Low-resource (Bangla) Sentiment Analysis: Challenges and Comparative Evaluations of Transformer-based Models

Tarikul Islam Tamiti
tarik2568@gmail.com
Rajshahi University of Engineering & Technology (RUET)
Rajshahi, Bangladesh

Prasenjit Mitra
pum10@psu.edu
L3S Research Center
Hannover, Germany

Shreya Ghosh
shreya.cst@gmail.com
The Pennsylvania State University
State College, Pennsylvania, USA

Sagor Chandro Bakchy
sagorchandro.10@gmail.com
Rajshahi University of Engineering Technology (RUET)
Rajshahi, Bangladesh

Rakibul Hassan
ruet.rakib.cse13@gmail.com
Rajshahi University of Engineering Technology (RUET)
Rajshahi, Bangladesh

ABSTRACT

In this paper, we aim to develop a sentiment analysis system for Bangla text. While deep learning provides promising results in varied NLP tasks including sentiment analysis, and opinion mining, a large numbered of labeled sentences is required for training these models. This in particular is difficult to obtain for a low-resource language like Bangla. Here, we have investigated several publicly available Bangla datasets and explored deep learning, and Transformer-based models to train and evaluate three types of sentiment: neutral, positive, and negative, as well as fine-grained sentiment labels such as, fear, happiness, disgust etc. We have reported the comparative analysis among different models (highest 86% accuracy for the binary classes and 71% accuracy for the ternary class) as well as analyzed the errors from a few mis-classified samples.

CCS CONCEPTS

• **Natural language processing**; • **Sentiment analysis**;

KEYWORDS

Bangla Sentiment Analysis, Bangla Transformer-based models, Low-resource NLP, CNN, BiLSTM

ACM Reference Format:
Tarikul Islam Tamiti, Prasenjit Mitra, Shreya Ghosh, Sagor Chandro Bakchy, and Rakibul Hassan. 2023. Low-resource (Bangla) Sentiment Analysis: Challenges and Comparative Evaluations of Transformer-based Models. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Sentiment analysis (SA) plays a pivotal role in various applications including product recommendations, public opinion towards an event or political agenda [3, 29]. Sentiment analysis is the domain of Natural Language Processing (NLP) in which texts are classified into different classes of sentiments (such as, negative, positive and neutral) and more fine-grained classifications (such as, disgust, joy, fear etc.). Understanding these expressions has drawn significant attention from researchers across different domains [8]. Several research efforts have been made in identifying sentiment using machine learning and deep learning models [9, 28]. However, in order to build a robust system, a huge volume of labelled data is required for training the models. It has been observed from the literature that most of the sentiment analysis works have been carried out in high-resource languages such as English [17, 42] and thus sufficient amount of labelled samples are available in high-resource language. On the contrary, Bangla is a low-resource language with a very few linguistic tool and labelled corpora.

Due to the rapid growth of Internet-based technologies, people are using various websites and social media platforms for communication and exchanging opinions [26]. Bangladesh Telecommunication Regulatory Commission (BTRC) estimated that there were around 96.2 million active Internet users in Bangladesh at the end of the fiscal year 2019 [1]. Thus, the necessity of automatically understanding Bangla textual data has been increasing exponentially. Bangla is the fifth most spoken language. Approximately 230 million people around the world use Bangla as their native language. They concentrate mostly in Bangladesh and some parts of India [25]. Limited research has been done on extracting meaning from Bangla texts. Thus, sentiment analysis for Bangla is still at a constructive stage [5]. Because of the limited amount of resources like annotated Bangla data sets, pre-trained models, and the complex nature of the Bangla language, Bangla SA is a challenging domain for researchers [24] [13] and the research works in this domain are somewhat fragmented.

Objectives of our paper: In this paper, we attempt to provide a comprehensive review on sentiment analysis on Bangla text and report the comparative evaluations on varied machine learning and deep learning modules. We train six machine learning (ML) models,

deep learning models (DL), and fine-tune several Bangla and Multilingual Transformer-based Bidirectional Encoder Representations from Transformers (BERT) models for predicting positive, negative (binary class), and positive, negative, neutral (ternary class) sentiment using SentNoB dataset, Bangla News Comments dataset and Ekattor TV's Twitter dataset.

- Using Gridsearch CV, we have eliminated the necessity of tuning the parameter values. This method tries all the possible combinations of machine learning model parameters and provides the best-performing parameters.
- We have implemented and reported the accuracy of ensemble techniques using the BERT models.
- We have also investigated the misclassification samples when models classify the same text into different classes causing the majority voting system to fail and reported the probable causes.

2 RELATED WORK

Low-resource languages like Bangla (or Bengali) still lag behind compared to high-resource languages like English when it comes to research in NLP. We searched papers using keywords like Bangla sentiment analysis, Bangla opinion extraction, Bangla opinion mining, and Bangla subjectivity analysis. In [4]authors investigated 18 papers and discussed general terminologies, processing steps, and flowcharts to visualize the process of SA, data collection processes, and inherent challenges of Bangla SA. 71 research papers were reviewed in [36] on 11 different Bangla NLP tasks. Among the papers, 9 papers were on Bangla SA. They also provided the basic architecture of Bangla SA.

After doing a comprehensive survey, we categorized Bangla SA into four major approaches. They are 1. Classical rule-based approach, 2. Machine learning approach, 3. Transfer learning-based approach, and 4. Hybrid approach

The very first work on Bangla SA was done by the classical rule-based approach. Conditional Random Field (CRF) was used in [11] for the classification of the MPQA and IMDB binary class datasets. They used SentiWordNet and Subjectivity Wordlist. Using the Samsad dictionary these lexical resources were translated into Bangla and got a 79.90% precision value. The clustering algorithm was used in [14] along with a feedback mechanism where users specify the top 100 features according to the maximum feature margin on Amazon product reviews and movie datasets. The result was evaluated against transductive Support Vector Machine (SVM) and suggested that user feedback is equivalent to annotating 275 documents. The contextual valence of verbs was used in [19] by translating sentences into English to use SentiWordNet 3.0 for the calculation of sentiment. Lexicon-based backtracking approach was implemented in [31]. The last three words of a sentence were represented by hash value and used in the calculation of sentiment. They found a 77.16% accuracy for 301 sentences. Collecting 5,100 Bangla sentiment words and assigning sentiment scores to each word in [16] these scores were normalized and adding these normalized scores sentiment scores of each sentence were calculated. They showed that their method outperforms Decision Tree (DT), Naive Bayes(NB), and SVM models. Latent Dirichlet Allocation (LDA) was used in aspect-based sentiment analysis [39]. For sentence and review levels, they got 80% and 73% F1 scores, respectively. Word2Vec's word co-occurrence score and sentiment polarity score on Microblogging comments and 75.5% accuracy were obtained.

The most widely used method in Bangla SA is the Machine learning(ML) approach. ML models rely on manual feature extraction thus requiring large annotated datasets for training.

LR was used as a baseline in [25] and showed that LR not being resilient due to the loss of discriminating features of various classes during feature extraction and provided a very poor F1 score of 67%. Naive Bayes algorithms were based on the Bayes theorem assuming independence among predictors. The Amazon review dataset was translated into Bangla using Google translator [20]. Applying Laplace smoothing and NB they got 85% accuracy. Mutual information and Multinomial Naive Bayes (MNB) were used on Amazon's watch review [30]. 84.78% accuracy was obtained in their approach. DT, MNB, and RF with Term Frequency-Inverse Document Frequency(TF-IDF) were implemented in [36]. MNB when 6-fold cross-validated, provided an accuracy of 80.48%.

Due to robustness, efficiency in computation, and memory, it appears that SVM is the most popular classifier used in Bangla NLP. SVM outperforms several ML models on the Facebook comments dataset [27]. For seven and three classes SVM provided 62% and 73% accuracy respectively. By varying the C parameter from 0.6 to 2.8, linear SVM performed better than nonlinear SVM, providing an accuracy of 91.684%. For escaping from cumbersome manual data annotation, semi-supervised self-training bootstrapping was used for annotation in the Twitter dataset [10]. They showed that SVM outperforms MaxEnt providing an F measures score of 93%. SVM and NB were implemented with IDF for vectorization on Amazon's watch review and Twitter dataset [34]. Again, SVM outperforms NB providing an 86.8% accuracy score, and found that confidence scores form a normal distribution. SVM, DT, and LR were implemented in Bangla and Romanized Bangla converted into Bangla reviews in [18]. The Amazon review dataset was translated into Bangla d unigram with SVM provided the highest accuracy of 79%.

Deep learning extracts features automatically thus eliminating the necessity of expertise required for feature selection. CNN with Bag of Words (BOW) and TF-IDF were implemented on a Twitter dataset in [35]. The result was evaluated against Deep Belief Network (DBN). The accuracy of CNN and DBN were respectively 46.8% and 43%. LSTM, GRU, and BiLSTM were implemented [40] in Bangla, English, and Romanized Bangla YouTube comments datasets . Word2vec's continuous bag of words (CBOW) and skip grams were used as word embedding. They found a 65.97% F1-score by using LSTM. A new dataset was introduced in [21] containing Bangla and Romanized Bangla texts. After conducting 32 experiments using LSTM they got 70% and 55% accuracy for two and three classes respectively. CNN-LSTM model with word2vec provided the best F1 score of 92.83% [22]. A new dataset was created from YouTube and Facebook containing 51,000 comments [33]. Informal Fasttext (IFT) with BiLSTM provided the best F1 score of 91%.

The Transformer concept was introduced in 2017 in a paper [41] which eliminates the vanishing and exploding gradient problems of RNN models. This paved the way for large, pre-trained language models. An attention layer was embedded in a CNN model [38].

This approach is very light but the performance was comparable with heavy pre-trained models and provided the best F measures of 66.02%.

In 2018, Google released Bidirectional Encoder Representation from Transformer (BERT) [15]. The BERT model is a real game changer in the NLP field. BanglaBERT and BanglishBERT were trained on two new datasets [5]. BanglaBERT outperformed all other models and provided an accuracy of 73%. Six ML, three DL, and four BERT models were implemented in [25] on the Bangla Hate Speech dataset. XLM-Roberta provided F1 score of 87% by using ensemble majority voting, they got the highest 88% F1 score.

The last category of approach, according to our survey, is hybrid approaches where rule-based, ML, and Transfer learning-based approaches are used together. Using SentiWordNet polarity score of each word was used as one of the input features in SVM [12]. The precision and recall scores of 70.04% and 63.02% were obtained. A new domain-based categorical weighted lexicon data dictionary (LDD) and a rule-based Bangla Text Score (BTSC) algorithm were formulated in [7]. LDD and BTSC were used to generate sentiment polarity scores. TF-IDF with these scores was used to train SVM, LR, KNN, and RF models. Bigram-SVM provided the best accuracy of 82.21%. Later on [6], the technique was used to train CNN, LSTM, and BERT models. BERT-LSTM combinedly provided the best 82.27% F1 score. The BEMoC dataset was developed in [23]. Among implemented models, the XLM-R provided the highest 69.73% accuracy. A new YouTube dataset was introduced and translated into English for classification by using lexical resources like VADER, TextBlob, SentiStrength and LR, SVM, RF, and Extremely Randomized Trees(ERT). SVM provided the highest accuracy of 93.5%.

In conclusion, all four approaches discussed above have their own limitations and further research is needed for developing more efficient models. Recently, Transformer based models are getting more attention and outperforming all other models. Specifically, XLM-RoBERTa and BanglaBERT provided state-of-the-art results according to our survey in Bangla SA.

3 METHODOLOGY

This section provides a brief overview of the procedures we followed in this paper. Figure 1 represents a schematic view of the steps that we used for the classification of Bangla texts according to sentiment polarity.

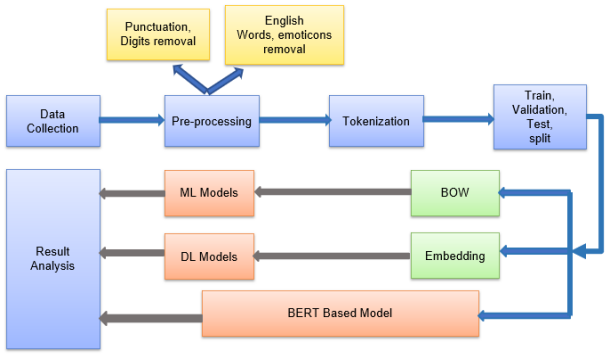

Figure 1: Schematic view of the proposed Methodology

3.1 Dataset Used

We have used the data set SentNoB [24] that was created by collecting comments from 'Prothom Alo' (Bangla Newspaper) and 'YouTube' videos. Comments were collected from 13 different domains. The comments (Data) in the dataset have 3 to 50 size word tokens. For ternary class (3 class) SA, we have used all three types of comments Neutral (0), Positive (1), and Negative (2). But in Binary class (2 class) we have used positive and negative comments i.e., '1' and '2'. 80-81% of data was used for training purposes, 9-10% was used for validation purposes, and 10% was used for testing purposes for DL and Transformer-based models. For ML models, 90% of the data was used for training purposes and 10% for testing purposes. Names of the classes, number of instances, labels, and one text as an example of each class and the English translation of each text of the SentNoB dataset are provided in Table 1.

The second dataset used in this paper was collected from Ekattor TV's Twitter dataset [32]. This dataset contains six emotion classes. However, many instances of the dataset were not correctly annotated which we have corrected. In table 2, the name of the classes, number of instances, labels, sample text of each class, and the English translation of each text is provided.

The final dataset used in this paper was on Bangla News Comments [2]. This dataset contains five classes (See Table 3).

3.2 Preprocessing and Tokenization

In the dataset, many words, and symbols were present that had little importance for sentiment analysis. Therefore, removing digits, various symbols used on the keyboard, pictographs, transport, and map symbols, flags(ios), Latin, general punctuations, and Bangla digits by using regular expressions(regex) will not affect the overall performance. However, as pre-trained BERT models provide better results on raw texts [25], we have restrained ourselves from further preprocessing. Tokenization is the process by which texts are broken down into smaller pieces (tokens). Tokens speed up the process of understanding the context of the words thus, classification can be done quickly. We have used the 'csebuetnlp/banglabert' [5] model's tokenizer to tokenize the words. Words are broken down into pieces of two or more subwords known as tokens as per the pre-trained BERT model. '##' is padded before the subwords except for the first subword. If the token is not present in the original vocabulary, it will be represented by a special token '[UNK]'. We have found that there are 32,000 tokens in the SentNoB dataset. In the following figure 2 one example of tokenization is demonstrated.

Figure 2: Example of Tokenization

3.3 Feature extraction

Feature extraction, feature encoding, or vectorization is the process by which textual data is converted into numeric data. ML and

Table 1: SentNoB Dataset overview

Class	Instances	Label	Text	Translation
Neutral	5709	0	আকবরিয়ার দই চট্টগ্রামে পাবো কি ভাবে	How can I get Akbaria yogurt in Chittagong?
Positive	6401	1	ভাই আপনার সব গুলো ভিড়িও আমার খুব ভাল লাগে	I like all your videos bro.
Negative	3609	2	হয়তো তার পদত্যাগ হবে অথবা কিছুই হবে না কিন্তু দেশ এতোটাই রসাতলে চলে গেছে যে এসব পাপ থেকে মুক্তি লাভ করা খুব কঠিন হয়ে পড়েছে	Maybe he will resign or nothing will happen but the country has gone to such an abyss that it is very difficult to get rid of these sins.

Table 2: Ekattor TV's Twitter dataset overview

Class	# of Instances	Label	Sample Text	Translation
Angry	1427	0	সিএসও কে বরখাস্ত করা হোক তিনি বিমানবন্দরে রাজনীতি শুরু করেছেন	The CSO should be sacked, he has started politics at the airport
Disgust	704	1	দুর্ভাগ্য শিশুটির যে তার পরিবার তার খেয়াল না রেখে কাকে নিয়ে আনন্দ ভাগাভাগি করল ফলে সে তার পরিবার থেকে হারিয়ে গিয়েছিল	The unfortunate child was lost from his family because his family did not care about who he shared his joy with
Fear	392	2	বাহিরের মানুষ দুরের কথা নিজের দেশের মানুষের নিরাপদ না আমাদের দেশ আমারা দেশে যেতে ভয় করে তো বাহিরে মানুষ কথা বাদ দিলাম	People from outside are not talking about the people of our own country. Our country is not safe. We are afraid to go to our country.
Happy	1924	3	আপনার এই লেখাটা আমার খুব ভাল লাগলো	I liked your writing very much.
Sad	1366	4	বহু বছরের পুরনো স্মৃতি মনে পারছে যখন ছোট বেলা দাদা যেতাম	I remember many years old memories when I used to go to my grandfather's house when I was a child.
Surprise	592	5	ঝাল না কি মিষ্টি কিছুই বুঝইতে পারি নাই	I can't understand whether it is salty or sweet.

Table 3: Bangla News Comment dataset overview

Class	Instances	Label	Text	Translation
Negative	3198	0	বাংলাদেশ ব্যাংকের রিজার্ভ চুরির ঘটনায় ফিলিপাইন যতটা সরব বাংলাদেশ ততটাই নীরবতদন্ত কমিটির রিপোর্ট ইহজনমে আলোর মুখ দেখবে কিনা সন্দেহ	It is doubtful whether the Philippines will see the light of day in the report of the investigative committee as much as Bangladesh in the theft of Bangladesh Bank's reserves.
Neutral	2951	1	চুক্তিটি নাভানা এর সাথে হয়েছে টয়োটা এর সাথে নয়	The deal is with Navana and not Toyota
Positive	1445	2	ম্যাকগ্রাই সেরা সে শচিন কেও বার আউট করছে আর এটাই সবচেয়ে বিখ্যাত	McGrey is the best he is bowling out Sachin Keo and this is the most famous
Very Negative	3928	3	গনতন্ত্রের কফিনে শেষ পেরেক মারছে সরকার	The government is putting the final nail in the coffin of democracy
Very Positive	2280	4	খুশিতে মনটা ভরে গেলো এগিয়ে যাও ভাইয়েরা	My heart was filled with happiness. Go ahead, brothers

DL models cannot understand semantic meaning from texts or tokens automatically [37]. For ML models BOW, TF-IDF (Unigram) and for DL models Keras embedding layers were used to extract features.

BOW: Bag of words (BOW) counts the frequency of any word in the document whether it is present in the document or not. BOW consists of two main parts. Corpus or vocabulary of the known words and frequency of the word in a sentence. For BOW, it is assumed that sentences or paragraphs having similar meanings will contain similar words. In BOW, the words that are present many times start to dominate but they may not contain much information that helps in the classification.

TF-IDF: Term Frequency-Inverse Document Frequency (TF-IDF) calculates scores to signify the importance of words. TF-IDF partially removes the problems that arise along with the BOW and emphasizes the rarity of a word. Words can be grouped together, known as grams. Grams are used for capturing more context information and reducing the size. But analyzing higher grams requires more computational power. Unigram considers one word at a time whereas Bigram and Trigram consider two and three consecutive words at a time in a sentence.

Keras embedding layer: For deep learning models, we used the Keras embedding layer by which tokens are converted into a sequence of integers by the Keras 'text_to_sequences' method. Keras 'pad_sequences' method was used so that all the sequences have the same size equal to 50. 'pad_sequences' method pad(add) '0' to the posterior of those tokens having a length less than 50. Keras embedding layer converts the token's semantic meaning into the predefined sized dimensional vectors in the vector space [1]. In DL models the embedding dimensions were set to 40,100,300. Thus, each token is converted into 40,100,300-dimensional vectors.

3.4 Machine Learning models

Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Multinomial Naive Bayes (MNB), k-nearest neighbors (KNN), and Stochastic Gradient Descent (SGD) models are used after converting tokens into numeric values by BOW and TF-IDF unigram. Table 4 provides different parameters used for ML models. We have used the GridSearchCV method to find the best parameters. Accuracy and F1-score have been used to report the models' performances.

Logistic Regression: Logistic regression(LR) is a type of supervised learning in which inputs are given as independent variables. The output of the logistic regression is a categorial dependent variable. Logistic regression provides an output that is probabilistic in nature, i.e. the value of the output lies between 0 to 1. For the successful operation of the LR model, it should be assured that inputs must not have multi-collinearity and the output should be categorical. LR was implemented by using scikit-learn's linear model LogisticRegression. We set the value of the C hyperparameter to different values 1,5,10,25. Hyperparameters are actually used to indicate the models and how to choose parameters. C hyperparameter is used for regularization purposes. The value of C must be a positive float-type value. The high value of C indicates assigning a high weight to training data, overlooking outliers, extreme data points, which lead to overfitting issues.

Support Vector Machine: Support Vector Machine (SVM) finds out the best hyperplane to accurately classify the data samples. SVM was implemented by us using scikit-learn library. The 'C' parameter was assigned with 1,10,20. The 'C' parameter is a regularization parameter and the strength is inversely proportional to the assigned value of 'C'. For each error, the penalty is a squared 'l2' penalty. 'kernel' parameter is assigned with 'rbf', 'linear', and 'sigmoid'. The 'kernel' parameter specifies the type of function used to convert the data points into higher dimensions from lower dimensions. Gama parameter assigned with 'auto' and 'scale' values. 'decision_function_shape' assigned with 'ovo' and 'ovr'. One-vs-rest('ovr') or one-vs-one('ovo') will be the return policy, which is defined by the 'decision_function_shape' parameter.

Decision Tree: A Decision Tree (DT) is a tree-structured, non parametric classifier used for classification. For developing the decision tree, Classification and Regression Tree algorithm (CART algorithm) is deployed. The complexity of the Decision tree is logarithmic and 'criterion' parameter is used for the measurement of the splitting done at each node. The value of the 'criterion' parameter is assigned to 'entropy'. 'max_depth' is used for indicating how deep the nodes will be expanded. The values assigned to the 'max_depth' parameter are 125,500,700,900. 'min_samples_split' indicates the lowest number of samples required to split the internal node. 'min_samples_split parameter are assigned with 15,55,95 values. 'max_features' defines the number of features should be considered before searching for the best split. The values assigned to 'max_features' are 'sqrt','log2'.

Multinomial Naïve Bayes: Naïve Bayes is a probabilistic method for classification based on the Bayes theorem. Multinomial and Gaussian are two major types of Naive Bayes classifiers. Multinomial Naïve Bayes (MNB) classifies the textual data by decision function. MNB also has been implemented using the scikit-learn library. The alpha parameter is used as the controlling parameter for MNB. The alpha parameter is an additive smoothing parameter. There are two types of smoothing. Laplace and Lidstone are two types of smoothing parameters. For the zero value of alpha, no smoothing will be conducted. We have assigned with 0.10, 0.25,0 .40, 0.75, 0.90 values for Alpha parameter.

K Nearest Neighbor: K Nearest Neighbor(KNN) considers K numbers of neighboring data points for classification. This algorithm uses instance-based learning where instead of adjusting weights from training data, it uses the whole spectrum of training instances. This technique is also called lazy learning. KNN is also nonparametric, which means no predefined form of the mapping function is used. While implementing KNN from scikit-learn we assigned 'n_neighbors' parameters for the number of neighbors selection. 'weights' parameters were used for the weight function and assigned with 'uniform', 'distance' values. 'uniform' weight means all the neighborhoods are weighted equally. Assigning the 'distance' value means weight points are inverse of their distance. 'algorithm' parameter computes the value of nearest neighbors.'ball_tree', and 'kd_tree' is assigned for generalization of N-point problems. 'brute' value uses a brute-force search algorithm.

Stochastic Gradient Descent: Gradient Descent finds optimal solutions to various problems and 'stochastic' means random probability linked with a process. The parameters of Stochastic Gradient Descent (SGD) are tuned iteratively for minimizing the cost function. For tuning hyperparameters, the 'loss' parameter was assigned with 'log' value, 'penalty' parameter was assigned with 'l2','l1','elasticent' values, 'alpha' parameter was assigned with 0.0001, 0.0005, 0.0009, 0.0012 values. The parameters used in Machine learning models are provided in table 4.

3.5 Deep learning models

Deep Learning (DL) models are used for learning the rules for language analysis and improve the ability of computers to process text and analyze human language [26]. Deep learning refers to the use of multilayer neural networks in machine learning [37]. We have used a Convolutional Neural Network (CNN) model and a concatenation of Convolutional Neural Network (CNN), and Bidirectional Long Short Term Memory (BiLSTM) called Conv-BiLSTM for sentiment classification. The parameters used in the DL models were provided in Table 5.

CNN: By using Keras embedding, the input textual data is converted into numeric data. Then, the numeric data is provided as input to the one-dimensional convolution layer. After the convolution operation, the next layer's input is the output of this convolution layer followed by the global_max_pooling layer. Here, the largest value among the neighbor values is chosen. Next, the dropout layer is used to drop the connection between some neurons to the next layer to avoid the contribution of some neurons to reduce the effect of overfitting. Then the output of this layer is fed to the input as the next fully connected dense layer. Then the following layers are two pairs of dropout and dense layers. Then the output is flattened in the flatten layer. Flatten layer converts the output into a one-dimensional array. Then at last output is taken from the fully connected dense layer. The output is provided as a probability. For binary classification, the sigmoid function is used,

Table 4: Machine Learning Models' Parameters

Model	Parameters
SVM	C: 1,10,20; kernel: rbf, linear, sigmoid; gamma: auto, scale; decision_function_shape: ovo,ovr
DT	max_depth : 125,500,700,900; max_features : sqrt,log2; min_samples_split: 15,55,95; criterion: entropy
LR	C: 1, 5, 10, 25
MNB	alpha: 0.10, 0.25, 0.40, 0.75, 0.90
KNN	n_neighbors: 3, 4, 5, 6, 7; weights: uniform, distance; algorithm : ball_tree, kd_tree, brute
SGD	loss: log; penalty: l2, l1, elasticnet; alpha: 0.0001, 0.0005, 0.0009, 0.0012

whereas, for ternary classification, the softmax function is used to select the highest probability class. The architecture of the implemented CNN model is provided in figure 3.

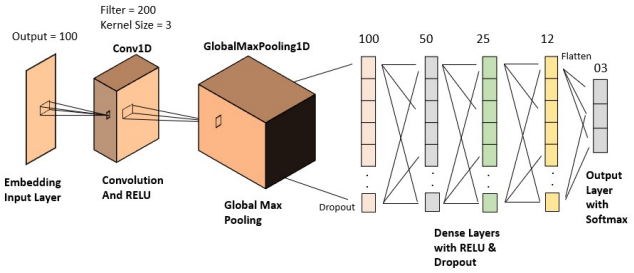

Figure 3: Architecture of implemented CNN model

Conv-BiLSTM: A concatenation of BiLSTM and CNN (Conv-BiLSTM) is implemented after hyperparameter tuning for efficient feature learning. In this model, a Conv1D layer of kernel size 3 and filter size 200 is applied with a Bidirectional LSTM of filter size 128 and 64 in two layers. A dropout layer of 0.25 is applied when the LSTM filter size is 200 and a dropout of 0.1 is used when the LSTM filter size = 64. Activation functions ReLU and Softmax are used for neuron selection of each layer. By Keras embedding layer, textual is converted into numeric input. Then the output of the previous layer is fed into the input of the next layer. The next layer is the conv1d layer. In the conv1d layer, one-dimensional convolution is performed on the data. The next layer is the bidirectional LSTM layer. Followed by a dropout layer for avoiding overfitting. Then again bidirectional LSTM and dropout layer was implemented. Subsequently, two pairs of dense and dropout layers were implemented. Lastly, in between two dense layers, one flatten layer was implemented. The architecture of the implemented Conv-BiLSTM model is provided in Figure 4.

3.6 Transformer-based models

In Transformers, every output element is connected to every input element, and the weightings between them are dynamically determined based on their connection. With just one extra output layer, the pre-trained Transformers model may be fine-tuned to provide state-of-the-art results for a variety of NLP tasks. For understanding a language, Transformers models are trained on vast unlabeled data. By fine-tuning just one layer, the vast knowledge that models acquired is implemented to do a particular task by transfer learning. Several Transformer models were selected from the Hugging

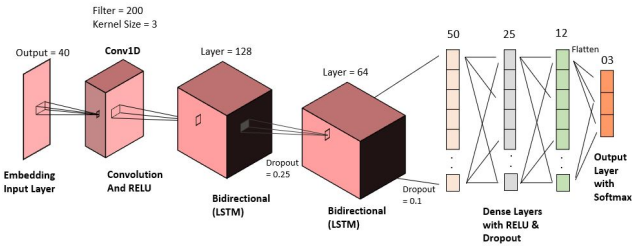

Figure 4: Architecture of implemented Conv-BiLSTM model

Table 5: Deep Learning Models' Parameters

Hyper parameters	CNN	Conv-LSTM
Embedding Dimension	40,100,300	40,100,300
Pooling type	Max	——
Batch size	32,64	32,64
Max length	50,100,200	50,100,200
Padding, Truncating	pre, post	pre, post
Activation Function	Relu, Softmax	Relu, Softmax
Learning rate	0.05,0.003,.008	0.05,0.003,.008
Dropout	0.1,.25,0.5	0.1,.25,0.5

Table 6: Transformer-based Models' Parameters

Hyperparameter	Value
class_weight	balanced
Learning Rate	1e-3 to 1e-5
Epochs	7,10,30
preproc	trans
Max length	50

transformers library and implemented using the Ktrain wrapper. The parameters used to fine-tune these pre-trained models were presented in Table 6.

4 RESULTS ANALYSIS

In this section, we report the experimental evaluations and findings on different Bangla text datasets.

Table 7: Evaluation of the models' performance on test data for SentNoB Binary class

Feature	Model	Accuracy	F1-score
BOW	SVM	0.738	0.737
BOW	DT	0.705	0.704
BOW	LR	0.717	0.717
BOW	MNB	0.733	0.732
BOW	KNN	0.702	0.702
BOW	SGD	0.715	0.715
Unigram	SVM	0.859	0.859
Unigram	DT	0.774	0.743
Unigram	LR	0.852	0.851
Unigram	MNB	0.843	0.842
Unigram	KNN	0.829	0.828
Unigram	SGD	0.829	0.828
Bigram	SVM	0.856	0.855
Bigram	DT	0.73	0.729
Bigram	LR	0.861	0.86
Bigram	**MNB**	**0.866**	**0.865**
Bigram	KNN	0.811	0.81
Bigram	SGD	0.84	0.838
Trigram	SVM	0.86	0.86
Trigram	DT	0.74	0.741
Trigram	LR	0.858	0.858
Trigram	MNB	0.863	0.862
Trigram	KNN	0.805	0.804
Trigram	SGD	0.842	0.841
Keras Embedding layer	CNN	0.84	0.84
Keras Embedding layer	Conv-BiLSTM	0.83	0.83
	csebuetnlp/banglabert	0.86	0.86
	Kowsher/bangla-bert	0.84	0.82
	monsoon-nlp/bangla-electra	0.82	0.82
	bert-base-multilingual-cased	0.82	0.82
	bert-base-multilingual-uncased	0.82	0.82
	sagorsarker/bangla-bert-base	0.84	0.84
	xlm-roberta-base	0.83	0.83
	Ensemble bert-base-multilingual-uncased, monsoon-nlp/bangla-electra, xlm-roberta-base	0.84	0.84

4.1 Experiment setup

Models were implemented using scikit-learn (version 1.21), Keras library (version 2.11.0), and Ktrain wrapper(version 0.31X) on Kaggle and Google Colab platforms. Huggingface library(version 4.26.0) was used for the implementation of Transformer-based models.

4.2 Result discussion

We have reported accuracy and F1-score for sentiment classification.

Table 8: Evaluation of the models' performance on test data for SentNoB ternary class

Feature	Model	Accuracy	F1-score
BOW	SVM	0.604	0.569
BOW	DT	0.522	0.494
BOW	LR	0.569	0.499
BOW	MNB	0.567	0.501
BOW	KNN	0.55	0.528
BOW	SGD	0.569	0.493
Unigram	**SVM**	**0.71**	0.665
Unigram	DT	0.594	0.565
Unigram	LR	0.698	0.658
Unigram	MNB	0.685	0.619
Unigram	KNN	0.669	0.638
Unigram	SGD	0.674	0.6
Keras Embedding layer	**CNN**	0.69	**0.67**
Keras Embedding layer	Conv-BiLSTM	0.66	0.61
	csebuetnlp/banglabert	0.70	**0.67**
	Kowsher/bangla-bert	0.69	**0.67**
	monsoon-nlp/bangla-electra	0.67	0.63
	bert-base-multilingual-cased	0.66	0.64
	bert-base-multilingual-uncased	0.66	0.65
	sagorsarker/bangla-bert-base	0.70	**0.67**
	Ensemble (bert-base-multilingual-uncased, monsoon-nlp /bangla-electra, bert-base -multilingual-cased)	0.68	**0.67**

In the binary classification of the SentNoB dataset, Multinomial Naive Bayes with Bigram feature extraction achieved the highest 86.6% accuracy score and 86.5% F1-score. Among the Machine Learning models, for the Bigram feature extraction technique, LR scored 86.1% accuracy. SVM and MNB scored 86.1% and 86.3% accuracy score, 86% and 86.2% F1-score respectively. Deep Learning and Transformer based models scored between 82% to 86% whereas, several Machine Learning models performed poorly, in the range from 70% to 86% accuracy. These results provide evidence of the supremacy of Deep learning and Transformer-based models over Machine learning based models. Table 7, summarizes the implemented models' performance on the Binary class of the SentNoB dataset is provided. The best accuracy and F1-score along with respective feature extraction are highlighted in the table.

Table 8 summarizes the implemented models' performance on the ternary class of the SentNoB dataset. For the ternary class, SVM provided the highest performance of 71% accuracy. Several DL and Transformer based scored highest 67% of F1-score. Overall, Transformer based models outperformed ML and DL models. From visualizing the Confusion Matrices, we found that the models used in this paper struggled to predict the neutral class correctly, which eventually dragged down the models' scores. In Tables 9 and 10,

Table 9: Results on Six class Ekattor TV Twitter dataset

Feature Extraction	Model	Accuracy	F1-score
BOW	SVM	0.427	0.242
BOW	DT	0.349	0.215
BOW	LR	0.412	0.275
BOW	MNB	0.409	0.257
BOW	KNN	0.37	0.237
BOW	SGD	0.415	0.252
Unigram	SVM	0.512	0.351
Unigram	DT	0.373	0.26
Unigram	LR	0.509	0.317
Unigram	MNB	0.501	0.296
Unigram	KNN	0.459	0.316
Unigram	SGD	0.508	0.337
Bigram	SVM	0.511	0.332
Bigram	DT	0.392	0.26
Bigram	LR	0.517	0.365
Bigram	MNB	0.506	0.319
Bigram	KNN	0.472	0.352
Bigram	SGD	0.509	0.31
Trigram	SVM	0.512	0.352
Trigram	DT	0.373	0.25
Trigram	LR	0.518	0.346
Trigram	MNB	0.5	0.307
Trigram	KNN	0.461	0.336
Trigram	SGD	0.5	0.297
Keras Embedding layer	CNN	0.45	0.21
Keras Embedding layer	Conv-BiLSTM	0.41	0.25
sagorsarker /bangla-bert-base embedding	Conv-BiLSTM	0.44	0.44
Attention embedding	BiLSTM	0.41	0.33
	csebuetnlp /banglabert	**0.56**	**0.45**
	bert-base-multilingual-cased	0.45	0.36
	bert-base-multilingual-uncased	0.43	0.35
	monsoon-nlp /bangla-electra	0.33	0.26
	xlm-roberta-base	0.51	0.38
	Kowsher/ bangla-bert	0.48	0.38
	Ensemble (bert-base-multilingual-uncased, monsoon-nlp/ bangla-electra, bert-base-multilingual-cased)	0.42	0.33

Table 10: Evaluation of the models' performance on Bangla News Comment dataset

Feature Extraction	Model	Accuracy	F1-score
Keras Embedding layer	**CNN**	**0.35**	0.22
Keras Embedding layer	Conv-BiLSTM	0.33	0.22
	sagorsarker /bangla-bert-base	0.31	**0.30**
	bert-base -multilingual-uncased	0.09	0.05
	monsoon-nlp /bangla-electra	0.27	0.09
	bert-base-multilingual-cased	0.26	0.25
	Ensemble (bert-base-multilingual-uncased, monsoon-nlp /bangla-electra, bert-base -multilingual-cased)	0.21	0.18

we summarized the results obtained from implementing the models on the six classes of Ekattor TV's Twitter dataset and on the five classes Bangla News Comment dataset. On the Ekattor TV dataset, csebuetnlp/banglabert model provided the highest accuracy of 56% and 45% F1-score. As previously mentioned, with increasing the number of classes, the models' accuracy and F1-score were also lowered.

It is reported that on five class Bangla News comment dataset, all of the models performed very poorly. The highest accuracy on this dataset was obtained from the CNN model and it was only 35% and highest F1-score of 30% from sagorsarker/bangla-bert-base model. In this dataset, Transformer-based models performed poorly than DL models. The probable reasons for these poor performances are (i) the poor quality of data annotation, (ii) the absence of sentiment polarity words on the datasets (iii) the complex nature of the Bangla language.

The models were also implemented on pure texts and token ID where tokens are converted into numeric values according to the vocabulary created from the corpus, along with tokens. However, we didn't notice any significant change in the result. The learning rate is one of the most important parameters for reaching the global minima i.e. the point where the loss is minimum. Thus, for finding the perfect learning rate we investigate the loss vs learning rate curve (log scale). The learning rate was taken where the loss was minimal. The learning rates of the Transformer based implemented models in this paper varied between 1e-3 to 1e-5.

4.3 Error Analysis

In this subsection, we investigate several cases of misclassification by manual inspection. The ensemble technique is used to increase

the overall performance by using majority voting. We used the aggregate function's mode parameter for enforcing majority voting. Some of the cases of misclassification have been discussed below:

Sentence 1: প্রত্যেক মাসে মাসে যেন চলে এই সব অভিযান তা না হলে এই সব বন্ধ হবে না ধন্যবাদ

Translation: If these expeditions are carried out every month, this will not stop. Thank you.

bert-base-multilingual-uncased's prediction: negative
monsoon-nlp/bangla-electra's prediction: negative
xlm-roberta-base's prediction: positive
Ensemble's prediction: negative
Actual sentiment: positive

Possible reason for Misclassification: Due to the presence of negation word 'not' two models misclassify into negative class resulting final outcome as negative

Sentence 2: একজন অসুস্থ মানুষ কিছু ফালতু ভিডিও বানিয়ে অনেক মানুষকে অসুস্থ করে ফেলছে এখন মজা করে খাবে আর শেষ বয়সে সরকারী অনুদানে চিকিৎসা করাবে

English translation: A sick man made some stupid videos and made many people sick, now he will eat for fun and in his old age he will be treated with government grants.

bert-base-multilingual-uncased's prediction: positive
monsoon-nlp/bangla-electra's prediction: negative
xlm-roberta-base's prediction: positive
Ensemble's prediction: positive
Actual sentiment: negative

Possible reason for Misclassification: In the sentence, there is no word that strongly expresses negative polarity. Thus, two of the models misclassify resulting in overall misclassification.

Sentence 3: আমি সবার কমেন্ট এ লাইক দেই বাট আমার কমেন্ট এ কেউ লাইক দেইনা

English translation: I like everyone's comments but no one likes my comments

bert-base-multilingual-uncased's prediction: positive
monsoon-nlp/bangla-electra's prediction: negative
xlm-roberta-base's prediction: positive
Ensemble's prediction: positive
Actual sentiment: negative

Possible reason for Misclassification: Here, models miss the important conjunction 'but' and the second part of the sentence.

Sentence 4: এই দেশের প্রতিটি ক্ষমতাবান মন্ত্রী নেতা ও বড়বড় কর্মকর্তা তাদের চিন্তা ভাবনা হলো দেশের মানুষ গুলো বাচুক আর মরুক কে মরলো কে বাচলো কিছু যায় আসেনা কে খাইলো আর না খাইলো কার বাড়ী ঘর আছে না আছে আমার ব্যাংক ব্যালেন্সটা যেনো কোটি কোটি টাকা দিয়া ভর্তি থাকে

English translation: Every powerful minister, leader and big official of this country, their thoughts are whether the people of the country live or die, who died, who lived, it doesn't matter who ate or not, who has a house or not, my bank balance should be filled with crores of rupees.

bert-base-multilingual-uncased's prediction: neutral
monsoon-nlp/bangla-electra's prediction: negative
bert-base-multilingual-cased's prediction: neutral
Ensemble's prediction: neutral
Actual sentiment: negative

Possible reason for Misclassification: Due to the length of the sentence and the complexity, models misclassify this comment.

Table 11: Comparison with previous works

Dataset	Best Model	Performance
SentNoB	SVM [24]	0.64 (F1-score)
SentNoB	csebuetnlp/banglabert [5]	0.72(F1-score)
SentNoB	Proposed SVM model	0.71 (Accuracy)
SentNoB	Proposed CNN model	0.69 (Accuracy)
SentNoB	Finetuned csebuetnlp/banglabert	0.70(Accuracy)

Sentence 5: যাতে তাদের কু থেকে কর্ম আমরা যেনো জানতে না পারি

English translation: So that we do not know their bad deeds bert-base-multilingual-uncased's prediction: positive
monsoon-nlp/bangla-electra's prediction: neutral
bert-base-multilingual-cased's prediction: negative
Ensemble's prediction: neutral
Actual sentiment: negative

Possible reason for Misclassification: This is a very special case of misclassification. The actual sentiment of this type of sentence is very subtle and very hard to classify. Three models classify these types of texts into three different classes.

Sentence 6: দার্জিলিং এ ভাল কেনাকাটা করা যায় এমন কয়েক টা মার্কেট নাম বলুন

English translation: Name some good shopping markets in Darjeeling bert-base-multilingual-uncased's prediction: positive
monsoon-nlp/bangla-electra's prediction: positive
bert-base-multilingual-cased's prediction: positive
Ensemble's prediction: positive
Actual sentiment: neutral

Possible reason for Misclassification: Due to the presence of positive polarity word 'good' all models misclassify it as positive sentiment.

4.4 Comparison with previous works

Islam et al. [24] implemented SVM, BiLSTM, and mBERT with various feature extraction techniques on SentNoB dataset and reported highest 64% F1 score. According to their findings, the neutral class was hardest to classify, the negative class was easy to classify for Politics and Economy class.

Two transformer-based BERT models on the Bangla language namely BanglaBERT and BanglishBERT were developed in [5] where they created two datasets for Natural Language Inference (NLI) and Question Answering (QA). The performance of the two BERT models was evaluated on four benchmark datasets on NLI, QA, Sentiment classification, and Name entity recognition. Combining four tasks, Bangla Language Understanding Benchmark (BLUB) was developed by them. Their pre-trained Bangla BERT model outperformed other transformer models and provided almost 73% of the F1 score[1]

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have carried out a comprehensive review of existing low-resource sentiment analysis task for Bangla. We have implemented six machine learning modules along with Transformer-based models on the SentNoB dataset to compare

[1]Our code is available here LINK.

the performance between binary and ternary sentiment classification. Furthermore, the performance of the model was evaluated on datasets with fine-grained sentiment (emotions such as, happy, sad, surprise etc.). To the best of our knowledge, no existing works in Bangla sentiment analysis have thoroughly evaluated the performances of transformer-based models on different datasets. It is observed that the performances of the models in three, five, and six labelled sentiment classifications are low due to the complex nature of the Bangla language and the quality of the annotated datasets. Overall `csebuetnlp/banglabert' model outperformed other models. The major limitations in the existing research of Bangla sentiment analysis are (i) the unavailability of quality labelled data and (ii) an efficient model that can handle multi-lingual data (for instance, Bengali and English comments, dialogues etc.) for efficient sentiment classification. In the future, we would like to create a Multi-class Bangla Text dataset with fine-grained sentiment labelling and investigate the challenges of BERT models in identifying the sentiments. Further, we aim to train Transformer based BERT models specifically focused on Bangla Sentiment analysis (where multi-lingual text is present) and deploy them in real-life scenarios for other tasks, such as stance detection and user's opinion mining.

REFERENCES

[1] Samrat Alam, Md Afnan Ul Haque, and Ashiqur Rahman. 2022. Bengali Text Categorization Based on Deep Hybrid CNN--LSTM Network with Word Embedding. In 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET). IEEE, 577--582.

[2] Md Akhter-Uz-Zaman Ashik, Shahriar Shovon, and Summit Haque. 2019. Data set for sentiment analysis on Bengali news comments and its baseline evaluation. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP). IEEE, 1--5.

[3] Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. 2013. Sentiment analysis of political tweets: Towards an accurate classifier. Association for Computational Linguistics.

[4] Nayan Banik, Md Hasan Hafizur Rahman, Shima Chakraborty, Hanif Seddiqui, and Muhammad Anwarul Azim. 2019. Survey on text-based sentiment analysis of bengali language. In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT). IEEE, 1--6.

[5] Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. arXiv preprint arXiv:2101.00204 (2021).

[6] Nitish Ranjan Bhowmik, Mohammad Arifuzzaman, and M Rubaiyat Hossain Mondal. 2022. Sentiment analysis on Bangla text using extended lexicon dictionary and deep learning algorithms. Array 13 (2022), 100123.

[7] Nitish Ranjan Bhowmik, Mohammad Arifuzzaman, M Rubaiyat Hossain Mondal, and MS Islam. 2021. Bangla text sentiment analysis using supervised machine learning with extended lexicon dictionary. Natural Language Processing Research 1, 3-4 (2021), 34--45.

[8] Mondher Bouazizi and Tomoaki Ohtsuki. 2019. Multi-class sentiment analysis on twitter: Classification performance and challenges. Big Data Mining and Analytics 2, 3 (2019), 181--194.

[9] Prerna Chikersal, Soujanya Poria, and Erik Cambria. 2015. SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning. In Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015). 647--651.

[10] Shaika Chowdhury and Wasifa Chowdhury. 2014. Performing sentiment analysis in Bangla microblog posts. In 2014 International Conference on Informatics, Electronics & Vision (ICIEV). IEEE, 1--6.

[11] Amitava Das and Sivaji Bandyopadhyay. 2009. Subjectivity detection in english and bengali: A crf-based approach. Proceeding of ICON (2009).

[12] Amitava Das and Sivaji Bandyopadhyay. 2010. Phrase-level Polarity Identification for Bangla. Int. J. Comput. Linguistics Appl. 1, 1-2 (2010), 169--182.

[13] Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H Sarker. 2021. Emotion classification in a resource constrained language using transformer-based approach. arXiv preprint arXiv:2104.08613 (2021).

[14] Sajib Dasgupta and Vincent Ng. 2009. Topic-wise, sentiment-wise, or otherwise? Identifying the hidden dimension for unsupervised text classification. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 580--589.

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

[16] Rajib Chandra Dey and Orvila Sarker. 2019. Sentiment analysis on bengali text using lexicon based approach. In 2019 22nd International Conference on Computer and Information Technology (ICCIT). IEEE, 1--5.

[17] Asif Ekbal and Pushpak Bhattacharyya. 2022. Exploring Multi-lingual, Multi-task, and Adversarial Learning for Low-resource Sentiment Analysis. Transactions on Asian and Low-Resource Language Information Processing 21, 5 (2022), 1--19.

[18] Fabliha Haque, Md Motaleb Hossen Manik, and MMA Hashem. 2019. Opinion mining from bangla and phonetic bangla reviews using vectorization methods. In 2019 4th International Conference on Electrical Information and Communication Technology (EICT). IEEE, 1--6.

[19] KM Azharul Hasan, Mosiur Rahman, et al. 2014. Sentiment detection from bangla text using contextual valency analysis. In 2014 17th international conference on computer and information technology (ICCIT). IEEE, 292--295.

[20] KM Azharul Hasan, Mir Shahriar Sabuj, and Zakia Afrin. 2015. Opinion mining using naive bayes. In 2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE). IEEE, 511--514.

[21] Asif Hassan, Mohammad Rashedul Amin, Abul Kalam Al Azad, and Nabeel Mohammed. 2016. Sentiment analysis on bangla and romanized bangla text using deep recurrent models. In 2016 International Workshop on Computational Intelligence (IWCI). IEEE, 51--56.

[22] Muntasir Hoq, Promila Haque, and Mohammed Nazim Uddin. 2021. Sentiment analysis of bangla language using deep learning approaches. In Computing Science, Communication and Security: Second International Conference, COMS2 2021, Gujarat, India, February 6--7, 2021, Revised Selected Papers. Springer, 140--151.

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

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

[25] Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 1--10.

[26] Lal Khan, Ammar Amjad, Noman Ashraf, and Hsien-Tsung Chang. 2022. Multi-class sentiment analysis of urdu text using multilingual BERT. Scientific Reports 12, 1 (2022), 5436.

[27] Md Serajus Salekin Khan, Sanjida Reza Rafa, Amit Kumar Das, et al. 2021. Sentiment analysis on bengali facebook comments to predict fan's emotions towards a celebrity. Journal of Engineering Advancements 2, 03 (2021), 118--124.

[28] Yinxia Lou, Yue Zhang, Fei Li, Tao Qian, and Donghong Ji. 2020. Emoji-based sentiment analysis using attention networks. ACM Transactions on asian and low-resource language information processing (TALLIP) 19, 5 (2020), 1--13.

[29] Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. ACM Transactions on Internet Technology (TOIT) 17, 3 (2017), 1--23.

[30] Animesh Kumar Paul and Pintu Chandra Shill. 2016. Sentiment mining from bangla data using mutual information. In 2016 2nd international conference on electrical, computer & telecommunication engineering (ICECTE). IEEE, 1--4.

[31] Tapasy Rabeya, Sanjida Ferdous, Himel Suhita Ali, and Narayan Ranjan Chakraborty. 2017. A survey on emotion detection: A lexicon based backtracking approach for detecting emotion from Bengali text. In 2017 20th international conference of computer and information technology (ICCIT). IEEE, 1--7.

[32] Ayubur Rahaman. 2013. Ayubur/bangla-sentiment-analysis-datasets: Different Bangla datasets for sentiment analysis on Bangla text. https://github.com/Ayubur/bangla-sentiment-analysis-datasets. [Online; accessed 01-Dec-2022].

[33] Nauros Romim, Mosahed Ahmed, Md Islam, Arnab Sen Sharma, Hriteshwar Talukder, Mohammad Ruhul Amin, et al. 2022. BD-SHS: A Benchmark Dataset for Learning to Detect Online Bangla Hate Speech in Different Social Contexts. arXiv preprint arXiv:2206.00372 (2022).

[34] Mir Shahriar Sabuj, Zakia Afrin, and KM Azharul Hasan. 2017. Opinion mining using support vector machine with web based diverse data. In Pattern

Recognition and Machine Intelligence: 7th International Conference, PReMI 2017, Kolkata, India, December 5-8, 2017, Proceedings 7. Springer, 673--678.

[35] Kamal Sarkar. 2019. Sentiment polarity detection in Bengali tweets using deep convolutional neural networks. Journal of Intelligent Systems 28, 3 (2019), 377--386.

[36] Ovishake Sen, Mohtasim Fuad, Md Nazrul Islam, Jakaria Rabbi, Mehedi Masud, Md Kamrul Hasan, Md Abdul Awal, Awal Ahmed Fime, Md Tahmid Hasan Fuad, Delowar Sikder, et al. 2022. Bangla Natural Language Processing: A Comprehensive Analysis of Classical, Machine Learning, and Deep Learning Based Methods. IEEE Access (2022).

[37] Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. Nlp-cuet@ dravidianlangtech-eacl2021: Offensive language detection from multilingual code-mixed text using transformers. arXiv preprint arXiv:2103.00455 (2021).

[38] Sadia Sharmin and Danial Chakma. 2021. Attention-based convolutional neural network for Bangla sentiment analysis. Ai & Society 36 (2021), 381--396.

[39] P Suresh and S Muthu Kumaran. 2015. Sentiment analysis of product reviews using LDA method based on customer text content. International Journal of Science, Engineering and Computer Technology 5, 12 (2015), 427.

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

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems 30 (2017).

[42] Chuhan Wu, Fangzhao Wu, Junxin Liu, Yongfeng Huang, and Xing Xie. 2019. Sentiment lexicon enhanced neural sentiment classification. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 1091--1100.