2021 International Conference on Information and Communication Technology
for Sustainable Development (ICICT4SD), 27-28 February, Dhaka

2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) | 978-1-6654-1460-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICT4SD50815.2021.9396910

Bengali Sentiment Analysis of E-commerce Product Reviews using K-Nearest Neighbors

Mst. Tuhin Akter [*], Manoara Begum [*] and Rashed Mustafa [†]
[*]Dept. of Computer Science & Engineering, Port City International University, Chattogram, Bangladesh.
[†] Dept. of Computer Science and Engineering, University of Chittagong, Chattogram, Bangladesh.
Email: mst.tuhinsimu@gmail.com, manoara.cse@portcity.edu.bd, rashed.m@cu.ac.bd

***Abstract*— The sentiment analysis of the Bengali language is converting into a trendy research topic nowadays. Sentiment analysis is a useful technique in opinion mining, emotion extraction, and trend predictions. By sentiment analysis, the actual sentiment of a text review can be extracted. Every day, every second's people use the internet for different purposes and leave their opinions or perspective views in various places on the internet as a text format. The opinion or review on the internet can contain the author's positive, negative, and neutral views of the statement. This study proposed a machine learning-based model to predict a user's sentiment (positive, neutral, and negative) of a Bangla text review. We have applied five machine learning algorithms in our dataset, which we manually gathered from a Bangladeshi e-commerce site called "Daraz." We have experimented with Random Forest classifier, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost algorithms with our dataset. KNN performs great among all these five algorithms in all the performance measures of accuracy, precision, recall, and f1-score. KNN shows 96.25% accuracy, 0.96 in each of precision, recall, and f1-score.**

Keywords*— *Sentiment Analysis, Machine Learning, SVM, KNN, TF-IDF, Bangla Sentiment Detection.

I. Introduction

The number of Bengali speaking people is increasing day by day, and the Bangla language is becoming famous worldwide. According to a report of [1], the Bangla language ranked seven among the 100 most spoken languages globally. There are 265 million natives and non-native speakers worldwide who speak in Bangla. This considerable amount of speakers express their opinion in Bangla on the internet for various purposes, such as writing a comment for a blog or news, leaving a product review, posting on the social media platform, etc. Sentiment detection is a technique to understand the users' sentiment or perception from a comment or review. As a result of Internet users' rapid growth, the number of opinions on the Internet increases. A sentiment detection system can use for sentiment analysis, better decision-making, and improvements in products, services, etc. This study proposes a machine learning-based model for sentiment detection of e-commerce product reviews for the Bangla language.

This study proposes a machine learning-based model for sentiment detection of Bangla reviews of e-commerce products. We propose five machine learning algorithms-Random Forest classifier, Logistic Regression, SVM, KNN, and XGBoost for sentiment analysis.

The reason behind this research is following:

- Lack of research works for sentiment analysis of product reviews in the Bengali language.
- The sentiment of a review can be used in business analysis to better a business.

The contribution of this study is to follow:

- Work with an emerging and new domain of Bengali Sentiment analysis.
- Improvement of some existing research work on the same domain.
- Proposed a Bengali sentiment analysis model and experimented with a different dataset.

We arrange the rest of the section of this paper as follows. In Section II, some literature review of similar study has been marked. Section III described our dataset property. In Section IV, we elaborately present our proposed methodology. In Section V and VI, we present the Result and discussion and conclusion, and future direction, respectively.

II. Literature Review

Hossain et al. [2] stated a two-class sentiment polarities detection - positive and negative for the Bengali book review. They collected 2000 textbook reviews from different internet sources, such as blogs, Facebook, and e-commerce sites. They choose Logistic Regression, Naive Bayes, SVM, and SGD algorithm for their study and found the highest accuracy in Multinomial Naive Bayes, 84%. Aspect-based sentiment analysis is proposed by Haque et al. [3]. For the Cricket dataset, SVM achieves a 37% f1-score, and LR achieves a 43% f1-score for the Restaurant dataset.

Sharmin et al. [4] has suggested Attention-based sentiment analysis using CNN. CNN achieves 72.06% accuracy for the Cricket dataset. Jiao et al. [5] used LSTM for driver sleepiness detection from EEG and EOG signals. 98% accuracy they have gained from that study. You et al. [6] detect automatic seizures using the generative adversarial network (GAN) by unsupervised learning. This experiment on 12 patients with various types of epilepsy achieves 96.6%. A machine learning-based empirical framework for sentiment analysis has been proposed by Tabassum et al. [7]. After applying unigram, POS tagging, and negation handling, RF achieves 87% accuracy. Sharif et al. [8] stated a machine learning-based sentiment polarity detection for Bengali restaurant reviews. The only algorithm they used for their study is Multinomial Naive Bayes for 1000 data, and the algorithm achieves 80.48%. They used positive and negative sentiment.

978-1-6654-1460-9/21/$31.00 ©2021 IEEE

Soumik et al. [9] presented a machine learning-based model for sentiment analysis of Google Play apps reviews of Bengali language for three sentiment classes- positive, negative, and neutral. They collected 10000 reviews from different apps such as Bkash, Daraz, Ubar. They use several machine learning algorithms- Naïve Bayesian Classifier, Support Vector Machine, Logistic Regression, Bagging Meta-estimator, Adaptive Boosting, Gradient Boosting, Extreme Gradient Boosting with TF-IDF Vectorizer. Gradient Boosting achieves 76.95% accuracy. Azmin et al. [10] stated a machine learning model with a Naive Bayes classifier with various features such as TF-IDF, POS tagger, n-grams, and stemmer for the Bangla language. They used 4200 data collected from Facebook posts and blogs for three sentiment classes happy, sad, and angry. Multinomial Naive Bayes algorithms achieve 78.60% accuracy. Wahid et al. [11] used deep learning for sentiment classification of Bangla text. Recurrent Neural Network with LSTM has applied for positive, negative, and neutral sentiment. They apply the algorithm with the ABSA dataset, which contains 10000 comments about cricket. LSTM attains 95% accuracy for this ABSA dataset.

Chowdhury et al. [12] experimented with a machine learning technique for movie reviews of the Bengali language. They apply SVM and LSTM algorithms for positive and negative sentiment with manually collected 4000 samples about movies. SVM achieves 88.90 % accuracy while LSTM achieves 82.42% accuracy. Sarowar et al. [13] suggested a hybrid machine learning model for the Bangla language sentiment analysis. They gathered 25000 data using web scraping from various e-commerce sites in Bangladesh. They experimented with many algorithms, for example, the Logistic Regression algorithm, PCA-based CNN, Random Forest, and KNN based SVM with their dataset for positive and negative sentiment. KNN based SVM achieves an 82.92% f1-score. Ma et al. [14] proposed aspect-based sentiment analysis using Sentic LSTM. They experiment with their study with SentiHood and SemEval 2015 dataset and achieves the highest 89.63% on SentiHood by Sentic LSTM of AffectiveSpace. Aspect-based sentiment analysis using CNN has been proposed by Rahman et al. [15]. They applied CNN for Cricket and Restaurant dataset for positive, negative, and neutral sentiment. The Cricket dataset contains 2900 records, while the restaurant dataset contains 2600 records. CNN achieves 51% and 64% f1-score for Cricket and Restaurant dataset, respectively.

Al-Amin et al. [16] stated a sentiment analysis model with Word2vec. Word2vec model achieves 75.5% accuracy for 16000 records with positive and negative sentiment. Alam et al. [17] presented a Convolutional Neural Network-based sentiment analysis technique for Bangla sentences. They use 850 data for the sentiment class of positive, negative, and neutral with the CNN algorithm. Their suggested CNN algorithm attains 99.87% accuracy.

III. Dataset Property

A. Dataset Properties

This study used a dataset, which has been compiled from Daraz [18]. This data are consists of different product reviews of the Bangla language. Our dataset contains 7905 unique data and contains nine columns – Username, Category, SubCategory, ProductType, ProductName, Comment, CommentDate, Rating, and Sentiment. The pie chart representation of sentiment distribution in the dataset is shown in Fig. 1.

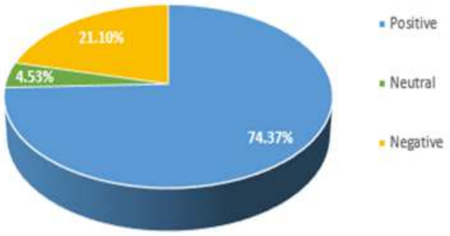

Fig 1. Pie chart representation of sentiment distribution.

IV. Prpposed Methodology

This study proposed a supervised machine learning-based model for sentiment analysis for e-commerce product reviews of the Bangla language. We follow the below steps throughout this study:

- Collection of data
- Label the sentiment
- Pre-process of data
- Feature extraction
- Oversampling
- Sentiment detection

We use Random Forest classifier, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost algorithms with our dataset followed by the steps mention above. Fig. 2 represents a diagram of our proposed methodology.

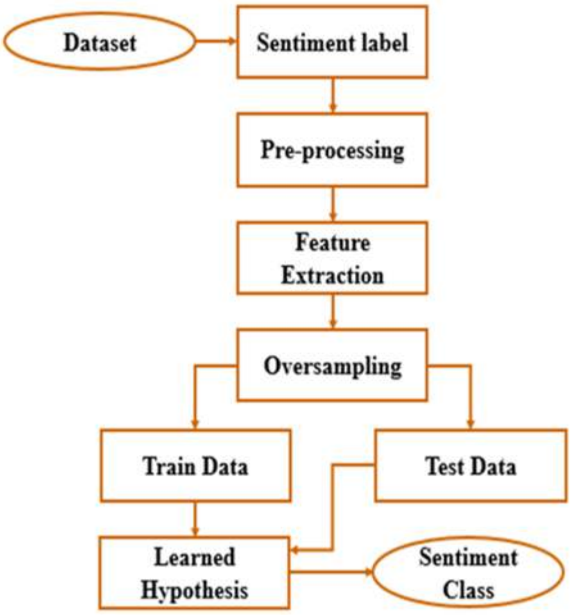

Fig. 2. Diagram of the proposed model.

Lack of research work on the Bengali sentiment analysis domain is the main reason for this study. Sentiment analysis can be an essential element in business analysis and help decide on products for shoppers.

A. *Collection of Data*

We've obtained datasets from "Daraz," an e-commerce website in Bangladesh. Our dataset contains reviews from various products such as electronics, clothes, and consumer goods. We have manually collected data from [18] where review text and rating are found. We only concern with product reviews, which are in the Bengali Language. We collect eight types of information, such as the Username of the reviewer, Category, Subcategory, ProductType, ProductName, Comment, CommentDate, and Rating. We collect a total of 7905 unique reviews about products. There is no sentiment data on that website.

B. *Label the Sentiment Class*

We collected dataset contains a label of rating ranging from 1 to 5. There is no sentiment label in this dataset. So we need to label our sentiment based on this ratio. We label the sentiment according to Table I.

TABLE I. SENTIMENT LABEL

Rating	Sentiment
Rating > 3	Positive
Rating == 3	Neutral
Rating < 3	Negative

C. *Pre-processing of Data*

We perform several tasks in the pre-processing stage. We first remove non-Bengali text from our review. Then we separate emoji, Bangla digit, and punctuation. After that, we tokenize the review produced from the above task. The next two final tasks are to eliminate stop words and stemming.

D. *Feature Extraction*

Feature extraction is a technique to reduce the dataset dimensionality. We need to extract the feature from our data because many dimensionalities need many computing resources to process. Feature extraction can reduce the dimensionalities of data efficiently and prepare our data to learn. For feature extraction, we use the TF-IDF vectorizer. TF-IDF has an advantage over the CountVectorizer. CountVectorizer only calculates the frequency of words in the corpus, while TF-IDF works with both term frequency and document frequency. We finally set the n-gram value of (1, 3) in TD-IDF after inspecting other n-gram values.

E. *Oversampling*

We see from Fig. 1 our dataset is imbalanced because all the sentiment label is not equally distributed. We can balance our dataset using a technique called Oversampling. Oversampling does the trick of equal distribution for all sentiment classes. The pie chart presentation after applying the oversampling technique to the dataset is in Fig. 3. By oversampling, the number of records increases from 7905 to 17637.

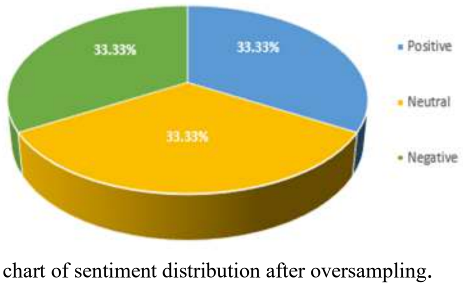

Fig. 3. Pie chart of sentiment distribution after oversampling.

F. *Sentiment Detection*

We divided the total dataset into 80% train data and 20% test data. We have applied five machine learning algorithms to the dataset.

1) Random Forest Classifier (RF): RF is the first algorithm for this study. The confusion matrix and ROC curve produced by RF are shown in Fig. 4 & Fig. 5.Here, we get an accuracy of 90.84%.

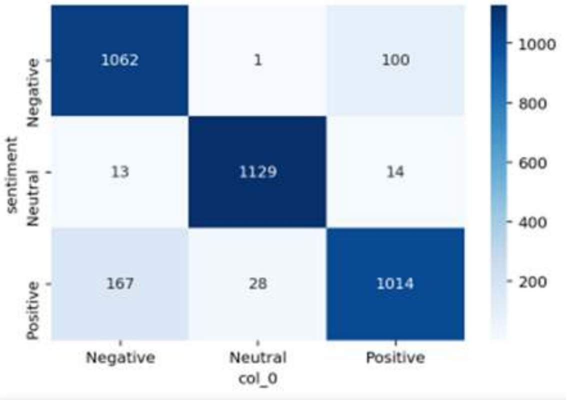

Fig. 4. Confusion Matrix of RF.

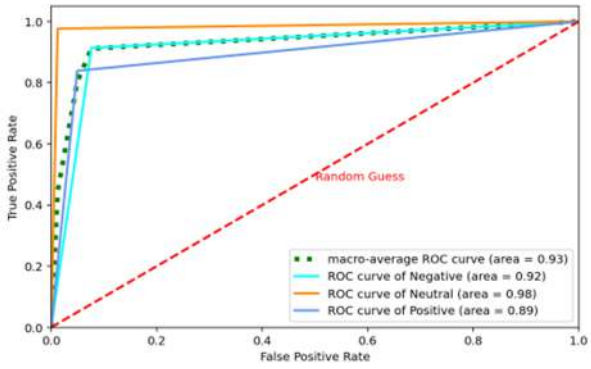

Fig. 5. ROC curve of RF.

2) Logistic Regression (LR): The confusion matrix and ROC curve generated by LR are shown in Fig. 6 & Fig. 7. At this stage, an accuracy of 90.33% is obtained.

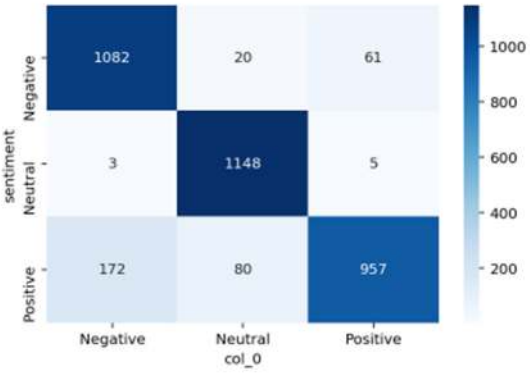

Fig. 6. Confusion Matrix of LR.

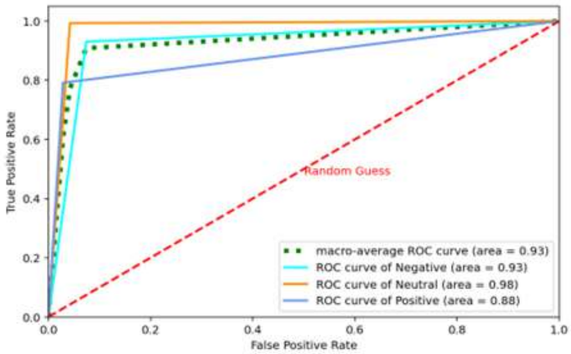

Fig. 7. ROC curve of LR.

3) Support Vector Machine (SVM): The confusion matrix and ROC curve created by SVM are shown in Fig. 8 & Fig. 9. At this time, we attain an accuracy of 94.35%.

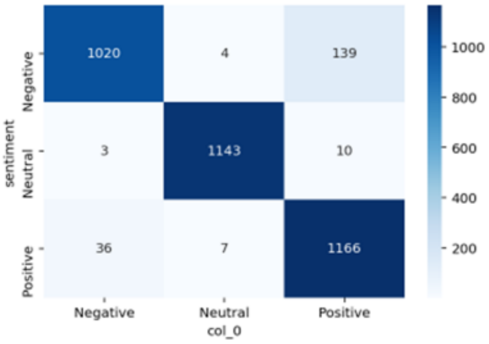

Fig. 8. Confusion Matrix of SVM.

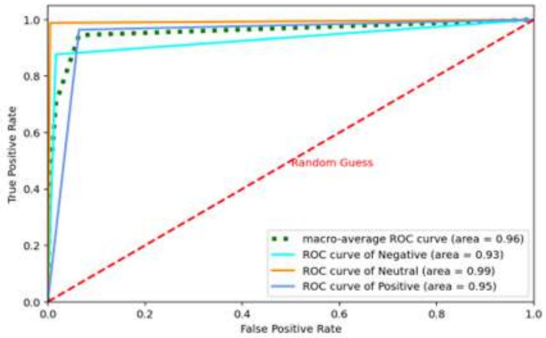

Fig. 9. ROC curve of SVM.

4) K-nearest neighbors (KNN): The confusion matrix and ROC curve produced by KNN are shown in Fig. 10 & Fig. 11. Here, an accuracy of 96.25% is received.

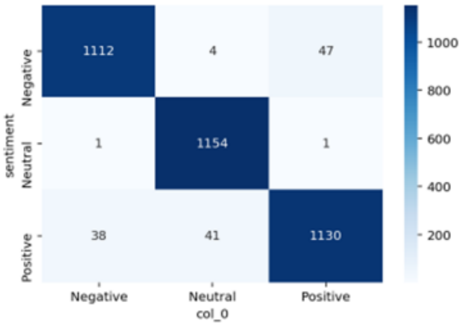

Fig. 10. Confusion Matrix of KNN.

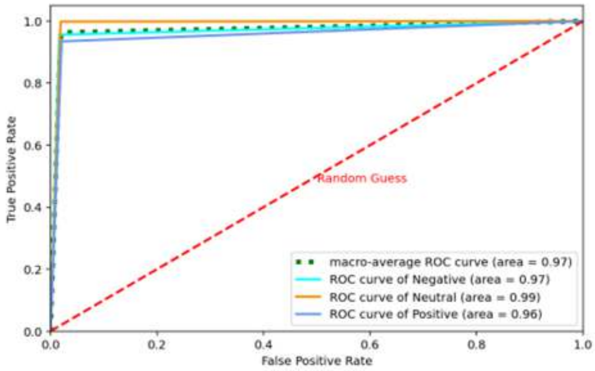

Fig. 11. ROC curve of KNN.

5) XGBoost: The confusion matrix and ROC curve generated by XGBoost are shown in Fig. 12 & Fig. 13. We get an accuracy of 90.56% at this point.

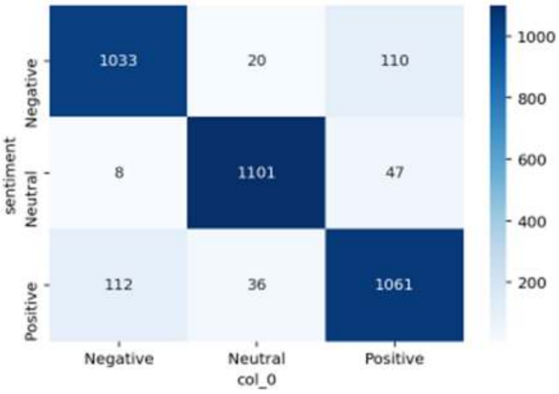

Fig. 12. Confusion Matrix of XGBoost.

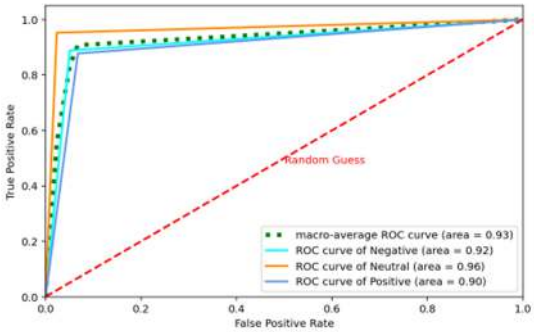

Fig. 13. ROC curve of XGBoost.

V. Result & Discussion

We have gone through many steps between data collection and sentiment detection. Five machine learning algorithms have been used for the sentiment detection of our dataset. We use a standard train test split ratio of 80% and 20% of our algorithms' data. Tenfold cross-validation has been done for all the algorithms. Table 2 represents the cross-validation score for every fold and the mean score for all the algorithms. Table 3 shows that KNN performs better and obtains a higher accuracy of 96.25% and a better f1-score of 96%.

TABLE II. CROSS-VALIDATION SCORE

K-fold	RF	LR	SVM	KNN	XG-Boost
1	82.93	87.87	81.75	94.84	69.61
2	85.26	87.59	84.69	95.80	81.80
3	88.32	90.99	86.85	97.22	85.03
4	94.78	92.74	98.47	98.07	94.55
5	94.16	92.12	98.02	96.42	94.39
6	92.97	90.70	96.60	97.17	92.80
7	95.46	90.42	97.39	97.22	94.44
8	93.59	92.40	97.90	96.54	94.55
9	91.15	86.16	94.90	96.03	89.05
10	94.84	93.02	98.99	97.62	95.34
Mean	91.35	90.40	93.55	96.69	89.16

TABLE III. ACCURACY, PRECISION, RECALL, AND F1-SCORE COMPARISON OF DIFFERENT ALGORITHMS

Algorithm	Accuracy	Precision	Recall	f1-score
Random Forest	90.84	0.91	0.91	0.91
Logistic Regression	90.33	0.90	0.91	0.90
SVM	94.35	0.94	0.95	0.94
KNN	96.25	0.96	0.96	0.96
XGBoost	90.56	0.91	0.91	0.91

VI. CONCLUSION & FUTURE WORKS

Sentiment detection can be an efficient technique to understand the author's sentiment from a text. This paper proposed a supervised machine learning technique to detect the sentiment from the Bangla language text. This study finds the most outstanding result from the KNN algorithm after applying five machine learning algorithms with the TF-IDF Vectorizer. KNN attains 96.25% accuracy on the applied dataset. In the future, we will use more datasets to train the model. Again, the inclusion of POS tagging, synonym analysis can improve model accuracy.

REFERENCES

[1] "Bangla ranked at 7th among 100 most spoken languages worldwide," Dhaka Tribune, 17-Feb-2020. [Online]. Available: https://www.dhakatribune.com/world/2020/02/17/bengali-ranked-at-7th-among-100-most-spoken-languages-worldwide. [Accessed: 13-Mar-2021].

[2] E. Hossain, O. Sharif, and M. M. Hoque, "Sentiment Polarity Detection on Bengali Book Reviews Using Multinomial Naive Bayes," *arXiv.org*, 06-Jul-2020. [Online]. Available: https://arxiv.org/abs/2007.02758. [Accessed: 13-Mar-2021].

[3] S. Haque, T. Rahman, A. K. Shakir, M. S. Arman, K. B. Biplob, F. A. Himu, D. Das, and M. S. Islam. "Aspect based sentiment analysis in Bangla dataset based on aspect term extraction." In International Conference on Cyber Security and Computer Science, pp. 403-413. Springer, Cham, 2020.

[4] S. Sharmin, and D. Chakma. "Attention-based convolutional neural network for Bangla sentiment analysis." AI & SOCIETY (2020): 1-16.

[5] Y. Jiao, Y. Deng, Y. Luo, and B. L. Lu, "Driver sleepiness detection from EEG and EOG signals using GAN and LSTM networks," Neurocomputing, vol. 408, pp. 100–111, 2020.

[6] S. You, B. H. Cho, S. Yook, J. Y. Kim, Y. M. Shon, D. W. Seo, and I. Y. Kim, "Unsupervised automatic seizure detection for focal-onset seizures recorded with behind-the-ear EEG using an anomaly-detecting generative adversarial network," Computer Methods and Programs in Biomedicine, vol. 193, p. 105472, 2020.

[7] N. Tabassum, and M. I. Khan. "Design an empirical framework for sentiment analysis from Bangla text using machine learning." In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1-5. IEEE, 2019.

[8] O. Sharif, M. M. Hoque, E. Hossain, "Sentiment analysis of Bengali texts on online restaurant reviews using Multinomial Naïve Bayes," In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) 2019 May 3 (pp. 1-6). IEEE.

[9] M. M. Soumik, S. S. Farhavi, F. Eva, T. Sinha, m. S. Alam, "Employing machine learning techniques on sentiment analysis of Google Play store Bangla reviews," In 2019 22nd International Conference on Computer and Information Technology (ICCIT) 2019 Dec 18 (pp. 1-5). IEEE.

[10] S. Azmin, K. Dhar, "Emotion detection from Bangla text corpus using Naïve Bayes classifier," In 2019 4th International Conference on Electrical Information and Communication Technology (EICT) 2019 Dec 20 (pp. 1-5). IEEE.

[11] M. F. Wahid, M. J. Hasan, M. S. Alom, "Cricket sentiment analysis from Bangla text using Recurrent Neural Network with Long Short-Term Memory model," In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) 2019 Sep 27 (pp. 1-4). IEEE.

[12] R. R. Chowdhury, M. S. Hossain, S. Hossain, K. Andersson, "Analyzing sentiment of movie reviews in Bangla by applying machine learning techniques," In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) 2019 Sep 27 (pp. 1-6). IEEE.

[13] M. G. Sarowar, M. Rahman, M. N. Ali, O. F. Rakib, "An automated machine learning approach for sentiment classification of Bengali E-Commerce sites," In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT) 2019 Mar 29 (pp. 1-5). IEEE.

[14] Y. Ma, H. Peng, T. Khan, E. Cambria, and A. Hussain, "Sentic LSTM: a Hybrid Network for Targeted Aspect-Based Sentiment Analysis," *Cognitive Computation*, vol. 10, no. 4, pp. 639–650, 2018.

[15] M. A. Rahman, E. K. Dey, "Aspect extraction from Bangla reviews using Convolutional Neural Network," In 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR) 2018 Jun 25 (pp. 262-267). IEEE.

[16] M. A. Amin, M. S. Islam, S. D. Uzzal, "Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words," In 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE) 2017 Feb 16 (pp. 186-190). IEEE.

[17] M. H. Alam, M. M. Rahoman, M. A. Azad, "Sentiment analysis for Bangla sentences using convolutional neural network," In 2017 20th International Conference of Computer and Information Technology (ICCIT) 2017 Dec 22 (pp. 1-6). IEEE.

[18] "Online Shopping In Bangladesh With Home Delivery," Online Shopping In Bangladesh With Home Delivery - Daraz.com.bd. [Online]. Available: https://www.daraz.com.bd/. [Accessed: 13-Mar-2021]