Aspect-based Sentiment Analysis for Bengali Text using Bidirectional Encoder Representations from Transformers (BERT)

Moythry Manir Samia[1*], Alimul Rajee[2*], Md. Rakib Hasan[3], Mohammad Omar Faruq[4], Pintu Chandra Paul[5]
Dept. of Information and Communication Technology, Comilla University, Cumilla, Bangladesh[1,2,3,5]
Dept. of Accounting and Information Systems, Comilla University, Cumilla, Bangladesh[4]

***Abstract*—Public opinion is important for decision-making on numerous occasions for national growth in democratic countries like Bangladesh, the USA, and India. Sentiment analysis is a technique used to determine the polarity of opinions expressed in a text. The more complex stage of sentiment analysis is known as Aspect-Based Sentiment Analysis (ABSA), where it is possible to ascertain both the actual topics being discussed by the speakers as well as the polarity of each opinion. Nowadays, people leave comments on a variety of websites, including social networking sites, online news sources, and even YouTube video comment sections, on a wide range of topics. ABSA can play a significant role in utilizing these comments for a variety of objectives, including academic, commercial, and socioeconomic development. In English and many other popular European languages, there are many datasets for ABSA, but the Bengali language has very few of them. As a result, ABSA research on Bengali is relatively rare. In this paper, we present a Bengali dataset that has been manually annotated with five aspects and their corresponding sentiment. A baseline evaluation was also carried out using the Bidirectional Encoder Representations from Transformers (BERT) model, with 97% aspect detection accuracy and 77% sentiment classification accuracy. For aspect detection, the F1-score was 0.97 and for sentiment classification, it was 0.77.**

***Keywords*—Sentiment analysis; Bengali sentiment analysis; Aspect-Based Sentiment Analysis (ABSA); Bengali ABSA; deep learning; Bidirectional Encoder Representations from Transformers (BERT)**

I. Introduction

Online communication has grown extremely prevalent all around the world due to the internet. Social media is used by about 51% of the world's population overall. They use social media on a daily basis for an average of 2.5 hours [1]. Social media platforms, for example, Twitter and Facebook have made it simple to convey comments. Online newspapers have replaced traditional newspapers, and readers now even share their opinions about many news stories in the news portal's comment section. People express their perspectives and ideas on a wide range of topics, including politics, business, and international issues. As a result, a significant volume of data is created daily and contains many insightful opinions on a variety of subjects on those online sites. For many widespread applications, it's vital to comprehend opinions from user comments. For instance, the quality of goods or services can be improved by studying client feedback from comments on e-commerce sites. Also, People frequently voice their opinions on many national issues in democratic nations like Bangladesh, India, USA on social media or news websites, which must be taken into consideration for a country's development. Studying all the text manually and trying to understand what the individuals are feeling is difficult. So, all of this data must be automatically studied to provide useful information as it is quite beneficial to understand the viewpoints of the general public on a particular topic. ABSA is a popular method for obtaining such useful data.

Sentiment analysis, often referred to as opinion mining, is a technique for determining if a given expression is negative, positive, or indifferent. There are three layers of sentiment analysis to consider, the first one to evaluate is the document level, the second is the sentence level, and the third is the level for aspect. At document level, the text is analyzed to see if it has a favorable or unfavorable sentiment. The polarity of every sentence is determined by sentence level. These two layers of analysis don't indicate what people liked and didn't like in particular. This shortcoming of earlier layers is resolved by the third layer. Sentiment analysis is known as ABSA on the aspect label. The main objective of ABSA is to find out the correlation of a piece of text with a set of aspects and deduce their sentimental polarities in the process. The most comprehensive version for document-level SA to retrieve key information is ABSA. ABSA consists of two key tasks. 1) Extricating the specific aspects addressed in the text. 2) For each aspect, identify the sentiment's polarity [2]. For example, "The fabric has decent quality, but the price is excessive." This evaluation focuses on two factors: 'price' and 'fabric'. 'Positive' feeling is indicated with the 'price' aspect, whereas a 'negative' feeling is shown with the 'fabric' aspect. Aspect categories are stated specifically inside this review. However, reviews can be implicit at times. For instance, "I asked the waiter three times for tomato sauce, but he didn't bring any". The implicit aspect here is a restaurant's 'service'.

Bengali is spoken by around 226 million people worldwide, with a large proportion of them using the internet [3]. There are very few ABSA datasets available in Bengali language. Not many Bengali-based ABSA works are available, to the extent that we are aware, and the insufficiency of Bengali ABSA datasets is one of the key barriers to the development of Bengali ABSA. The present Bengali ABSA research is based on a variety of deep learning and machine learning strategies, for instance: Support Vector Machine (SVM), Random Forest (RF), and Convolutional Neural Network (CNN). BERT is

well-known for resolving many shortcomings of these machine learning and deep learning techniques. In a self-supervised fashion BERT learns the language model by utilizing Transformer's encoder. But to the best of our knowledge, BERT was never used in any prior ABSA works for Bengali text.

As we previously discussed, there isn't much research done on ABSA in Bengali. We committed our work to the Bengali language by making the following contributions:

- For implementing ABSA in Bengali text, we collected heterogeneous Bengali data from news portals, YouTube channels and comments from social media.
- We annotated the collected dataset in five aspects: Technology, Corruption, Economy, Politics, and Sports, and three sentiment categories: Neutral, Positive, and Negative.
- We performed aspect detection and sentiment classification using our dataset by implementing BERT, which provided the best performance in terms of Bengali ABSA to the best of our knowledge.

The paper is organized as follows. The literature review is covered in Section II, and the research procedures for our experiment are described in Section III. We present and analyze the experimental results in Section IV. Finally, Section V provides the conclusion along with potential future directions.

II. Literature Review

A fast-growing and improving sub-field of Natural Language Processing (NLP) is sentiment analysis. Subramaniam *et al.* conducted a sentiment analysis-based survey in their research [4]. They studied the IMDB dataset and observed that sequence neural models had the best results, but they come at a high cost in terms of classifier complexity. By comparison, CNN had a preferable performance with less complexity over models of sequence neural. YouTube videos were used for a study on sentiment analysis by Alhujaili *et al.* in their research work [5]. The development of sentiment analysis is being accelerated by cutting-edge deep learning techniques like LSTM and BERT. Althobaiti suggested an automated technique for extracting coarse-grained hate speech and offensive language from Arabic tweets [6]. BERT was contrasted with two traditional machine learning methods SVM and Logistic Regression (LR). Along with the textual content of the tweets, the use of sentiments and text-based emoji descriptions was also looked at. They had an F1-score of 84.3% for offensive language detection, 81.8% for hate speech detection, and 45.1% for fine-grained hate-speech recognition by applying BERT. BERT was used by Nijhawan *et al.* for sentiment classification in their research [8]. Experiments revealed that, despite its comparatively simpler structure, the BERT model outperforms many prominent models in this subject.

The majority of Bengali sentiment analysis works are based solely on determining the text's polarity. Recurrent Neural Network (RNN) with Long Short-term Memory (LSTM) was used to analyze sentiment in Bengali data to avoid long-term reliance by Ahmed *et al.* [3]. Demonstration of the benefits of tuned hyper-parameter and how it can aid sentiment analysis on a dataset was given in their paper. An attention-based CNN algorithm was proposed by Basri *et al.* for analyzing sentiment on the Bang-lish Disclosure in their paper [10]. Islam *et al.* introduced two fresh datasets for sentiment analysis in Bengali in their research that was manually tagged [11]. $BERT_{BSA}$, which is a deep learning method of sentiment analysis for the Bengali language that surpassed all the previous models, was also introduced in their research.

ABSA was first discussed as a chapter in Liu, B.'s research paper. There were discussions of ABSA's methods and related issues [12]. The goal of SemEval2014's Task 4 was to advance the field of ABSA. Pontiki *et al.* developed and published benchmark datasets for ABSA with manually labeled reviews from the restaurant and laptop sectors [13]. Four subtasks, Aspect term extraction, Aspect category detection, Aspect term polarity, and Aspect category polarity—were used to analyze the 163 entries from 32 teams. Karimi *et al.* research on adversarial training in ABSA was discussed in their paper [14]. The studies with the proposed architecture show that using adversarial instances during network training enhances the efficiency of the common purpose BERT, while BERT which is in-domain and post- trained, brought better results for feature extraction and sentiment classification. Suciati *et al.* suggested a methodology to ensemble aspect extraction focused on Part-of-speech (POS) tagging, dependency parsing, and dense neural networks [7]. They found that deep learning integrated with conventional methods can yield better outcomes than lexicon-based approaches.The aspect-agnostic issue, which is pervasive in the context modeling of ABSA, was identified and discussed by Xing *et al.* [9]. They claimed that during context modeling, the semantics of the provided aspect should be taken into account as a new indication outside the context. In order to produce the aspect-aware masked states specifically designed for the ABSA work, they presented context encoders that are aspect-aware, such as Aspect-Aware BERTs (AABERTs), Aspect-Aware Graph Convolutional Networks (AAGCN), and Aspect-Aware LSTM (AALSTM).

Rahman *et al.* constructed two freely accessible datasets, 'Cricket' and 'Restaurant', to be used for the ABSA in Bangla, which function as a benchmark for the Bangla ABSA field [15]. It was the first ABSA work in Bengali. Customer reviews of restaurants make up one of the datasets, while manually labeled user comments on cricket make up the other. 2900 comments about cricket in five aspect categories are included in the 'Cricket' dataset, while 2600 restaurant reviews are included in the 'Restaurant' dataset. They designed an aspect category extraction model which is based on CNN architecture in the same year [2]. They used their suggested datasets [15] to compare their model with well-known machine learning techniques. For the 'Cricket' dataset, they achieved 51% F1-Score while KNN, SVM, and RF classification only achieved 35%, 34%, and 37% accuracy, accordingly. For the 'Restaurant' dataset, KNN, SVM, and RF classification achieved 42%, 38%, and 38% F1-Score, respectively, while CNN outperformed them with 64% F1-Score. In [16], Boidini implemented three stacked Auto-encoders (AEs) models based on the datasets from [15] to categorize aspects in the Bengali text. The stacking network's layers were individually trained to comprehend the encoded data of the layer that came before them. The three layers that were trained in a stacked manner were AE, Sparse AE (SAE), and Contractive AE (CAE). In comparison to the studies of Rahman *et al.* [2,15], all of the suggested models exhibit improved precision, F1 score, and

recall. The CAE model especially generated the best F1-score on the 'Restaurant' and 'Cricket' datasets with 0.87 and 0.91 respectively. Based on the dataset provided by Rahman *et al.* [15], Haque *et al.* constructed the Bangla ABSA model [17]. The traditional supervised aspect classification machine learning approach was employed in this work with minimal preprocessing. They observed that if the data is less preprocessed, then the result has a higher F1 Score. To perform aspect extraction on the dataset provided by Rahman *et al.* [15], F.A. Naim presented a novel technique called PSPWA (Priority Sentence Part Weight Assignment) [1]. PSPWA is a phase of the data pre-processing process that leads to better performance. CNN and traditional supervised learning methods were utilized to show the performance. Compared to other learning algorithms, CNN performed best. CNN's F1-score for the 'Cricket' dataset was 0.59, and for the 'Restaurant' dataset it was 0.67.

The summary of this literature review is presented in Table I.

TABLE I. SUMMARY OF LITERATURE REVIEW

Author	Year	Dataset	Used Method	Result
Ahmed *et al.* [3]	2020	Prothom Alo, Cricket [15], etc.	LSTM	94%
Subramaniam *et al.* [4]	2021	IMDB dataset	RNN,CNN-LSTM, etc.	CNN-LSTM: 96%
Althobaiti [6]	2022	Data set from [20]	LR, SVM, BERT	Offensive language: 84.3%, Hate speech: 81.8%, Fine-grained hate speech: 45.1%
Suciati *et al.* [7]	2020	from PergiKuliner platform	SVM, LR, DT,ET	Food:GRU: 88.16%, BiLSTM: 88.16% Price:RF: 89.54%, Service: BiLSTM: 89.03% Ambience: BiLSTM: 84.78%
Nijhawan *et al.* [8]	2022	100042 tweets	RF,DT, LR,BERT	97.78%
Basri *et al.* [10]	2021	Dataset with 5000 short paragraphs	Attention -based CNN	Funny sentiment: 90% Multiclass data: 83%
Islam *et al.* [11]	2020	Prothom alo	Multi-lingual BERT	3-class: 60% 2-class: 71%
Pontiki *et al.* [13]	2014	Restaurant and laptop sectors reviews	CRF with features extracted	Laptop: 74.55% Restaurant: 84.01%
Karimi *et al.* [14]	2020	SemEval-2016, SemEval-2014	BERT	Laptop: 85.57% Restaurant: 81.50%
Bodini *et al.* [16]	2019	Cricket Restaurant [15]	AE, SAE, CAE	Cricket- CAE: 0.88 Restaurant- CAE: 0.87
Naim *et al.* [1]	2021	Cricket Restaurant [15]	CNN, RF, SVM etc.	Cricket-CNN: 0.59, RF: 0.41 Restaurant- CNN: 0.67, , RF: 0.35
Haque *et al.* [17]	2020	Cricket Restaurant [15]	SVM, RF, LR etc.	Cricket- SVM: 0.35, RF: 0.37 Restaurant- LR: 0.43, RF: 0.35
Rahman *et al.* [2]	2018	Cricket Restaurant [15]	CNN, RF, SVM etc.	Cricket-CNN: 0.51, RF: 0.37 Restaurant- CNN: 0.64 , RF: 0.38

III. METHODOLOGY

This work can be divided into two subtasks: determining a Bengali sentence's aspect and identifying sentiments. For the first subtask, five aspects can be defined: Technology, Corruption, Sports, Politics, and Economy. As for the second subtask, there will be three sentiment categories: Positive, Neutral, and Negative. BERT was used for both subtasks due to its exceptional features and efficiency. We began by collecting data, then labeled it and cleaned the dataset by removing stop words, punctuations, and decimal values. Following that, we preprocessed the dataset using tokenization and special token addition because BERT requires the data in a specific format. The training dataset was then fed into the BERT model, and the model was tested using the test dataset. Other deep learning models, such as CNN, Bidirectional LSTM (BiLSTM), LSTM, RNN and Gated Recurrent Unit (GRU) models, were also built, trained, and tested on our dataset to cross-validate the BERT model's results. Our model's performance is evaluated using a variety of performance metrics. The process as a whole is depicted in Fig. 1 below.

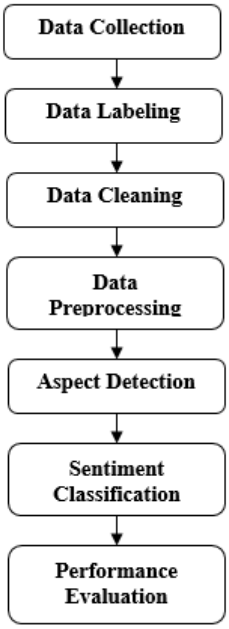

Fig. 1. Aspect-based sentiment analysis process

A. Data Collection

The dataset is a critical component of a machine learning algorithm since it determines the outcome based on its size and quality. Data was collected from social media, news portals, and Youtube comments of Bangladeshi users on five topics: Technology, Corruption, Sports, Politics, and Economy. The comments were stored in a Microsoft excel file.

B. Data Labeling

Some multi-lined sentences were simplified to single sentences by removing unnecessary sentences manually. Following that, 10042 sentences were left to be labeled. After that each author individually annotated the dataset. Each sentence was assigned to one of the five aspect categories—technology, corruption, sports, politics, and the economy—and was then annotated for sentiment polarity into three categories: positive, negative, and neutral. Each participant added an annotation to every comment. To figure out the ultimate aspect category and

the polarity of a statement, voting percentage was calculated. In the final dataset, there are three columns for 10042 annotated sentences. One is for text or comments, the second one is for aspect and the final one is for sentiment category. Fig. 2 contains a sample of our dataset.

Text	Category	Polarity
রন্ধ্রে রন্ধ্রে জড়িয়ে পড়েছে যে দেশে দুর্নীতি ও সুদ ঘুষ কিছুতেই জেন লাগাম টানা যাচ্ছেনা.!!!	Corruption	Negative
ইলিশ রপ্তানি করে প্রচুর বৈদেশিক মুদ্রা অর্জন করা সম্ভব হচ্ছে	Economy	Positive
দুর্নীতি দমন কমিশন ক্ষমতাবানদের বিরুদ্ধেও স্বাধীন তদন্ত করবে এমনই আমাদের প্রত্যাশা	Corruption	Positive
গ্রামে আমরা টুজি নেটওয়ার্কই ঠিকমতো পাইনা, তারা আবার আসছে ফোরজি নিয়ে। যত্তসব।	Technology	Negative
আওয়ামীলীগ জনগণের জন্য কাজ করে বলেই তারা সফল।	Politics	Positive
আর কোন দিন বাংলাদেশ ক্রিকেট দল জয় এর মুখ দেখবে না যত দিন পাপন সভাপতি থাকবে।	Sports	Negative
কোন ব্যাংক থেকে এ ঋণ পাওয়া যাবে?	Economy	Neutral

Fig. 2. Sample of the labeled dataset

The Zipf's law was implemented in our dataset. According to Zipf's Law, there should be an inverse relationship between the frequency of various words in the dataset and their positions [18]. Fig. 3 demonstrates how Zipf's Law applies to our dataset. It suggests that the words in our dataset are not arbitrary.

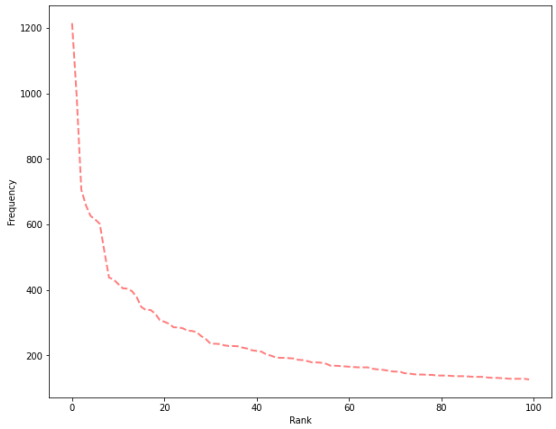

Fig. 3. Word frequency of our dataset using Zipf's law

Every effort was made for balancing the dataset, however for the chosen topics, people are more likely to comment on negative instances than positive ones, and comparatively less comments were without opinion. Because of this there are a few more negative comments with sentiment labeling and number of Neutral comments is comparatively less. As for the aspect labeling, all five aspects have nearly identical numbers of comments. Tables II and III display the dataset's statistics for aspect labeling and sentiment labeling, respectively.

TABLE II. STATISTICS OF ASPECT LABELING

Aspect category	No. of data
Technology	2003
Economy	2011
Corruption	2001
Sports	2021
Politics	2006
Total	**10042**

TABLE III. STATISTICS OF SENTIMENT POLARITY LABELING

Sentiment polarity	No. of data
Positive	3507
Negative	3909
Neutral	2626
Total	**10042**

C. Data Cleaning

NLP relies significantly on data preparation. To adapt the machine learning algorithm to the Bengali data, data cleansing is necessary. To clean the dataset, a few procedures were used.

Step 1: The dataset was first changed to lowercase.

Step 2: Several unnecessary decimal digits appeared in the text. So, they were taken out.

Step 3: We eliminated all types of punctuation, extra characters, and components.

Step 4: Bengali stop words come in a variety of forms. These cause issues when the data was examined. Therefore, all stop terms from the dataset were eliminated using a corpus of Bengali stop-words.

Fig. 4 displays the flow chart for data cleaning.

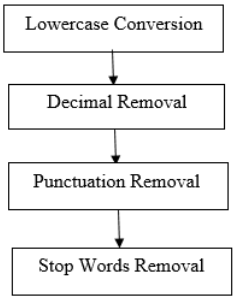

Fig. 4. Flow chart for data cleaning

D. Data Preprocessing

Text preprocessing is an important step in achieving better accuracy in Natural language processing tasks. The dataset should be preprocessed in a particular way for pre-trained models like BERT. Two steps were followed for preprocessing, tokenization and special token addition. The BERT model was trained to transform sentences into tokens. WordPiece tokenization was used in this case. This type of tokenization is useful when dealing with words that are not in our dictionary. After that special tokens, like [CLS], [SEP] were included in the data. The main dataset was then divided into two subsets: the training dataset and the test dataset. Data splitting was carried out in this research using the scikit-learn library's Train test split function.

E. Model Development

For both of our subtasks, aspect detection and sentiment classification, we used the BERT model. We began the aspect

detection task by training the BERT model with training data and testing the model with test data. We used the same method to detect sentiment. By fully understanding both the left and right context, BERT aims to pre-train from the unlabeled text. Making use of Transformer's attention mechanism, BERT learns contextual relationships between words.

BERT: BERT is an acronym that stands for Bidirectional Encoder Representations from Transformers [20]. It is a free and open-source NLP framework. Transformer serves as an inspiration for BERT as it studies the connections between words in a sentence. Transformer comes with an encoder and a decoder that reads the input data and predicts the outcome. BERT only requires the encoder component of the transformer to complete its task. In order to clarify the meaning of vague terms in data, BERT attempts to provide context to the surrounding data. The BERT framework was pre-trained using a large volume of Wikipedia texts. BERT considers the context of each word of the text. A word's vector in early word embedding techniques was always the same, regardless of where the word appeared in the text. Contrary to them, BERT provides different vectors for the same words depending on their position. For input representation, the first sentence starts with a classification token, [CLS], and each subsequent sentence ends with a separation token, [SEP]. BERT includes two steps, pre-training and fine-tuning.

a. Pre-Training: In this step, the model is trained on unlabeled text data. Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) are associated with the pre-training task. MLM and NSP are applied together to reduce the total loss function of the two methods [26].

- **Masked Language Modeling (MLM):** The first step is to arbitrarily mask out 15% words in the input text data and replace them with [MASK] tokens [26]. The BERT attention-based encoder is then applied to the entire sequence, which only predicts the masked words based on the context that the other, non-masked words in the sequence provided. Next, the attention-based encoder of BERT processes the input. It only predicts masked words using the context given by the other non-masked words in the sentence.
- **Next Sentence Prediction(NSP):** NSP predicts the connection between two sentences [26]. The primary task of NSP is to determine whether the second sentence comes after the first sentence or not. If the second sentence comes after the first sentence, it is referred to as IsNext; otherwise, it is referred to as NotNext. The sentence is IsNext 50% of the time and NotNext 50% of the time.

b. Fine Tuning: BERT can be fine-tuned to perform a variety of language tasks. By simply placing a single layer over the base model, we can adjust the base model according to our own dataset [26]. The pre-trained parameters are used to initialize the BERT model at first. Then, using annotated data from the required tasks, all of the parameters are fine-tuned. Multilingual BERT is trained in 104 languages using Masked Language Modeling (MLM) [27], and Bengali is one of them. In our work, we used the multilingual BERT model to identify aspects and sentiments in order to gauge the scope of this language model. The pre-trained model weights and implementation for HuggingFace's Transformers have been made public, and it was used in our work. The two parts of BERT are encoding and decoding, respectively. Encoding is used to read text data, and decoding is used to make predictions. The classification layer is then placed over the encoding layer. In our work, in the classification layer, the loss function was computed using binary cross entropy, and the Adam optimizer was applied. The ktrain library was used to calculate the proper learning rate. The model was fed with training data. Then, the model was trained with the previously calculated learning rate, weight, and 10 epochs. Then, testing was done using the test dataset. Fig. 5 depicts how BERT processes a Bengali sentence.

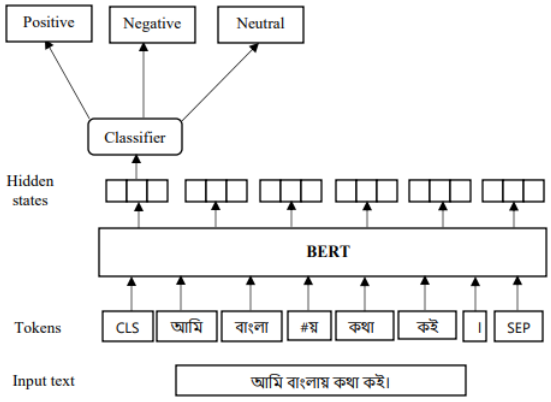

Fig. 5. BERT processing on Bengali text

Along with BERT, a number of deep learning models were also used in this ABSA work. Word2Vec [28] preprocessing technique was used in those models for feature extraction. Some brief descriptions of the algorithms are given below:

- **CNN:** CNN is a member of the family of ANN in deep learning. Using a variety of building pieces, including convolution layers, fully connected layers, and pooling layers, CNN is intended to adaptively and instinctively grasp spatial feature hierarchies by backpropagation [21].
- **RNN:** The outcome from the prior step is provided as input to the ongoing step in an RNN, a form of neural network [22]. In contrast to conventional neural networks, RNN input and output are interdependent. Many times, when forecasting the upcoming word of a statement, the prior words are necessary, and thus the preceding words must always be memorized. RNN was designed to address this issue by utilizing a Hidden Layer. The Hidden state is the most essential characteristic of RNN. It stores some sequence information in memory.
- **LSTM:** An RNN variant that can pick up dependency of order in predicting sequence issues is the LSTM network [23]. Every step uses the output from the previous one as its input in this technique. It resolved the problem of RNN long-term reliance, which is

that RNN can anticipate words looking at recent data but cannot predict words kept in long-term memory. The substantial number of memory cells and 4 neural networks, which are arranged in a chain pattern, make up the LSTM. The four components of a typical LSTM unit: Cell, Input gate, Output gate, Forget gate. Three gates regulate the data flow in and out of the cell, as well as the cell retains contents for arbitrary time periods. Time series with indeterminate duration can be examined, categorized, and predicted with the LSTM algorithm.

- **BiLSTM:** Two models are integrated into the bidirectional LSTM rather than training only one [24]. The input sequence is acquired by the first model, while its reverse is understood by the second model. Context from both sides enhances the performance of language-based models in numerous applications, including language translation, speech recognition, etc. Bi-Directional LSTMs are employed to use those features.
- **GRU:** GRU is a variant of RNN [19]. Similar to LSTM, it was created to address the vanishing gradient issue with RNN. It only has three gates, as opposed to the four in LSTM, and it doesn't keep track of the internal state of the cell. The hidden state of the GRU incorporates similar information as from the internal cell state of the LSTM recurrent unit. This set of information is sent to the subsequent GRU.

IV. Experiment and Result

A. Preparing Environment

Deep learning-based experiments need powerful CPU and GPU. In order to conduct experiments, a system with a core i5 processor with 8 GB of RAM was used. KERAS has been programmed in Python while TensorFlow has served as the backend.

B. Result

The results were evaluated for both subtasks using well-known evaluation matrices. They are F1-score, accuracy, recall, and precision. They are assessed using True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Equations for accuracy, precision, recall, and F1-score are [25]:

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

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

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

$$F1 - score = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{4}$$

Both aspect detection and sentiment classification are multi-label classification. Tables IV and V display the findings of our analysis of the dataset for aspect extraction and sentiment classification.

TABLE IV. Aspect Detection Result

Algorithm	Accuracy	Precision	Recall	F1-score
BERT	**0.97**	**0.97**	**0.97**	**0.97**
BiLSTM	0.95	0.95	0.96	0.96
LSTM	0.94	0.95	0.94	0.95
GRU	0.88	0.94	0.90	0.92
RNN	0.71	0.85	0.75	0.78
CNN	0.91	0.93	0.92	0.93

TABLE V. Sentiment Classification Result

Algorithm	Accuracy	Precision	Recall	F1-score
BERT	**0.77**	**0.78**	**0.77**	**0.77**
BiLSTM	0.74	**0.78**	0.74	0.75
LSTM	0.73	**0.78**	0.74	0.75
GRU	0.62	0.77	0.64	0.69
RNN	0.52	0.68	0.51	0.58
CNN	0.63	0.74	0.67	0.69

Tables IV and V show that BERT got the highest score in both subtasks. Its F1-Score was 0.97 and it had a 97% accuracy rate for aspect detection. As for sentiment classification BERT accurately classified 77% sentiments, with an F1-Score of 0.77. BiLSTM achieved a 95% accuracy in aspect detection, closely followed by LSTM at 94%. Their respective F1 scores for these subtasks were 0.96 and 0.95. In terms of sentiment classification, BiLSTM predicted 74% sentiment correctly, while for LSTM it was 73%. CNN and GRU performed almost similarly in both subtasks. Their F1-score for those subtasks were same (0.75). Despite having the same F1-score in sentiment classification (0.69), CNN outperformed GRU in aspect detection (0.93). The subtasks with the poorest performance were completed by RNN. It achieved an F1-Score of 0.78 and an accuracy of 71% for aspect detection. Only about half of the sentiments were correctly predicted by RNN (accuracy: 52%). The RNN's F1-Score for detecting sentiments was 0.58.

C. Discussion

Based on the findings in the result section, the BERT model outperformed every other deep learning algorithm, in the both subtasks of ABSA, aspect detection and sentiment classification. They achieved the best scores in all the evaluation matrices due to its "self-attention" quality. Other classifiers' use of the word embedding method, which is not context-based, led them to miss some of the word's context and produce some inaccurate predictions. Contextual embedding, used by BERT, ensures that words are understood in relation to one another. It is one of the main factors influencing BERT's exceptional performance.

The results of LSTM and BiLSTM were close, but BiLSTM outperformed LSTM in terms of its ability to accurately detect more aspects and sentiments due to its capacity to understand the context from both sides of a statement, whereas LSTM only comprehends context from the beginning of a sentence. GRU and CNN performed comparably. RNN performed rather poorly as a result of vanishing gradient issues.

It can also be seen that aspect detection produces better results than sentiment classification since the dataset is evenly

distributed in terms of aspects but slightly unbalanced in terms of sentiments.

D. Comparison with Previous Works

The results of this study were compared with four earlier Bengali ABSA based papers, [1], [2], [15], [17] and [18]. Table VI shows a comparison of our dataset to datasets used in their studies. Tables VII and VIII, respectively, provide performance comparisons for aspect detection and sentiment classification among them.

TABLE VI. COMPARISON OF BENGALI ABSA DATASETS

Dataset	Total comments	Aspect category	Sentiment polarity
Cricket [15]	2958	Batting, Bowling, Team, Team management, Other	Positive Negative Neutral
Restaurant [15]	2053	Food, Price, Service, Ambiance, Miscellaneous	Positive Negative Neutral
BAN-ABSA [18]	9009	Politics, Sports, Religion, Others	Positive Negative Neutral
Our Dataset	10042	Technology, Economy, Corruption, Sports, Politics	Positive Negative Neutral

TABLE VII. COMPARISON OF ASPECT DETECTION WORKS FOR BENGALI ABSA

Research	Dataset	Algorithm	Accuracy	F1-score
Rahman *et al.* [2, 15]	Cricket	RF	0.25	0.37
		SVM	0.19	0.35
		CNN	0.81	0.51
Rahman *et al.* [2, 15]	Restaurant	RF	0.30	0.33
		SVM	0.29	0.38
		CNN	0.83	0.64
Haque *et al.* [17]	Cricket	RF	-	0.37
		SVM	-	0.35
		LR	-	0.34
Haque *et al.* [17]	Restaurant	RF	-	0.35
		SVM	-	0.39
		LR	-	0.43
		RF	-	0.41
F. A. Naim [1]	Cricket	SVM	-	0.48
		CNN	-	0.59
F. A. Naim [1]	Restaurant	RF	-	0.35
		SVM	-	0.52
		CNN	-	0.67
Masum *et al.*[18]	BAN-ABSA	RF	-	0.65
		SVM	0.69	0.66
		CNN	0.79	0.75
		LSTM	0.77	0.76
		BiLSTM	0.80	0.78
This Research	Our Dataset	CNN	0.91	0.93
		RNN	0.71	0.78
		GRU	0.88	0.92
		LSTM	0.94	0.95
		BiLSTM	0.95	0.96
		BERT	**0.97**	**0.97**

Tables VI, VII, and VIII clearly show:

- Among all the datasets, the CNN algorithm gave the best accuracy and F1-score (0.93 and 0.93) in aspect detection for the dataset used in our research. A clear illustration of the performance of CNN algorithm in these datasets can be found in Fig. 6.

TABLE VIII. COMPARISON OF SENTIMENT CLASSIFICATION WORKS FOR BENGALI ABSA

Research	Dataset	Algorithm	Accuracy	F1-score
Masum *et al.*[18]	BAN-ABSA	SVM	0.65	0.41
		CNN	**0.72**	0.60
		LSTM	0.70	0.61
		BiLSTM	0.71	**0.62**
This Research	Our Dataset	CNN	0.63	0.69
		RNN	0.52	0.58
		GRU	0.62	0.69
		LSTM	0.73	0.75
		BiLSTM	0.74	0.75
		BERT	**0.77**	**0.77**

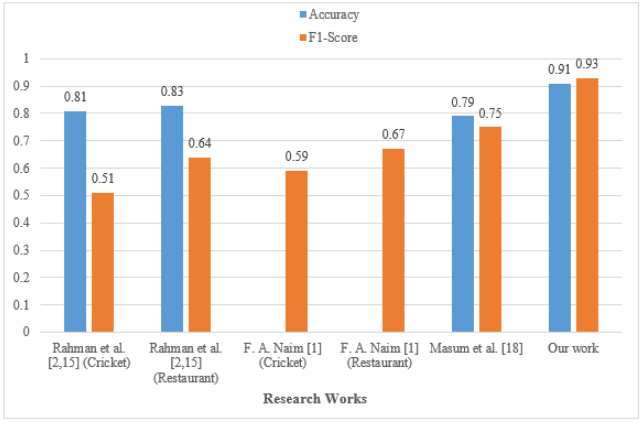

Fig. 6. Representation of CNN performance in research works for aspect detection

- It also can be seen that for the BiLSTM algorithm, that was applied by Masum *et al.* [18] and in our work for aspect detection, it scored better in this work with an accuracy score of 0.95 and F1-score of 0.96.
- To the best of our knowledge, among all Bengali ABSA works, for subtask 1, aspect detection, this research got the best accuracy score and the best accuracy and F1-score for BERT (0.97 and 0.97). Fig. 7 depicts the best aspect detection scores from the research works for clear comparison.
- To the best of our knowledge, this research received the best accuracy and F1-score for BERT (0.77 and 0.77) for subtask 2, sentiment classification, in comparison to the only other ABSA subtask 2 work that has been done in [18]. The comparison of the highest sentiment classification scores from the research works is shown in Fig. 8.
- When it comes to performance of the model, the dataset's richness and quality are important. Given that the dataset used in this research is larger than the Cricket, Restaurant, and BAN-ABSA datasets and includes more diversity, the same algorithms (e.g. CNN, BiLSTM) scored higher in this dataset.

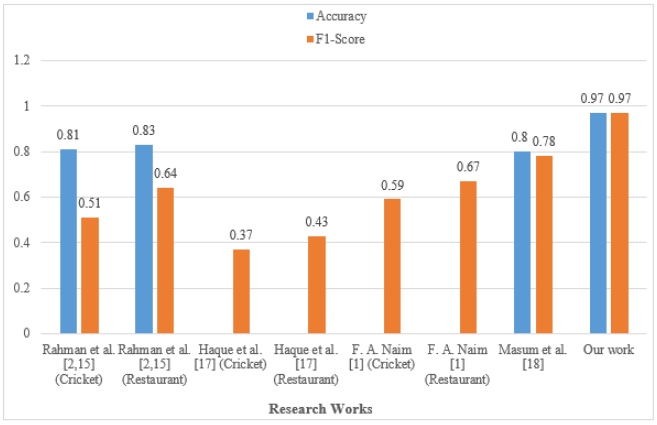

Fig. 7. Comparison of best performances of research works for aspect detection

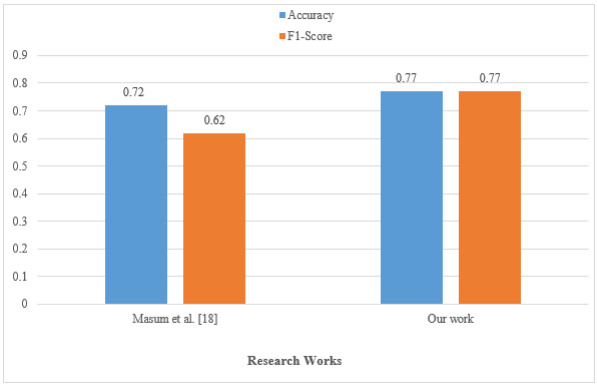

Fig. 8. Comparison of best performances of research works for sentiment classification

V. Conclusion and Future Work

A. Conclusion

Since there are very few ABSA research works on Bengali text currently available, more Bengali language research in this field is still needed. Our model to perform ABSA was built by applying BERT, which is known for getting over the limitations of the algorithms used in prior Bengali ABSA works. Also, a new Bengali dataset was introduced that is annotated with five aspect categories (Technology, Politics, Sports, Economy, and Corruption) and three sentiment categories (Negative, Positive, and Neutral), and these aspects and categories are particularly popular in our everyday life. After comparing it to five additional models that were applied in this study, it is evident that our BERT-based model outperformed the others. Due to BERT's ability to understand words in relation to each other in a phrase, which CNN, RNN, GRU, LSTM, and BiLSTM lack, it received the highest score in both subtasks. For aspect detection, BERT accurately detected 97% of the aspects and the F1-Score was 0.97. As for sentiment analysis, the F1-Score was 0.77 and accuracy was 77%.

B. Future Work

Our future research will be concentrated on researching hybrid methodologies, which will integrate a number of models and practices to improve the efficacy of this research. The models used in this study adhere to supervised techniques. We'll use unsupervised techniques like Latent Dirichlet Allocation (LDA) in the future as well. We will focus more on reducing time complexity and memory space requirements. To more effectively train our models, we will conduct our study in the future using more sophisticated equipment. In order to enhance this dataset and add further preparation steps, we will add extra data before training models. Romanized Bengali and code-switched Bengali data will also be added to increase the diversity of the dataset. The sentiments can be classified into a wider variety of human emotions, including joy, sorrow, hate, and anxiety.

References

[1] F. A. Naim, "Bangla Aspect-Based Sentiment Analysis Based On Corresponding Term Extraction". International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2021.

[2] M. A. Rahman, E. K. Dey, "Aspect Extraction from Bangla Reviews using Convolutional Neural Network." 7th International Conference on Informatics, Electronics & Vision and 2nd International Conference on Imaging, Vision & Pattern Recognition. 2018.

[3] A. Ahmed, M. A. Yousuf, "Sentiment Analysis on Bangla Text Using Long Short-Term Memory (LSTM) Recurrent Neural Network." Proceedings of International Conference on Trends in Computational and Cognitive Engineering. 2020.

[4] R. R. Subramaniam, N. Akshith, "A survey on sentiment analysis." 11th International Conference on Cloud Computing, Data Science and Engineering, 2021.

[5] R. F. Alhujaili, W. M.S. Yafooz, "Sentiment Analysis for Youtube Videos with User Comments: Review." International Conference on Artificial Intelligence and Smart Systems (ICAIS). 2021.

[6] M. J. Althobaiti, "BERT-based Approach to Arabic Hate Speech and Offensive Language Detection in Twitter: Exploiting Emojis and Sentiment Analysis." (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 13, No. 5. 2022.

[7] A. Suciati, I. Budi,"Aspect-Based Sentiment Analysis and Emotion Detection for Code-Mixed Review." (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 9. 2020

[8] T. Nijhawan , G. Attigeri, T. Ananthakrishna, "Stress detection using natural language processing and machine learning over social interactions." Journal of Big Data 9, Article number: 33, 2022.

[9] B. Xing, I. W. Tsang, "Out of Context: A New Clue for Context Modeling of Aspect-Based Sentiment Analysis." Journal of Artificial Intelligence Research 74, 2022.

[10] R. Basri, M.F. Mridha, M. A. Hamid, M. M. Monowar, "A Deep Learning based Sentiment Analysis on Bang-lish Disclosure." National Computing Colleges Conference (NCCC), 2021.

[11] K. I. Islam, M. S. Islam, M. R. Amin, "Sentiment analysis in Bengali via transfer learning using multi-lingual BERT." 23rd International Conference on Computer and Information Technology (ICCIT). 19-21 December. 2020.

[12] B. Liu, "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies, Vol. 5, No. 1. 2012

[13] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, et al, "SemEval-2014 Task 4: Aspect Based Sentiment Analysis." Association for Computational Linguistics, pp 27-35. 2014.

[14] Akbar Karimi, Leonardo Rossi, Andrea Prati, "Adversarial Training for Aspect-BasedSentiment Analysis with BERT." 25th International Conference on Pattern Recognition (ICPR). 2020.

[15] M. A. Rahman, E. K. Dey, "Datasets for aspect-based sentiment analysis in bangla dataset." MDPI Journals, 2018, doi: 10.3390/data3020015.

[16] M. Bodini, "Aspect Extraction from Bangla Reviews Through Stacked Auto-Encoders." MDPI Journals, 2019, doi:10.3390/data4030121.

[17] S. Haque, T. Rahman, A. K. Shakir, M. S. Arman, K. B. B. Biplob, F. A. Himu, and, D. Das, "Aspect Based Sentiment Analysis In Bangla Dataset Based On Aspect Term Extraction." 2nd International Conference on Cyber Security and Computer Science, 2020.

[18] M. A. Masum, S. J. Ahmed, A. Tasnim, M. S. Islam, "BAN-ABSA: An Aspect-Based Sentiment Analysis dataset for Bengali and it's baseline evaluation." Proceedings of International Joint Conference on Advances in Computational Intelligence, pp 385–395. 2021.

[19] P. B.Weerakody,K. W. Wong, G. Wang, W. Ela, "A review of irregular time series data handling with gated recurrent neural networks." Neurocomputing, pp 161-178. 2021.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding." NAACL-HLT, pp 4171–4186.2019.

[21] J. Alawad, B. Alburaidi, A. Alzahrani, F. Alflaj, "A Comparative Study of Stand-Alone and Hybrid CNN Models for COVID-19 Detection." (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 5. 2021.

[22] A. Flores, H. Tito, D. Centty, "Recurrent Neural Networks for Meteorological Time Series Imputation." (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 3. 2020.

[23] S. Alhagry, A. A. Fahmy, R. A. El-Khoribi, "Emotion Recognition based on EEG using LSTM Recurrent Neural Network." (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10. 2017.

[24] M. Chihab, M. Chiny, N. M. H. Boussatta,Y. Chihab, M. Y. Hadi, "BiLSTM and Multiple Linear Regression based Sentiment Analysis Model using Polarity and Subjectivity of a Text." (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 13, No. 10. 2022.

[25] Z. Vujovic, "Classification Model Evaluation Metrics." (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 6. 2021.

[26] K. Ghosh, A. Senapati, "Technical Domain Classification of Bangla Text using BERT." Proceedings of Intelligent Computing and Technologies Conference (ICTCon2021). 2021.

[27] T. H. V. Phan, P. Do, "BERT+vnKG: Using Deep Learning and Knowledge Graph to Improve Vietnamese Question Answering System." (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 7. 2020.

[28] A. Samih, A. Ghadi, A. Fennan, "ExMrec2vec: Explainable Movie Recommender System based on Word2vec." (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 8. 2021.