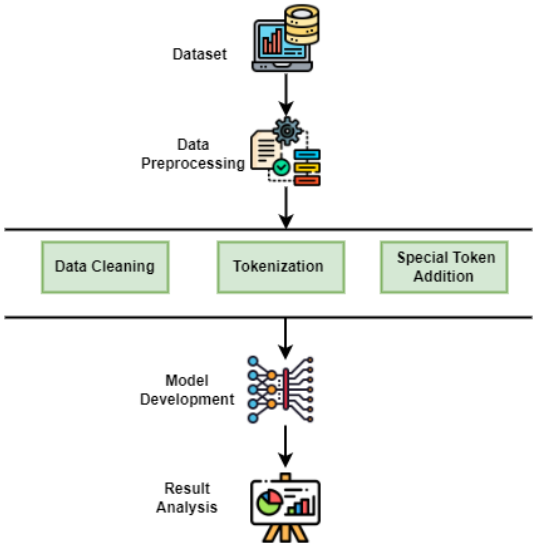

Figure 1. Methodology utilized in this study.

A. Data Preprocessing

Data preprocessing is a crucial step in preparing data for use with BERT. Though we use the pre-trained BERT model but data preprocessing is still necessary. Because pre-trained BERT models are trained on large amounts of data and fine-tuned for specific tasks, but the input data still needs to be preprocessed and formatted in a way that the model can understand and process. The following is the list of preprocessing processes which are used in this study.

- **Tokenization:** Converting the text into a list of tokens, or words and symbols, that can be easily processed by our proposed model. Here WordPiece tokenization was used.
- **Special Token Addition:** Adding special tokens, such as [CLS] and [SEP], to indicate the start and end of a sentence and separate individual sentences within a document.

After this, the training dataset and test dataset were created from both of the primary datasets. The scikit-learn library's Train test split function was utilized in this study to split the data.

V. Model Development

In our study, we utilized the powerful BERT model for both sentiment classification and aspect detection tasks. To begin with, we trained the BERT model on a training dataset for sentiment classification and aspect detection. We then evaluated the performance of the model on a separate test dataset. The BERT model is equipped with the Transformer's attention mechanism, which enables it to understand the contextual relationships between words. Additionally, BERT is pre-trained on a large amount of unlabeled text, considering both the left and right context, making it a highly efficient language model. In our research, we employed the pre-trained multilingual BERT model to classify aspects and sentiments in Bengali text. By leveraging the power of BERT, we aimed to achieve accurate and effective sentiment analysis and aspect detection in the Bengali language. We employed the HuggingFace Transformers implementation and pre-trained weights for our task. The BERT model consists of two parts, encoding and decoding, which respectively read and make predictions from text data. Over the encoding layer, a classification layer was added, with the Adam optimizer being used to calculate the loss function and optimize it using binary cross entropy. The n-train library was used to determine the optimal learning rate, and the model was trained on the training data for 100 epochs. Finally, the model was evaluated on the test dataset, as shown in the Figure: 2, which illustrates the BERT process for a Bengali sentence in our research.

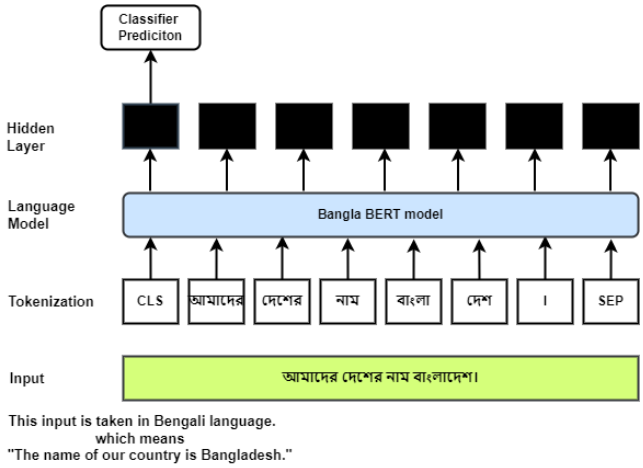

Figure 2. BERT processing on Bangla text in this study.

In addition to the BERT model, other deep learning models are used in this study, including the Gated Recurrent Unit (GRU) [6], Bi-directional Long Short Term Memory (Bi-LSTM) [7], and Long Short Term Memory (LSTM) [8]. In this study, BERT is used as a feature extractor by taking the hidden state outputs of the model and using them as inputs for a downstream task, such as sentiment analysis or aspect detection. This process is known as transfer learning and allows us to leverage the pre-trained BERT model to improve the performance of the downstream task [19]. The extracted

features from BERT can be fine-tuned with additional training data to further improve the performance. In figure: 3 is the architecture of this proposed model.

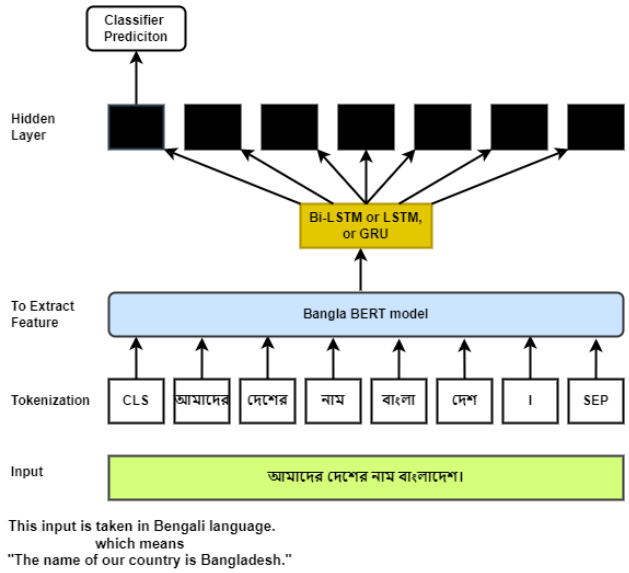

Figure 3. Bi-LSTM or LSTM, or GRU processing on Bangla text in this study.

VI. Result & Analysis

The result analysis for Bangla text classification using BERT and Bi-LSTM, LSTM, and GRU involves evaluating the performance of the model by calculating various metrics such as Confusion matrix, Accuracy [20], Precision [21], Recall [21] and F1-Score.

Confusion Matrix : Confusion matrices are a popular tool used to evaluate the performance of classification models. These matrices are used to measure the performance of both binary and multiclass classification problems [22]. The confusion matrix is a table that displays the predicted and actual values of a classification problem. The matrix consists of four different sets of values, including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

True Positive (TP) represents the number of observations that are predicted to be positive and are actually positive. True Negative (TN) represents the number of observations that are predicted to be negative and are actually negative. False Positive (FP) represents the number of observations that are predicted to be positive but are actually negative, and False Negative (FN) represents the number of observations that are predicted to be negative but are actually positive.

Accuracy: Accuracy measures the proportion of correct predictions made by the model overall predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision : Precision measures the number of correct positive results divided by the number of positive results predicted by the model.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall : Recall measures the number of correct positive results divided by the number of all actual positive results.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-Score : The F1 score is the harmonic mean of precision and recall and provides a balance between both.

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

These metrics help us understand how well the model is performing in terms of correctly identifying aspects and sentiments. By comparing the results of BERT with Bi-LSTM, LSTM, and GRU, we can determine which architecture is the most effective in classifying Bengali text. The result analysis is an important part of any research work and helps to determine the validity and reliability of the findings. In bellow Table III, we can see the summary of the result.

Table III
Accuracy, Precision, Recall and F1-Score

Model	Target	Accuracy	Precision	Recall	F1-Score
BERT	**Sentiment**	92.5	90.2	87.3	87.8
	Aspect	90.4	90.9	88.1	88.1
Bi-LSTM	**Sentiment**	90.1	89.7	90.2	90.1
	Aspect	90.2	89.2	88.7	89.1
LSTM	**Sentiment**	89.5	89.4	89.2	89.3
	Aspect	88.3	89.6	89.4	89.5
GRU	**Sentiment**	91.5	90.1	91.3	91.5
	Aspect	91.4	90.5	91.5	91.4

The above table III summarizes the results of the evaluation of the BERT model, Bi-LSTM, LSTM, and GRU for both sentiment analysis and aspect detection on Bengali text. The results show that the models performed well, with the highest accuracy and F1-Score achieved by the GRU model for sentiment analysis. The highest accuracy for aspect detection was achieved by the Bi-LSTM model. However, the performance of the BERT model was also noteworthy, as it achieved high precision, recall, and F1-Score values for both subtasks.

A. Confusion Matrix

In this study, confusion matrices were used to evaluate the performance of the BERT, Bi-LSTM, LSTM, and GRU models that were utilized for sentiment analysis and aspect detection in Bengali language. Figure 4 depicts the confusion matrices for all four models, and the results show that the BERT model outperforms the other models in terms of correctly predicting the sentiment and aspect of the Bengali sentences. The BERT model predicted 92.21% of the samples correctly, with only 7.79% of the samples being incorrectly classified.

On the other hand, the GRU model was found to have the highest misclassification rate, with around 9.29% of the data being missclassified. Specifically, 3.58% of the samples were predicted as positive sentences but actually needed to be classified as neutral sentences. This indicates a limitation of the GRU model in accurately detecting the sentiment and aspect of Bengali text.

Regarding the Bi-LSTM and LSTM models, they were found to predict almost the same percentage of samples correctly, with 87.36% and 87.13% respectively. Although the performance of these models is not as good as the BERT model, they still perform reasonably well in sentiment analysis and aspect detection of Bengali text.

Overall, the confusion matrices provide a clear visualization of the model's performance in classifying Bengali text based on sentiment and aspect, and demonstrate the effectiveness of BERT in comparison to other traditional deep learning models.

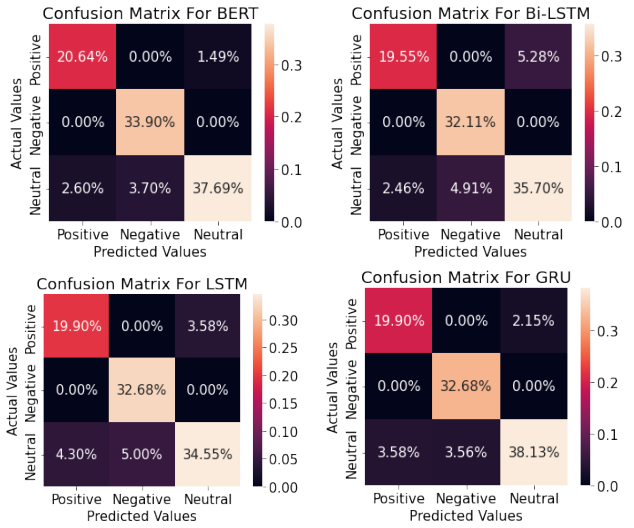

Figure 4. Confusion Matrix for the models utilized in this study to predict the aspect of the sentence.

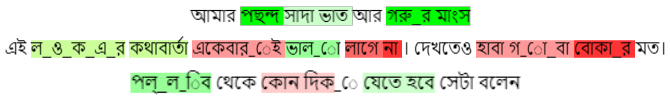

Figure 5. Example sentences highlighting the important words.

For a better interpretation of the model we explore and show word importance with three example sentences. This is obtained for BERT model trained on dataset [15] for sentiment analysis. These representations are produced using Captum [23]. In Figure 5, the first sentence, the model puts a high emphasis on the positive word পছন্দ (like). First sentence "আমার পছন্দ সাদা ভাত আর গরুর মাংস।" (I like beef and plain rice) is detected as positive with 98.78% probability by BERT. In the second sentence "এই লোকের কথাবার্তা একেবারেই ভালো লাগেনো। দেখতেও হাবাগোবা বোকার মত।" (I don't like the way this guy talks. He looks like a fool) is detected as negative senetce with 95.46% proability. Here the words ভালো লাগেনো (don'tlike), বোকা (fool) emphasize this negative. In the third sentence "পল্লবি থেকে কোন দিকে যেতে হবে" (Tell, which direction to go from Pallavi) is detected as neutral with 94.78% probability.

VII. Conclusion and Future Work

In conclusion, in this research, we proposed a model for classifying Bengali text into sentiments and aspects using pre-trained BERT and recurrent neural networks including Bi-LSTM, LSTM, and GRU. The results showed that the BERT model outperforms traditional recurrent neural networks in both sentiment and aspect classification tasks. The highest performance was observed with the GRU model with an accuracy of 91.5% in sentiment classification and 91.4% in aspect classification.

For future work, we plan to investigate other pre-trained models such as RoBERTa, ALBERT, etc. and compare their performance with BERT. Additionally, we aim to extend our study to other languages and domains, such as medical and legal text classification. Furthermore, we also plan to explore the use of different attention mechanisms and fine-tuning techniques to improve the performance of the models.

References

[1] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*, 2015.

[2] Shreyas Agrawal, Sumanto Dutta, and Bidyut Kr. Patra. Sentiment analysis of short informal text by tuning bert - bi-lstm model. In *IEEE EUROCON 2021 - 19th International Conference on Smart Technologies*, pages 98–102, 2021.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Moshiur Rahman Faisal. Bengali ekman's six basic emotions corpus, 2022.

[5] Sudipta Kar, Saiful, and Khondoker I. Islam. Sentnob, 2022.

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[7] Shouxiang Wang, Xuan Wang, Shaomin Wang, and Dan Wang. Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting. *International Journal of Electrical Power & Energy Systems*, 109:470–479, 2019.

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

[9] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[10] Nafis Irtiza Tripto and Mohammed Eunus Ali. Detecting multilabel sentiment and emotions from bangla youtube comments. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–6. IEEE, 2018.

[11] Houshmand Shirani-Mehr. Applications of deep learning to sentiment analysis of movie reviews. In *Technical report*. Stanford University, 2014.

[12] Md Mahbubur Rahman, Md Aktaruzzaman Pramanik, Rifat Sadik, Monikrishna Roy, and Partha Chakraborty. Bangla documents classification using transformer based deep learning models. In *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pages 1–5. IEEE, 2020.

[13] Md. Arid Hasan, Jannatul Tajrin, Shammur Absar Chowdhury, and Firoj Alam. Sentiment classification in bangla textual content: A comparative study. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–6, 2020.

[14] Hasmot Ali, Md. Fahad Hossain, Shaon Bhatta Shuvo, and Ahmed Al Marouf. Banglasenti: A dataset of bangla words for sentiment analysis. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–4, 2020.

[15] Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. Sentnob: A dataset for analysing sentiment on noisy bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271, 2021.

[16] MD Asif Iqbal, Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H Sarker. Bemoc: a corpus for identifying emotion in bengali texts. *SN Computer Science*, 3(2):135, 2022.

[17] Moythry Manir Samia, Alimul Rajee, Md. Rakib Hasan, Mohammad Omar Faruq, and Pintu Chandra Paul. Aspect-based sentiment analysis for bengali text using bidirectional encoder representations from transformers (bert). *International Journal of Advanced Computer Science and Applications*, 13(12), 2022.

[18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[19] K. Mouthami, S. Anandamurugan, and S. Ayyasamy. Bert-bilstm-bigru-crf: Ensemble multi models learning for product review sentiment analysis. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, pages 1514–1519, 2022.

[20] John A Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988.

[21] Michael Buckland and Fredric Gey. The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19, 1994.

[22] Md Asif Bin Khaled, Md Junayed Hossain, Saifur Rahman, and Jannatul Ferdaus. Multiclass classification for gvhd prognosis prior to allogeneic stem cell transplantation. In *AI 2022: Advances in Artificial Intelligence: 35th Australasian Joint Conference, AI 2022, Perth, WA, Australia, December 5–8, 2022, Proceedings*, pages 487–500. Springer, 2022.

[23] N Kokhlikyan, V Miglani, M Martin, E Wang, B Alsallakh, J Reynolds, et al. Captum: a unified and generic model interpretability library for pytorch. arxiv200907896 cs stat. september 16, 2020. *Back to cited text*, (55), 2022.