Journal of
Information Systems Engineering
and Business Intelligence

Vol.9, No.2, October 2023
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Advancement in Bangla Sentiment Analysis: A Comparative Study of Transformer-Based and Transfer Learning Models for E-commerce Sentiment Classification

Zishan Ahmed [1)], Shakib Sadat Shanto [2)], Akinul Islam Jony [3)*]

[1)2)3)]*Department of Computer Science, American International University - Bangladesh, Dhaka, Bangladesh*

[1)] zishanahmed599@gmail.com, [2)] shakibsss080@gmail.com, [3)] akinul@aiub.edu

Abstract

Background: As a direct result of the Internet's expansion, the quantity of information shared by Internet users across its numerous platforms has increased. Sentiment analysis functions at a higher level when there are more available perspectives and opinions. However, the lack of labeled data significantly complicates sentiment analysis utilizing Bangla natural language processing (NLP). In recent years, nevertheless, due to the development of more effective deep learning models, Bangla sentiment analysis has improved significantly.
Objective: This article presents a curated dataset for Bangla e-commerce sentiment analysis obtained solely from the "Daraz" platform. We aim to conduct sentiment analysis in Bangla for binary and understudied multiclass classification tasks.
Methods: Transfer learning (LSTM, GRU) and Transformers (Bangla-BERT) approaches are compared for their effectiveness on our dataset. To enhance the overall performance of the models, we fine-tuned them.
Results: The accuracy of Bangla-BERT was highest for both binary and multiclass sentiment classification tasks, with 94.5% accuracy for binary classification and 88.78% accuracy for multiclass sentiment classification.
Conclusion: Our proposed method performs noticeably better classifying multiclass sentiments in Bangla than previous deep learning techniques.

Keywords: Bangla-BERT, Deep Learning, E-commerce, NLP, Sentiment Analysis

Article history: Received 26 June 2023, first decision 22 September 2023, accepted 3 October 2023, available online 28 October 2023

I. INTRODUCTION

Sentiment refers to individuals' general attitude toward a particular interaction [1]. Sentiment Analysis involves comprehending people's opinions. A text's sentiment may be either positive or negative [2]**.** On occasion, a neutral viewpoint is also considered for classification. In recent years, the field of natural language processing (NLP) has given an important focus on sentiment analysis [3, 4]**.** Due to the rapid development of e-commerce platforms, it is now essential for businesses to understand the opinions that customers have regarding a product through customer reviews. Monitoring public sentiment and evaluating user feedback on well-known e-commerce websites such as 'Daraz,' 'Bikroy,' and 'Chaldal.com' are examples of the diverse applications of a powerful Bangla sentiment analysis system [5, 6]. Companies can use sentiment analysis to obtain insight into customer satisfaction, identify areas for development, and make data-driven decisions to improve their products and services.

However, sentiment analysis comes across unique challenges when it comes to languages with limited resources, such as Bangla [7]. The lack of resources, such as annotated datasets and language parsers, limits the accurate analysis of sentiments in Bangla text. Despite these challenges, recent advancements in machine learning and deep learning methods have shown promise in overcoming these difficulties [8]. In recent years, the field of Bangla sentiment analysis has witnessed significant advancements by utilizing various deep learning methods, in particular RNN-based models like GRU, LSTM and several transformer-based techniques like BERT [9, 10].

In this study, we annotated a dataset of approximately one thousand pure Bangla comments collected from the popular e-commerce platform “Daraz” in both binary (positive, negative) and multiclass (very positive, positive, negative, and very negative) categories. We collected the user reviews from “Daraz” since it has a wide range of products accessible compared to the other e-commerce websites, and so the collected reviews contain greater

* Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.9.2.181-194

variations of user emotion. The dataset consists of 1,000 pure Bangla comments on multiple product types. All the comments were collected and annotated manually. There are 2 attributes or features: Comment and Sentiment. Using our curated dataset, we then assess user sentiment across binary and multiclass classes. Binary and multiclass classifications were implemented and compared to highlight performance differences across deep learning architectures. The analysis specifically aimed to evaluate the models' ability to represent the linguistic complexity and grammatical nuances of the Bangla language under both binary and multiclass settings. This study addresses the need for more Bangla-specific research and resources in sentiment analysis. Several prior works have applied transformers like BERT and Bangla-BERT for binary sentiment classification in Bangla [17,18]. They show good accuracy in binary classification but do not explore multiclass sentiment analysis. Similarly, RNN-based transfer learning approaches like LSTM and GRU have also been commonly applied for Bangla sentiment analysis [11,12,13]. These studies are mostly limited to a binary classification of positive vs. negative sentiment. Additionally, there aren't many research that compare the effectiveness of models based on transfer learning and transformers in the same context. So, this study aims to compare the performance of the popular RNN-based models GRU and LSTM with the highly effective Transformer-based model Bangla-BERT designed specifically for the Bangla language in classifying both binary and multiclass user sentiments. The research objective also includes improving the accuracy of existing multiclass sentiment classification algorithms for the Bangla language using deep learning techniques.

This study contributes by the curation of a well-balanced Bangla e-commerce dataset for both Binary and Multiclass sentiment classification, which is made publicly available for further research and improvements. This paper provides a comparative performance analysis of fine-tuned Transfer learning models (GRU, LSTM) and Transformer based method (Bangla-BERT). Finally, we present an improved accuracy of existing multiclass Bangla sentiment classification.

The paper consists of the following sections: Section 2, or the literature review section, presents a comprehensive review of existing studies on sentiment analysis in the Bangla language and e-commerce. Section 3 is divided into three parts: The dataset description, which covers data collection, labeling, preprocessing, and analysis, then the Experiment section, which includes embedding techniques, transfer learning-based models, transformer-based models, and the optimization of hyperparameters. Finally, Evaluation measures are discussed within this section as well. Section 4 presents the findings of the experiments, with separate evaluations for both multiclass and binary models. Section 5 offers insights and implications drawn from the results, as well as a discussion of comparison and limitation of used approaches. Finally, Section 6, or the Conclusion section, summarizes the study's key findings and outlines potential directions for future research in the domain of Bangla sentiment analysis for e-commerce.

II. Literature Review

The classification of sentiment in Bangla texts using the deep learning method has been a research priority. Deep learning approaches generally outperform conventional machine learning methods [8]. In their work, the authors present a framework for analyzing sentiments in Bangla-written texts [11]. They construct a classification model using Bangla comments. A neural network variant, Convolutional Neural Network, generates the model. The classification accuracy of the classifier model is 99.87%, which is 6.87% higher than the current state-of-the-art Bangla sentiment classifier.

Specifically, deep recurrent models are frequently used for Bangla sentiment classification tasks among the deep learning techniques [12]. The authors propose a technique based on deep learning to classify Bangla restaurant reviews [13]. They employed a dataset consisting of 8,435 reviews written in native Bangla, the majority of which were either positive or negative. They compared their method to other machine learning techniques. Their proposed LSTM model had the highest accuracy at 91.35 percent. In their study, the authors created a system to categorize online food reviews based primarily on positive and negative user sentiment labels [14]. They gathered over a thousand Bangla Food reviews from various online resources, including Foodpanda, HungryNaki, Shohoz Food, and Pathao Food. They preprocessed the data and tested machine learning techniques using Count Vectors, TF-IDF, and N-gram. Among all machine learning and deep learning techniques, LSTM, a deep learning model with word2sequence feature extraction, provided the highest accuracy at 90.89 percent. After refining a massive dataset, the authors used a set of 6,600 data for sentiment analysis in Bangla [15]. Their study's objective was to develop a sentiment classification framework to evaluate the performance of various deep learning models with various parameter calibration combinations. With an accuracy of 94%, the proposed LSTM model with advanced layers attained superior performance in resolving the sentiment polarity of the targeted entities. In their research, the authors compiled a dataset of 10,000 cricket-related comments in Bangla, categorizing them as positive, negative, or neutral [16]. Then, for the vectorization of each word, they used the word embedding method, and their proposed system using LSTM achieved an accuracy of 95% higher than the accuracy of all previous methods.

The use of Transformers models to categorize Bangla sentiments has also gained popularity among researchers. In their study, the authors employed Bangla-Bert to classify sentiment on binary classes [17]. Their approach significantly outperforms all embeddings and sentiment classification algorithms for binary classes. In their research, the authors classify Bangla fake news using Bangla-BERT, Sentiment Analysis on Bengali News Comments, and Cross-lingual Sentiment Analysis in Bengali [18]. The authors fine-tune multilingual transformer models for classification tasks on Bangla text in several areas, such as sentiment analysis, emotion detection, news categorization, and authorship attribution [19]. On six benchmark datasets, they achieve cutting-edge outcomes, outperforming the prior findings by 5-29% accuracy in various tasks. In this study, the authors employ various deep learning techniques to perform 2-class and 3-class sentiment analysis on Bangla text [20]. BERT-GRU performed the best, achieving an accuracy of 60% on their 3-class dataset and 71% on their 2-class dataset.

According to our background study, insufficient Bangla datasets with balanced annotations for E-commerce product reviews make it difficult to compare and classify binary and multiclass sentiments for business understanding and other use cases. This study aims to create a dataset of e-commerce product reviews and compare the performance of Transfer learning models and Transformer-based models in classifying user sentiment across binary and multiclass categories.

III. METHODS

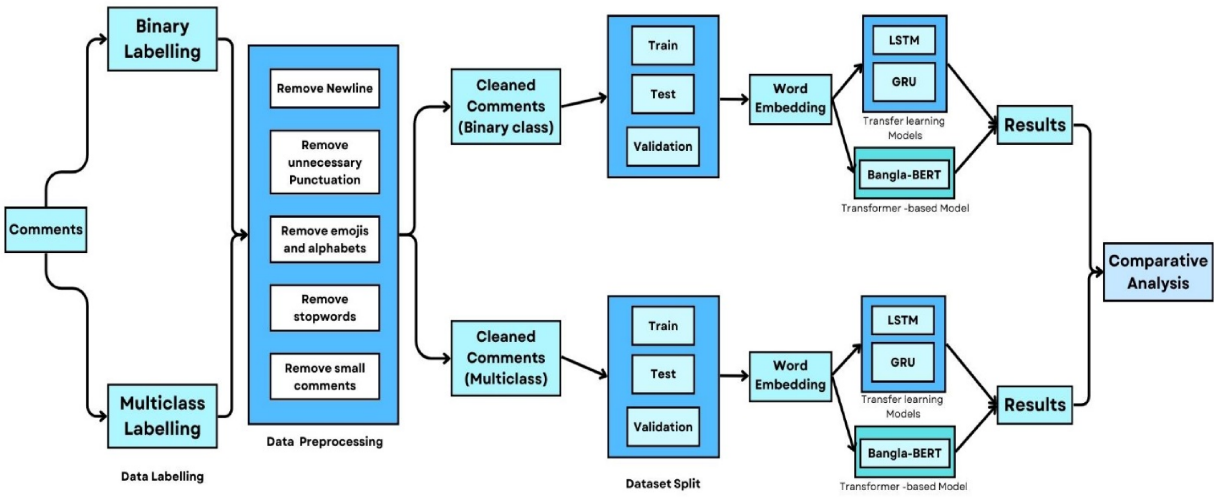


Fig. 1 Proposed System Design

Fig. 1 depicts the system architecture of our study, highlighting the various phases of our research. We initially gathered comments by creating a relevant dataset, which were then labeled for binary and multiclass classification tasks. The comments were then subjected to a phase of preprocessing to ensure data quality, providing a set of clean comments suitable for analysis with both classification approaches. Next, the dataset was divided into training, testing, and validation subsets to facilitate a reliable evaluation of models. We utilized advanced word embedding techniques to enhance the semantic representation of the comments. Then, transfer learning and transformer-based models were used to train and assess the efficacy of both classification approaches. Finally, we compared the outcomes of the various models.

A. Dataset Description

In our research, we collected 1000 pure Bangla comments from 'Daraz.' For our study, we collected data, manually annotated the dataset, preprocessed it, and then analyzed it. Each of these procedures was essential in helping us extract useful information from the dataset, thus allowing us to look more closely at how they worked.

1) Data Collection

Online purchasing in Bangladesh has increased dramatically over the past several years, leading to the explosive expansion of the country's e-commerce sector. In an ever-changing environment, a firm's ability to read and respond

to customer sentiments and views is crucial for success. In order to effectively utilize the insights provided by our customers, we undertook a tedious data-gathering approach to gather a sizable collection of Bangla reviews. Our data collection process was designed to ensure a diverse representation of product types within the dataset. Recognizing the importance of capturing a broad range of customer experiences across various product categories, we made concerted efforts to collect comments from a wide variety of product types available on the e-commerce platform.

We concentrated our data gathering efforts on 'Daraz,' a well-known e-commerce site in Bangladesh, to ensure the accuracy and clarity of the information we collected. Daraz's massive user base and wide selection of items made it a great place to get authentic, representative reviews written in Bangla. We chose this platform to record the vast spectrum of user and consumer sentiments and experiences.

Carefully, we read every single one of the thousands of customer reviews that were posted on Daraz. We took great care only to include reviews written entirely in pure Bangla since we wanted to be sure that the language used truly reflected the sentiments and experiences of those who speak Bangla as their native tongue. We conveyed the complexity and subtlety of customer comments thanks to Bangla's emphasis on purity.

We used rigorous criteria in the selection process to ensure that the reviews we collected were genuine and of high quality. The criteria were: giving extra weight to reviews that avoided spammy or repetitive language, collecting comments that expressed themselves clearly and comprehensively, and collecting comments from a wide range of product types such as fashion, electronics, household stuff, clothing, beauty products, etc. We focused our study on a high-quality dataset that accurately reflected the varied opinions and sentiments of Bangla-speaking consumers by eliminating irrelevant and low-quality reviews.

2) Data Labelling

The accurate labeling of the collected dataset is an essential stage in sentiment analysis, as it sets the groundwork for training effective machine learning models. In our study, we assigned sentiment labels to reviews based on the thoughts and sentiments conveyed in each comment. The classification process consisted of four categories: positive, very positive, negative, and very negative, capturing a wide range of user sentiments.

For binary classification, we merged the positive and very positive categories into a single label, 'Positive.' Similarly, the negative and very negative categories were merged under the label 'Negative.' This consolidation simplified the classification task and aligned with the dataset's objective of distinguishing between positive and negative sentiments.

We assigned each review's sentiment labels ourselves to ensure consistency and dependability in our labeling. We, the authors of this study, were involved in annotating the dataset. As we, the authors, are all native Bengali speakers, the contextual meaning of the comments was properly understandable to us, which made the annotation process more precise. The annotated comments were cross-checked by each author to avoid inconsistencies or errors. This process of cross-checking helps ensure the quality and accuracy of the annotations and promotes consistency among the annotators. We initially labeled the dataset according to the multiclass labeling format. As annotators, we used the below criteria for labeling each comment:

TABLE 1
CRITERIA FOR LABELLING COMMENTS

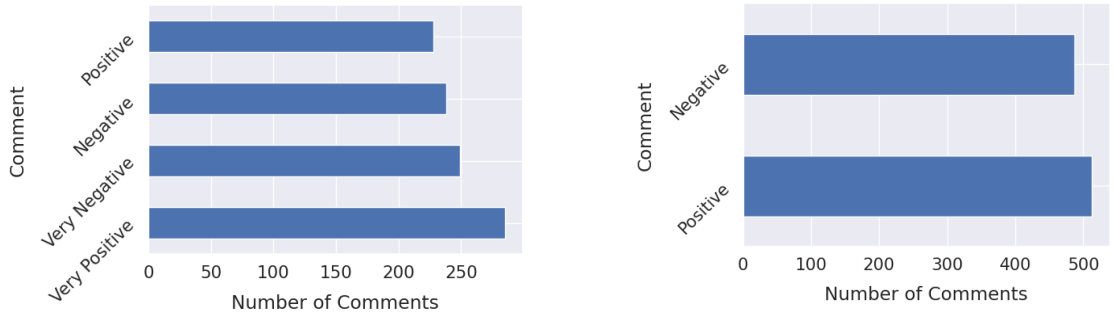
Comment Type	Criteria
Positive	Conveys mild satisfaction, praise, and appreciation
Negative	Conveys mild dissatisfaction, anger, negative emotions, disappointment
Very Positive	Conveys a strong degree of positivity, praise, endorsement, and enthusiasm
Very Negative	Conveys strong degree of negativity, slang, harsh language, condemnation

The criteria shown in Table 1 were used by the annotators while annotating the comments. Since each of our collected comments held a distinctive tone that represented a particular sentiment, the manual labeling with a set of common criteria made the labeling more precise. This strategy allowed us to analyze the content of each comment in detail, discern the underlying sentiments, and designate labels accordingly. By manually labeling the dataset, we maintained high control and precision in capturing the reviewers' intended emotions and thoughts. The table 2 shows the labelling examples for both multiclass and binary classification process.

Fig. 2 demonstrates the distribution of the collected comments in the preceding graph. As evidenced by the data, we maintained relatively equal-sized comment categories during the data collection phase. The process involved collecting equal comments for each sentiment category, including positive, very positive, negative, and very negative. While collecting the comments, we made sure comments from all kinds of sentiments were equally taken. By ensuring a proportionate representation of sentiments, we aimed to reduce any potential bias resulting from an overabundance of one sentiment category.

TABLE 2
COMMENT LABELLING SAMPLES

Original Comment	English Translation	Multiclass Sentiment	Binary Sentiment
মধু আলহামদুলিল্লাহ মোটামুটি ভালো লাগছে ❤️❤️	The Honey feels pretty good Alhamdulillah ❤️❤️	Positive	Positive
অনেক উপকার পাচ্ছি 😍, সেলার ন্যাচারাল আইটেম দিয়ে বানিয়েছেন কথাটা সত্যি, ঘ্রানটাও অনেক সুন্দর।	I am getting a lot of benefits 😍, it is true that the seller has made it with natural items, the smell is also very nice.	Very Positive	Positive
ছবি দেখে মনে করেছিলাম একটু বড় হবে। ফ্যানটি ছোট 😔।	Looking at the picture, I thought it would be a little bigger. The fan is small 😔.	Negative	Negative
খুবই খারাপ অভিজ্ঞতা 😡😡😡😡😡 দারাজ মলের থেকে কিনলাম কিন্তু খাবারের সাথে সাবানের গুঁড়ো এক বক্স আশা করিনি।	Very bad experience 😡😡😡😡😡 bought from Daraz mall but didn't expect a box of soap powder with the food.	Very Negative	Negative



(a) Distribution of multiclass comments

(b) Distribution of binary class comments

Fig. 2 The distribution of both categories (a) Multiclass and (b) Binary comments in the collected dataset

3) Data Preprocessing

The preprocessing of the collected dataset was a key step in getting the text data ready for the next step, which was the classification of sentiment. First, we had to get rid of punctuation, numbers, emoji, pictorial icons, and letters that could have added more complexity to the classification of sentiment. By getting rid of these parts that were not important, we aimed to improve the text data and focus on the most important information for research.

Initially, sentiment classification was done with separate lists of stop words. These stop-word lists consisted of commonly used words that do not convey significant sentiment or category-specific details. To ensure the effectiveness of stop-word filtering, we employed a novel approach. Instead of only using existing stop-word lists, we made our own by finding the words that were used the least in the dataset. With this method, we could include domain-specific stop words that might not be on general lists. This process made our filtering process more accurate.

After stop words were removed, the stemming technique was used to get similar words with different endings back to their basic forms. This method lowers the number of dimensions in the dataset and ensures that words with similar meanings are looked at as a single unit.

TABLE 3
SAMPLES OF MINOR REVIEWS

Minor Reviews	English Translation
কাটার মানের	Cutting quality
শাড়িটার সুতা	Saree thread
চশমাটা ভাই	The glasses are brother
ব্যাগ ছিলো	There was bag
মেট্রেস নিঃসন্দেহে	The mattress, of course

Moreover, minor reviews that did not substantially contribute to the aggregate sentiment or category classification were eliminated from the dataset. The minor reviews are those that had less than 3 words in them and provided no precise sentiment. They could also be referred to as short or inconclusive reviews. This step intended to eliminate noise and ensure that the dataset only contained reviews that provided meaningful insights and representative

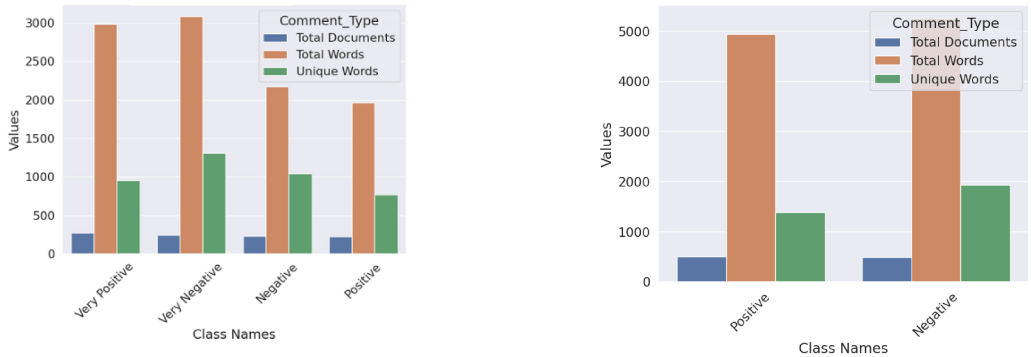
expressions of sentiment. Eliminating minor reviews improved the quality and dependability of the dataset, ensuring that the subsequent analysis was centered on reviews that contained considerable data. Table 3 shows a few examples of some minor reviews. The resulting preprocessed, few examples of sentiment-classification-ready comments are displayed in Table 4.

TABLE 4
PREPROCESSED COMMENT SAMPLES

Original Comment	English Translation of Original Comment	Cleaned Comment	English Translation of Cleaned Comment
এক কথাই অসাধারণ একটা শাড়ি❤❤❤যা বলার মতো কথা সাথে কাজে ১০০% মিল আছে আপনার এই চাইলে চোখ বন্ধ করে বিশ্বাসের সাথে শাড়ি টা নিতে পারে ধন্যবাদ❤❤ ধারাজ ধন্যবাদ❤❤ সেইলার	One word is a wonderful saree❤❤❤ which is 100% matching with words and actions, if you want you can close your eyes and take the saree with faith. Thank you❤❤ Thank you very much❤❤ Seller	এক কথাই অসাধারণ একটা শাড়ি ধন্যবাদ সেইলার	One word is a wonderful saree thanks Seller
কম দামে ভালো একটা প্রোডাক্ট। ❤️ ❤️ দাম অনুযায়ী অনেক ভালো ছিলো, আজি হাতে পেলাম, ১১.১১ তে অর্ডার ছিলো। মাত্র ৪৮০/- টাকা পরলো ২ টা 🤟 🤟 ধন্যবাদ সেলার কে। ❤️ ❤️ বালিস মাথায় দিয়ে রিভিউ দিচ্ছি😁 😁	A good product at a low price. ❤️ ❤️ It was very good according to the price, got it today. The order was on 11.11. Only Tk 480/- wore 2 🤟 🤟 Thank you seller. ❤️ ❤️ I am giving a review with a pillow on my head 😁 😁	কম দামে ভালো একটা প্রোডাক্ট দাম অনুযায়ী অনেক ভালো ছিলো, আজি হাতে পেলাম ধন্যবাদ সেলার কে	A good product at a low price It was very good according to the price, I received it today, thanks to the seller
আলহামদুলিল্লাহ 😘 মোজা গুলো পছন্দ হইছে, মানে যেমন দেখে অর্ডার করছি তেমনি পেয়েছি। 😍 😍 শুকরিয়া 🤍	Alhamdulillah 😘, I like the socks. I mean, I received it as I ordered it. 😍 😍 Thank you 🤍	আলহামদুলিল্লাহ মোজা পছন্দ হইছে যেমন অর্ডার করছি তেমনি পেয়েছি শুকরিয়া	Alhamdulillah, I like the socks, I received them as ordered, thank you

4) Data Analysis

In this section, we delve into the analysis of the preprocessed data to gain valuable insights into customer sentiments and preferences. We examine various aspects of the dataset, including sentiment distribution, comment lengths, and lexical diversity.



(a) Data statistics for multiclass comments (b) Data statistics for binary class comments

Fig. 3 Data statistics for both category (a) Multiclass and (b) Binary comments in the preprocessed dataset

In Fig. 3 the data statistics for both multiclass and binary class comments are visualized. Here, the total number of comments is represented by 'Total Documents'. The total number of words in each sentiment class is represented by 'Total Words', and the total number of unique words is represented by 'Unique Words'. Unique words mean the number count of distinct words from the total words of each sentiment class.

It is observed in Fig. 3 that the total word count is higher for the very negative and very positive classes in the multiclass sentiment classification. In contrast, the negative and positive classes exhibit lower counts. In the binary classification, the total word count for both classes are comparable.

Another significant observation is that the number of unique words is lower than the total word count in each class for both classifications. This reduced lexical diversity can be beneficial when dealing with misspelled or inappropriate words present in the corpus. Fewer unique words make it easier for models to handle such variations in the text and contribute to more efficient computation and memory utilization by reducing the dimensionality of the vector space representation. However, it is worth noting that the very positive and positive classes have fewer unique words than the very negative and negative classes. This limited lexical diversity may introduce ambiguity, particularly if multiple words are assigned to the same vector representation.

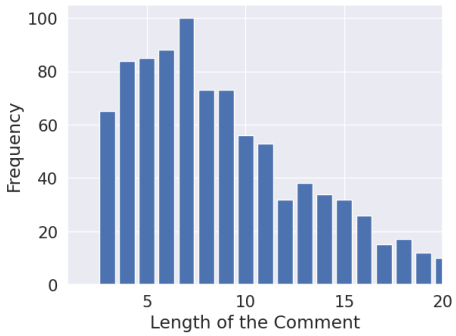


Fig. 4 Length-frequency distribution of the preprocessed dataset

The analysis of the comment lengths relative to the number of comments, as illustrated in Fig. 4, reveals an interesting distribution pattern. The majority of the comments fall within the range of two to nine words in length. However, there needs to be more lengthy comments in the dataset because lengthier comments would have added more word variation to the dataset and lessened the number of minor comments after preprocessing.

In summary, the data analysis after preprocessing demonstrates that the preprocessed dataset primarily consists of comments with shorter lengths, ranging from two to nine words. The presence of significantly fewer lengthy comments suggests a tendency for concise expressions among users. Though general and domain-specific stop-word removal made most comments shorter in length, they conveyed the clear sentiment of the reviewer.

B. Experiment

This experiment's primary objective is to compare the performance of Transfer Learning-based and Transformer-based models for classifying binary and multiclass sentiment in Bangla using fine-tuning. We conducted experiments using Google Collaboratory, which is extensively used for developing deep learning applications. We used TensorFlow == 2.12.0, Keras == 2.12.0, and Transformers == 4.30.2 in order to develop our deep-learning models. After preprocessing the dataset and removing small reviews, the final dataset contains a total of 979 cleaned reviews. We split the preprocessed dataset into training, test, and validation datasets. The training set should be the largest split as it is used to actually train the models. A bigger training set allows the models to learn more robustly and generalize better. 72% (704 reviews) for training ensures sufficient data for this purpose. The validation set is used during training to tune hyperparameters and evaluate model performance on data not seen during training. The validation set contains 18% (177 reviews), which is a reasonable size - large enough to get a good estimate of model generalization but not so large it significantly reduces the training data. The test set is used only for the final evaluation of the fully trained model. It should be a representation of real-world unseen data. The test set contains 10% (98 reviews) of total reviews. 10% is a commonly used ratio for testing in machine learning or deep learning. Using an approximate of 70/20/10 ratio is a commonly accepted rule of thumb for splitting datasets into train/validation/test. So, the percentages are aligned with standard practices.

1) Embedding

Word embedding is a metric for language modeling and feature learning in neural networks that converts textual words into dense, low-dimensional vectors. The Word2vec embedding approach is used to extract the feature. Word2Vec has achieved considerable success in sentiment analysis in several languages, including Bengali [21, 22]

2) Transfer learning Models

In our study, we use RNN-based Transfer learning models GRU and LSTM. A form of recurrent neural network (RNN) layer called the GRU (Gated Recurrent Unit) layer is ideally suited for processing sequential input, such as text data [23, 24]. Long-term dependencies in sequential data are captured by the LSTM (Long Short-Term Memory)

layer, a form of recurrent neural network (RNN) layer that is created to address the vanishing gradient problem [25, 26]

Each word in the input set is first mapped to a dense vector representation in the model's first embedding layer. The vocabulary words ' low-dimensional representations (embeddings) are learned and assigned at this layer. Next comes the Bidirectional GRU Layer for GRU and the Bidirectional LSTM layer for the LSTM approach. The model may consider the context of both the words that come before and after the input sequence because of the bidirectional GRU layer [27, 28]. The bidirectional long short-term memory (LSTM) layer is a recurrent layer that employs three separate gates (input gate, forget gate, and output gate) to record long-term associations and retain information in a memory cell [16, 29]. The output of the GRU and LSTM layers is subjected to an application of a dense layer with 24 units and the ReLU activation function. This layer receives the output from the GRU and LSTM layers and extracts additional features and non-linear transformations. The output from the previous dense layer must be flattened since the following dense layer requires a one-dimensional tensor. The last dense layer uses a softmax activation function with two units for binary and four for multiclass, depending on the number of emotion categories. Constructing a probability distribution across the categories calculates the probability of each sentiment category for a given input. The multiclass capable loss function of sparse categorical cross-entropy was used to create this model. The Adam optimizer is used in gradient descent.

3) Transformer-based Model

We employed the Bangla-BERT transformer-based model, which was pre-trained with 2.18 billion tokens [18, 30] because of BERT's recent success in various NLP tasks [31, 32].

A language model for the Bangla language called Bangla-BERT (Bidirectional Encoder Representations from Transformers) is based on the Transformer architecture [17]. The Bangla-BERT model's design is similar to the original BERT model; however, Bangla-BERT was trained on a sizable corpus of Bangla text to capture language-specific patterns and semantic representations. In order to handle words that are not recognized and gather details at a subword level, Bangla text is broken down into WordPieces, which are smaller units of words. The Transformer layers, or several self-attention and feed-forward neural networks, comprise the model. Each layer engages in self-attention, which enables the model to focus on different elements of the input sequence and recognize contextual relationships between words. The model can evaluate the relative weights of various words in the input sequence, capturing both local and global dependencies, thanks to the self-attention mechanism within each Transformer layer. The input embeddings are subjected to positional encoding to convey the positioning information of the words in the sequence and help the model comprehend the sequential order of the words.

4) Hyperparameters

Hyperparameters are beneficial for determining any model's network architecture and key characteristics, including the number of layers and nodes. It is also crucial for network training and how it is trained based on learning rates, batch size, and dropout rates [33]. The hyperparameter settings for the proposed study are displayed in Table 5. The entire hyperparameter space was explored to determine the optimal value of each hyperparameter. As demonstrated in Table 5, these models for our proposed study are trained with optimal hyperparameter values.

TABLE 5
HYPERPARAMETER SETTINGS

Hyperparameters	Hyperparameter Space	Optimal Value
Batch Size	16,32,64,128	64
Embedding Dimension	8,16,32,64,128,256	128
Dropout Rate	0.1,0.15,0.2,0.25,0.30,0.35,0.40,0.45,0.5	0.30
Optimizer	SGD, Adam	Adam
Learning Rate	0.6,0.3,0.1,0.01,0.001,0.0001	0.0001
Epochs	10,15,20,25,30	10

C. Evaluation Measures

Deep learning model performance is measured using evaluation metrics. We evaluated the models based on their accuracy, precision, recall, and F1 score.

Accuracy [34] is a performance metric that evaluates how well a machine learning model performs. Which is defined as:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (1)$$

Precision [34] quantifies the ratio of true positives (accurately predicted positive occurrences) to the total number of cases that fell into the positive category. Which is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (2)$$

Recall [34] calculates how many positive actual depictions, according to our criteria, are marked as positive (true positive). Which is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3)$$

The harmonic mean of accuracy and recall, or F1-Score [34], is widely used as a single statistic to evaluate a model's efficacy. The F1-score is defined as:

$$\text{F1} - \text{Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (4)$$

IV. Results

A. Model Evaluations for Multiclass Classification

The result comparison of various deep learning approaches with different evaluation metrics of Multiclass sentiment classification is shown in Table 6.

TABLE 6
MULTICLASS SENTIMENT CLASSIFICATION RESULTS

Approach	Accuracy	Precision	Recall	F1-Score
GRU	84.7	87.6	84.8	86.1
LSTM	85.0	82.0	83.7	85.5
Bangla-BERT	**88.78**	**88.78**	**88.68**	**89.77**

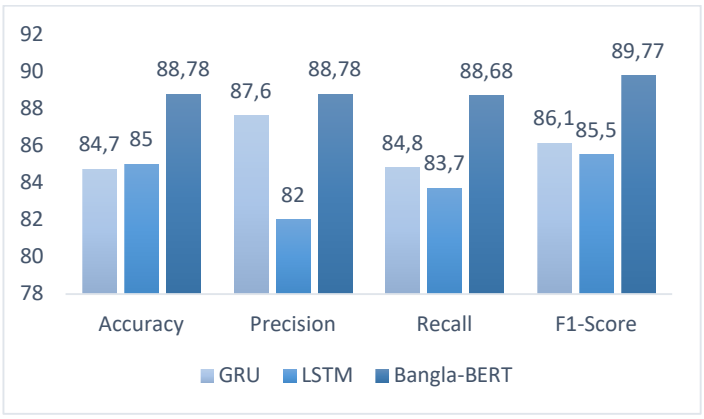


Fig. 5 Performance Comparison of deep learning approaches for multiclass classification

In terms of multiclass sentiment classification, Bangla-BERT outperformed the other models, as shown in Table 6 and Fig. 5. The accuracy of 88.78% demonstrated that the classification of sentiment categories was accurate. Additionally, Bangla-BERT's accuracy of 88.78% demonstrated its ability to classify instances accurately as having the desired sentiment.

Moreover, Bangla-BERT's recall rate of 88.68% indicates that it effectively identifies instances belonging to each sentiment category (positive, negative, very positive, and very negative). The F1-score of 89.77% attained by Bangla-BERT demonstrates its superior performance in balancing precision and recall. While the Bangla-BERT model achieved an accuracy of 88.78%, the LSTM and GRU models only achieved an accuracy of 85% and 84.7%, respectively, with inferior precision, recall, and F1-score.

Results indicate that when accuracy, precision, recall, and F1-score are compared, Bangla-BERT outperforms GRU and LSTM models. Bangla-BERT has proved to be the most effective model for sentiment classification across multiple classes for our dataset.

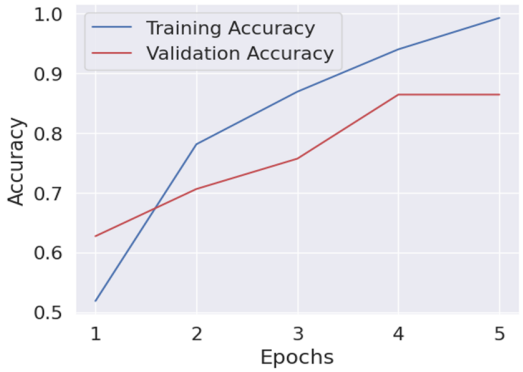


Fig. 6 Training and Validation Accuracy Curve for Bangla-BERT (Multiclass)

Fig. 6 depicts the accuracy of training and validation across various epochs. The number of epochs is displayed on the x-axis, while the accuracy is displayed on the y-axis.

The training accuracy line begins at a lower number and progressively increases with each epoch as the model learns and becomes better at its task. As epochs progress, the training accuracy line gradually approaches 1.0. This trend demonstrates that the model becomes proficient at predicting the correct labels for training data.

On the other hand, the validation accuracy line begins at a lower number, which is typically lower than the training accuracy, because it measures how well the model performs on validation data that it has never seen before. Initially, the accuracy of the validation improves at the same rate as the accuracy of the training, but it gradually increases. The validation accuracy line peaks near 0.90, and this result indicates that the model is accurate in predicting sentiment classes.

B. Model Evaluations for Binary Classification

The result comparison of various deep learning approaches with different evaluation metrics of Binary sentiment classification is shown in Table 7.

TABLE 7
BINARY SENTIMENT CLASSIFICATION RESULTS

Approach	Accuracy	Precision	Recall	F1-Score
GRU	90.97	91.06	91.10	90.97
LSTM	91.67	92.01	91.88	91.67
Bangla-BERT	**94.5**	**94.42**	**94.49**	**94.44**

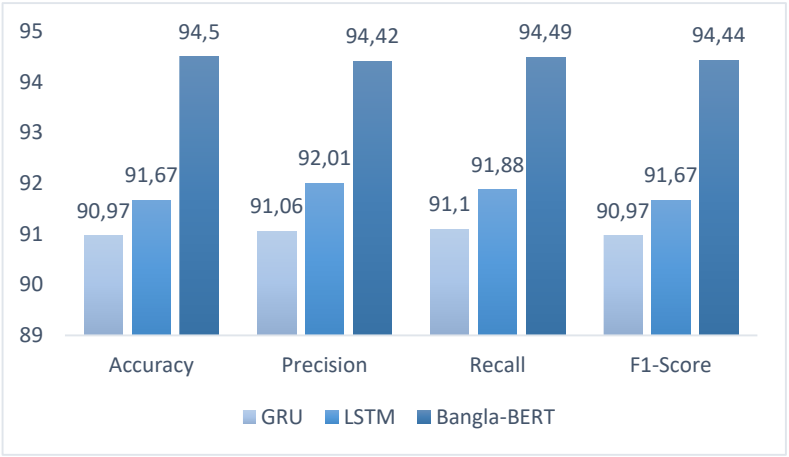


Fig. 7 Performance Comparison of deep learning approaches for Binary classification

Based on the information in the Table 7 and the Fig. 7, Bangla-BERT has the highest overall performance. It exhibits 94.5 percent accuracy, 94.4 percent precision, 94.4 percent recall, and 94.4 percent F1 score. This result indicates that it excels at accurately identifying and categorizing Binary sentiments in our data set.

GRU and LSTM also demonstrate excellent performance but lag slightly behind Bangla-BERT. LSTM performs admirably with 91.67% accuracy, 92.01% precision, 91.88% recall, and 91.67% F1-score. F1-score, precision, accuracy, and recall all add up to 90.97 percent for GRU.

As the results demonstrate, Bangla-BERT performs better than GRU and LSTM at accurately categorizing emotions. Bangla-BERT provides superior sentiment categorization in terms of accuracy, precision, recall, and F1 score owing to its superior capacity to comprehend intricate patterns and dependencies within Bangla text.

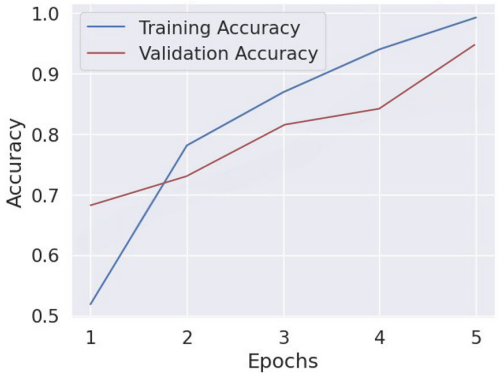


Fig. 8 Training and Validation Accuracy Curve for Bangla-BERT (Binary class)

Fig. 8 illustrates training and validation accuracy across multiple epochs for sentiment classification of binary classes using Bangla-BERT. As epochs progress, training accuracy increases in an ascending pattern. Furthermore, it achieves its highest point near 1. This trend demonstrates that the model performs exceptionally well at accurately predicting labels for training data.

The validation accuracy, on the other hand, begins at a lower level and then increases over time. The validation accuracy curve ends close to 0.95, demonstrating that the model can accurately predict unseen labels.

V. Discussion

A. Insights and Implications

The acquired results provide valuable insights into sentiment classification in the context of the e-commerce industry and the Bangla language. The superiority of Bangla-BERT in both multiclass and binary sentiment classification highlights its effectiveness in accurately categorizing sentiments and capturing intricate patterns in Bangla text. The high accuracy, recall, precision, and F1-score attained by Bangla-BERT emphasize its ability to understand and interpret sentiment expressions across different categories. These findings underline the importance of leveraging advanced transformer-based models for sentiment analysis tasks, enabling businesses to understand customer sentiments comprehensively, enhance product review analysis and make informed decisions to improve customer experiences in the dynamic e-commerce landscape.

Compared to the few studies conducted on multiclass sentiment analysis in Bangla using deep learning, the efficacy of our proposed approach with Bangla-BERT is significantly improved. In our dataset, Bangla-BERT classified multiclass sentiments with 88.78% accuracy, which is higher than previously proposed deep learning methods [15,35,36].

Our findings revealed substantial improvements in categorizing sentiments into two categories. By applying transfer learning and transformers methods on our dataset, we compare the categorization accuracy of Bangla text as positive or negative. Bangle-BERT was the most accurate method for classifying binary sentiment in Bangla texts, with an accuracy of 94.5%.

B. Comparison and Limitations

There are some noteworthy advantages of Bangla-BERT and some drawbacks of RNN-based sequential models for the task of classifying sentiments expressed in the Bangla language. Firstly, GRU and LSTM are recurrent neural network architectures that process sequence data sequentially. This limits their ability to model long-range

dependencies in long text sequences compared to BERT's transformer architecture. RNN-based models like GRU and LSTM can suffer from vanishing gradient problems during training, which can degrade performance. In contrast, BERT uses attention, which alleviates this issue. Self-attention in BERT allows explicitly modeling relationships between non-consecutive words. RNNs implicitly simulate such relationships through their hidden state. Bangla-BERT is pre-trained on a large Bangla corpus, allowing it to have better Bangla language understanding. However, GRU and LSTM lack this pre-trained initialization. BERT tokenizes text into WordPieces, which can better handle misspellings or unseen words compared to word-level tokenization used in GRU or LSTM. Overall, BERT's transformer architecture, pre-training, WordPiece tokenization, bidirectional encoding, and higher parameter capacity allow it to learn better representations of text for sentiment analysis. The limitations of RNN sequentiality, lack of pre-training, word-level tokenization, and smaller size likely contributed to lower performance compared to BERT. The results validate Bangla-BERT's superiority for the Bangla sentiment analysis task over to GRU and LSTM.

VI. Conclusion

This study makes significant contributions to Bangla sentiment analysis in the e-commerce domain. In response to the insufficiency of labeled data in the Bangla language, we have compiled a well-balanced dataset for Bangla e-commerce sentiment analysis using the "Daraz" platform. Our experiments contrasted transfer learning (LSTM, GRU) and transformer-based (Bangla-BERT) sentiment classification approaches. Our results demonstrate that Bangla-BERT performs superiorly in binary and multiclass sentiment classification. With an accuracy of 94.50% for binary classification and 88.78% for multiclass sentiment classification, Bangla-BERT has demonstrated its ability to classify sentiments accurately and capture the nuances of Bangla text.

Regarding future works, there are numerous avenues to investigate. The expansion of the dataset can improve the efficacy and generalizability of sentiment analysis models. In addition, investigating additional transformer-based models and fine-tuning techniques may yield even better outcomes. Exploring ensemble techniques that incorporate multiple models could also be advantageous for enhancing sentiment classification accuracy. In addition, investigating sentiment analysis in other domains and integrating domain-specific knowledge could yield valuable insights for specialized applications.

This study establishes the groundwork for advanced sentiment analysis in Bangla for the e-commerce industry. This research's findings can aid businesses in comprehending and capitalizing on customer sentiments to improve their products, services, and overall customer experiences.

Author Contributions: *Zishan Ahmed*: Conceptualization, Programming, Data Curation, Methodology, Writing - Original Draft. *Shakib Sadat Shanto*: Conceptualization, Programming, Data Curation, Methodology, Writing - Original Draft. *Akinul Islam Jony*: Conceptualization, Methodology, Writing -Review & Editing, Supervision, Data Curation.

All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The data that support the findings of this study are openly available in https://github.com/shakib-sadat/Bangla-E-commerce-Dataset

Informed Consent: There were no human subjects.

Animal Subjects: There were no animal subjects.

ORCID:
Zishan Ahmed: https://orcid.org/0009-0004-9598-917X
Shakib Sadat Shanto: https://orcid.org/0009-0009-8798-9010
Akinul Islam Jony: https://orcid.org/0000-0002-2942-6780

REFERENCES

[1] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093-1113, 2014.

[2] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pp. 579-586, 2005.

[3] A. Adak, B. Pradhan, and N. Shukla, "Sentiment analysis of customer reviews of food delivery services using deep learning and explainable artificial intelligence: Systematic review," *Foods*, vol. 11, no. 10, 1500, 2022.

[4] A. Iqbal, R. Amin, J. Iqbal, R. Alroobaea, A. Binmahfoudh, and M. Hussain, "Sentiment Analysis of Consumer Reviews Using Deep Learning," *Sustainability*, vol. 14, no. 17, 10844, 2022.

[5] S. Zulfiker, A. Chowdhury, D. Roy, S. Datta, and S. Momen, "Bangla E-Commerce Sentiment Analysis Using Machine Learning Approach," in *4th International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pp. 1-5, 2022.

[6] M.J. Hossain, D.D. Joy, S. Das, and R. Mustafa, "Sentiment Analysis on Reviews of E-commerce Sites Using Machine Learning Algorithms," in *International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pp. 522-527, 2022.

[7] K.A. Hasan, S. Islam, G. M.E. Elahi, and M.N. Izhar, "Sentiment recognition from Bangla text," in *Technical Challenges and Design Issues in Bangla Language Processing*, pp. 315-327, 2013.

[8] O. Sen et al., "Bangla Natural Language Processing: A comprehensive analysis of classical, machine learning, and deep learning based methods," *IEEE Access*, vol. 10, pp. 38999-39044, 2022.

[9] N.R. Bhowmik, M. Arifuzzaman, and M.R.H. Mondal, "Sentiment analysis on Bangla text using extended lexicon dictionary and deep learning algorithms," *Array*, vol. 13, 100123, 2022.

[10] M.R. Khan, S.N. Rahmatullah, M.F. Islam, A.R.M. Kamal, and M.A. Hossain, "Sentiment analysis of COVID-19 vaccination in Bangla language with code-mixed text from social media," in *12th International Conference on Electrical and Computer Engineering (ICECE)*, pp. 76-79, 2022.

[11] M.H. Alam, M.M. Rahoman, and M.A.K. Azad, "Sentiment analysis for Bangla sentences using convolutional neural network," in *20th International Conference of Computer and Information Technology (ICCIT)*, pp. 1-6, 2017.

[12] A. Hassan, M.R. Amin, A.K. Al Azad, and N. Mohammed, "Sentiment analysis on bangla and Romanized Bangla text using deep recurrent models," in *International Workshop on Computational Intelligence (IWCI)*, pp. 51-56, 2016.

[13] E. Hossain, O. Sharif, M.M. Hoque, and I.H. Sarker, "Sentilstm: a deep learning approach for sentiment analysis of restaurant reviews," in *International Conference on Hybrid Intelligent Systems*, pp. 193-203, 2020.

[14] M.I.H. Junaid, F. Hossain, U.S. Upal, A. Tameem, A. Kashim, and A. Fahmin, "Bangla Food Review Sentimental Analysis using Machine Learning," in *IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC),* pp. 0347-0353, 2022.

[15] A. Ahmed and M.A. Yousuf, "Sentiment analysis on Bangla text using long short-term memory (LSTM) recurrent neural network," in *Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020,* pp. 181-192, 2020.

[16] M.F. Wahid, M.J. Hasan, and M.S. Alom, "Cricket sentiment analysis from Bangla text using recurrent neural network with long short term memory model," in *International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1-4, 2019.

[17] N.J. Prottasha et al., "Transfer learning for sentiment analysis using BERT based supervised fine-tuning," *Sensors*, vol. 22, no. 11, 4157, 2022.

[18] M. Kowsher, A.A. Sami, N.J. Prottasha, M.S. Arefin, P.K. Dhar, and T. Koshiba, "Bangla-BERT: transformer-based efficient model for transfer learning and language understanding," *IEEE Access*, vol. 10, pp. 91855-91870, 2022.

[19] T. Alam, A. Khan, and F. Alam, "Bangla text classification using transformers," arXiv preprint arXiv:.04446, 2020.

[20] K.I. Islam, M.S. Islam, and M.R. Amin, "Sentiment analysis in Bengali via transfer learning using multilingual BERT," in *23rd International Conference on Computer and Information Technology (ICCIT),* pp. 1-5, 2020.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:. 2013.

[22] M. Al-Amin, M.S. Islam, and S.D. Uzzal, "Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words," in *International Conference on Electrical, Computer and Communication Engineering (ECCE),* pp. 186-190, 2017.

[23] Y. Santur, "Sentiment analysis based on gated recurrent unit," in *International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1-5, 2019.

[24] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:. 2014.

[25] G. Murthy, S. R. Allu, B. Andhavarapu, M. Bagadi, and M. Belusonti, "Text based sentiment analysis using LSTM," *Int. J. Eng. Res. Tech. Res*, vol. 9, no. 5, pp. 299-303, 2020.

[26] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM," *Neural Computing and Applications*, vol. 32, pp. 9713-9729, 2020.

[27] R. Rahman, S. A. Hasan, and F. A. Rubel, "Identifying Sentiment and Recognizing Emotion from Social Media Data in Bangla Language," in *12th International Conference on Electrical and Computer Engineering (ICECE),* pp. 36-39, 2022.

[28] M.M. Abdelgwad, T.H.A. Soliman, A.I. Taloba, and M.F. Farghaly, "Arabic aspect based sentiment analysis using bidirectional GRU based models," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 9, pp. 6652-6662, 2022.

[29] A.A. Sharfuddin, M. N. Tihami, and M. S. Islam, "A deep recurrent neural network with bilstm model for sentiment classification," in *International conference on Bangla speech and language processing (ICBSLP),* pp. 1-4, 2018.

[30] A. Bhattacharjee et al., "BanglaBERT: Language model pre-training and benchmarks for low-resource language understanding evaluation in Bangla," arXiv preprint arXiv:.00204, 2021.

[31] J.D. M.W.C. Kenton and L.K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019.

[32] A. Zhao and Y. Yu, "Knowledge-enabled BERT for aspect-based sentiment analysis," *Knowledge-Based Systems*, vol. 227, 107220, 2021.

[33] G.I. Diaz, A. Fokoue-Nkoutche, G. Nannicini, and H. Samulowitz, "An effective algorithm for hyperparameter optimization of neural networks," *IBM Journal of Research and Development*, vol. 61, no. 4/5, pp. 1-9, 2017.

[34] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science*, vol. 152, pp. 341-348, 2019.

[35] E.A.E. Lucky, M.M.H. Sany, M. Keya, S.A. Khushbu, and S.R.H. Noori, "An attention on sentiment analysis of child abusive public comments towards bangla text and ml," in *12th international conference on computing communication and networking technologies (ICCCNT),* pp. 1-6, 2021.

[36] M. Rahman, M.R.A. Talukder, L.A. Setu, and A.K. Das, "A dynamic strategy for classifying sentiment from Bengali text by utilizing word2vector model," *Journal of Information Technology Research*, vol. 15, no. 1, pp. 1-17, 2022.