Bangla Text Sentiment Analysis Using Supervised Machine Learning

Abdullah Al Mahdi
Faculty of Science and Technology
American International University - Bangladesh (AIUB)
Dhaka, Bangladesh

Muhammad Shahriar Zaman
Faculty of Science and Technology
American International University - Bangladesh (AIUB)
Dhaka, Bangladesh

Muhammad Rafid Rahman
Faculty of Science and Technology
American International University - Bangladesh (AIUB)
Dhaka, Bangladesh

Syeda Faiza Karima
Faculty of Science and Technology
American International University - Bangladesh (AIUB)
Dhaka, Bangladesh

***Abstract*— Natural language processing researchers are increasingly interested in sentiment analysis (SA) as a result of the growth of social digital material on the Internet (NLP). Bangla has a complicated grammatical structure in text, which has led to some substantial studies in this area. In the context of Bangla, this essay focuses on SA. In this study, phrase polarity is identified using a brand-new rule-based method called Bangla Text Sentiment Score (BTSC). This system takes stop words and special characters into account when calculating a word's score, as well as sentences and blogs. The BTSC method is used to generate scores for a Bangla dataset and extract sentiments from it. Third, using the respective BTSC scores, two feature matrices are created for the two datasets. The feature matrices are then subjected to supervised machine learning classifiers such as SVM, KNN, Logistic Regression, and Decision Tree. According to the results, the logistic regression had the maximum accuracy, which is close to 73%. These demonstrate the BTSC algorithm's efficiency in Bangla SA.**

Keywords—Sentiment analysis, Bangla NLP, Tf-Idf, SVM, BTSC, N-grams, Bi-grams, KNN, Logistic Regression, Decision Tree

I. INTRODUCTION

Sentiment analysis (SA), also known as opinion mining [1], is a branch of study that uses peoples' feelings, attitudes, and other characteristics to forecast the polarity of public opinion or textual data from microblogging sites [2] on a hotly debated issue. Researchers are rapidly gaining interest in SA due to the topic's relevance to natural language processing (NLP) in the machine learning field and the abundance of opinionated data on the Internet. Nowadays, individuals share their opinions on a certain product or item via social media platforms, newspapers, blogs, etc. Additionally, there are forum discussions, opinions on particular topics, comments, and feelings. If a phrase or text lacks any fundamental opinion words, there may be several obstacles in the way of recognizing binary or ternary class feeling, such as subjectivity or opinion-based identification. The machine learning framework has drawn a lot of attention since it can categorize this data into good, negative, or neutral classes based on sentiment. This is due to the construction of models using adaptable feature extraction, alternating, forecasting using probabilistic theory, and constructing useful feature matrix representations across numerous language domains [4]. For this kind of study, several aspects have been noted, including the bag of words (Bo W) model, lexical analysis, and semantic features [5]. Language affects this matrix feature. More than 250 million people speak Bangla, an ancient Indo-European tongue [6]. The ability to extract sentiment from the Bangla language will thus be crucial for NLP researchers to advance machine learning significantly. The Bangla Text Sentiment Score (BTSC), a powerful and original rule-based system, is utilized in this study to identify sentence polarity and to better extract sentiment from chunks of Bangla text. We develop an automated system that can extract opinions from reviews of the Bangla dataset, and that automated system will be categorized using supervised machine learning techniques [8] and the N-grams model. This is due to the fact that the author [9] discovered this model to be effective in text categorization. The remaining sections of the paper are structured as follows: they summarize the relevant research studies, present our system methodology, and illustrate the evaluation outcomes using trained and tested datasets using supervised classification techniques like SVM, logistic regression (LR), K-nearest neighbors (KNN), etc. An overview of the research's conclusions is given at the end.

II. RELATED WORK

SA is now a hot issue among scholars in the era of the growth of social media and microblogging websites. A lot of linguistic areas, including English, French, Chinese, Arabic, etc., apparently practice SA. Due to several technological and empirical limitations, its advancement in the Bengali language has been minimal [10]. This publication [11] has been a major source of inspiration for us because, as far as we are aware, no research has used SA in Bengali while employing an enlarged lexicon. The outcomes of experiments utilizing an Arabic lexicon-based data dictionary have thus far been more successful [12]. Authors Alshari et al. [13] employed an emotional lexicon dictionary based on word2vec to execute SA and labeled SentiWordNet (SW) as a curse of dimensionality. Additionally, the author of the Bangla text [14] used a TF-IDF vectorizer to preprocess the data for a SA and then classified the data using an SVM algorithm, but they did not measure the polarity by determining the score of the text. As a result, it is necessary to identify the polarity of each sentence using a particular rule-based [15] algorithm. To execute a SA employing SW

by translating Bangla phrases to English, the author of Chowdhury and Chowdhury [16] developed a semi-supervised bootstrapping strategy in SVM and maximum entropy (MaxEnt) classifier. They have only counted positive and negative word polarity by SW in their bootstrapping rule-based technique, which only functions for short texts. In Azharul Hasan et al. [17], authors suggested a way of employing an XML-based POS tagger and SW to determine the sentiment from Bangla text using valency analysis. They have employed SW and WordNet (WN), which were created exclusively for the English language. In order to determine the word score or polarity from the text, a lexicon-weighted Bangla word dictionary is required. In addition, in Islam et al. [18], writers extracted positive and negative (bi-polar) polarity from the Facebook text by tokenizing adjective terms using POS tagger, performing valence shifting negative words at the right side of a phrase, and replacing them with antonym words using SW. Due to SW's flaw in providing accurate polarity in Bangle text, the authors in [19] discussed an automated system for emotion detection by mapping each text to an emotion class. Their accuracy was 90%, but labeling the data took more time, and their phrase patterns were only formed for three sub-categories of sentiment that aren't used in complex sentences. Only positive and negative terms from their feature word list dictionary were used in Tabassum and Khan's [20] framework for SA. In Zhang et al. [21], the authors built an expanded sentiment dictionary and used a rule-based classifier to categorize the field of text polarity by calculating a sentence's score. By examining the occurrences of an emotive feature word in tagging each phrase, the authors Akter and Aziz [22] described a lexicon-based dictionary model.

III. Methodology

The major objective of this study is to use a machine-learning technique and a special rule-based algorithm to assess the sentiment from Bangla text. We have broken down our entire project into three pieces in order to determine sentiment polarity from the raw text. Our suggested strategy is depicted in Figure 1 and these phases are discussed below. The following goals have been determined in order to reach the objective:

a. Data extraction and preprocessing should be done appropriately.

b. To extract scores from a section of Bangla text in order to create a new and efficient rule-based system for identifying sentence polarity categorization.

c. To examine the feature matrix using the intended dataset, assess our theoretical assertion, and then assess how well our work aligns with some previously published studies on the supervised machine learning technique.

IV. Data Preprocessing

Our SA uses supervised machine learning to classify documents based on their sentiment. Following the collection of corpus data, we must preprocess them using the different processes outlined below.

A. Data Cleaning

After extracting data from the source CSV file, we stored it in a data frame. Then all the unnecessary characters were removed by matching them with regular expression patterns.

B. Data Tokenization

Splitting the sentence into a word list is called a tokenization process. Each token is called a word. For example: “আচারের সংযোজন খুব ভালো ছিল।” [The addition of the pickle was very good], after tokenizing this sentence it will create a list, as like [“আচারের” [pickle], “সংযোজন” [addition], “খুব” [very], “ভালো” [good], “ছিল” [was]]. While doing the tokenization process we also finished normalizing the data. Normalizing means removing characters [“,”, “.”, “!”, “@”, “#”, “%”], etc. and stop words [28] from the sentence. The characters and stop words will have on producing training and test data and building machine learning models.

C. Data Stemming

After performing the tokenization procedure, we applied stemming which refers to selecting the root word from the provided word list. For this research we have removed words like “র”,“এর”,“গুলি”,“গুলো”,“টার”, “টি” etc.,

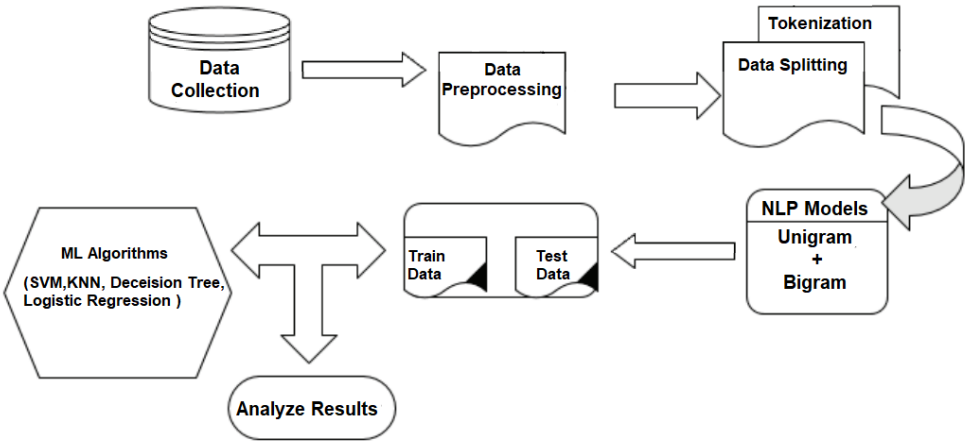

Figure 1: Visualization of proposed system architecture

For example: “স্বাধীনতার” [Independent], “বাংলাদেশের” [Bangladesh], “দুর্বলতাগুলির” [Weaknesses] words convert the root word into respectively “স্বাধীনতা”, “বাংলাদেশ”, “দুর্বলতা” by stemming process. After completion of preprocessing the data frame was stored in a new CSV file. The columns in this case were a label, sentence, clean sentence, and clean data.

D. Vectorization using the n-gram approach

The feature extraction process for data used in machine learning models, known as vectorization, is crucial. By translating text to numerical vectors, the goal in this situation is to extract some identifiable characteristics from the text for the model to train on. A collection of text documents may be turned into a matrix of token counts using the count vectorization method with the help of the sci-kit-learn module CountVectorizer, which was imported from the open-source Python sklearn library. To assist in identifying recurring text patterns in a corpus of text documents, such as web pages or product descriptions, CountVectorizer may output a matrix listing the frequency of each word or phrase.

We vectorized both the words and the phrases in our datasets for this study.

V. Data Splitting

For this step, the training set will receive 80% of our data. Using the train test split() function, which was imported from the sklearn package, a total of 2449 records were split into 1959 training records and 490 testing records.

VI. Model Training

We have used ML techniques including SVM, Logistic Regression, KNN, and Decision Tree for the purposes of this research.

All of them were imported from the sklearn package, and after setup, the fit() method was used to train the models. When using the fit() function to train the imported classifiers, we supplied two arrays, x train, and y train, as parameters because we are using supervised learning.

Since there are only three classes, we picked a linear kernel for SVM, and we selected five as the n number for KNN.

VII. Result Evaluation

Here, a function to compute accuracy and depict the confusion matrices has been developed. To do this, we utilized the accuracy score() function to determine the accuracy and imported metrics from the sklearn package. We have also loaded the Seaborn and matplotlib.pyplot modules because we are working with statistical data representation.

After testing, we found that the accuracy of the logistic regression model was 73%, the SVM was 69.5%, the KNN was 43.85%, and the decision tree was 57.9%.

	A	B
1	**polarity**	**sentence**
2	ntr	সরকারের উচিৎ টাইগারদের দেশে ফিরিয়ে নিয়ে আসা।
3	neg	অভিনয় কত আর করবি! সমালোচনা ভয়ে অবসর।
4	neg	জনগনকে একটা খেলা দেখিয়ে এখন আবার দলে ফিরলো। এখন কোন মুখে ফেরার ঘোষনা দেয়
5	pos	আপনাকে অনেক অনেক শুভেচ্ছা,আপনি আমাদের গর্ব
6	ntr	এটা অন্যতম কারণ
7	pos	নিঃসন্দেহে প্রসংশনীয়।
8	pos	শুভ বুদ্ধির উদয় হোক , আত্মা শুদ্ধ হোক এই কামনাই করি

Figure 2: Source dataset preview

	A	B	C	D
1	**polarity**	**sentence**	**cleaned_sentence**	**clean_data**
2	neg	"এ ব্যর্থতা সরকারেরই!"	এ ব্যর্থতা সরকারেরই	ব্যর্থতা সরকারেরই
3	pos	"<3 ওস্তাদ তোমাকে ভালবাসি খুব লেখাগুলো তার থেকেও বেশি!!!"	ওস্তাদ তোমাকে ভালবাসি খুব লেখাগুলো তার থেকেও বেশি	ওস্তাদ তোমাকে ভালবাসি লেখাগুলো থেকেও বেশি
4	pos	"অন্তর থেকে ধন্যবাদ কবি।।"	অন্তর থেকে ধন্যবাদ কবি	অন্তর ধন্যবাদ কবি
5	neg	":D চুরি করার মধ্যে এক নাম্বার !!!"	চুরি করার মধ্যে এক নাম্বার	চুরি করার এক নাম্বার
6	ntr	"আঘোরীদেরকে বলা হয় পৃথিবীর সেরা কালো জাদুকর।"	আঘোরীদেরকে বলা হয় পৃথিবীর সেরা কালো জাদুকর	আঘোরীদেরকে বলা পৃথিবীর সেরা কালো জাদুকর

Figure 3: Source dataset after preprocessing

```
Decision Tree accuracy :  0.5795918367346938
confusion_matrix:
 [[165  70  37]
 [ 28 112   7]
 [ 32  32   7]]
```

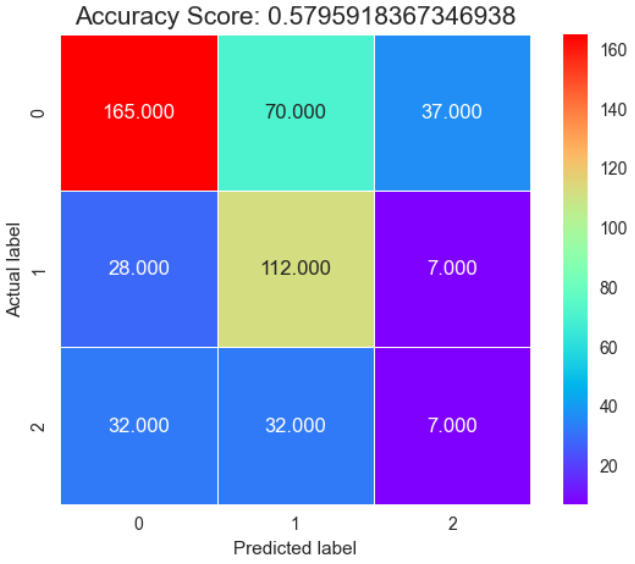

Figure 4: Decision tree accuracy and confusion matrix

```
KNN accuracy :  0.4387755102040816
confusion_matrix:
 [[ 87 129  56]
 [ 24 113  10]
 [ 24  32  15]]
```

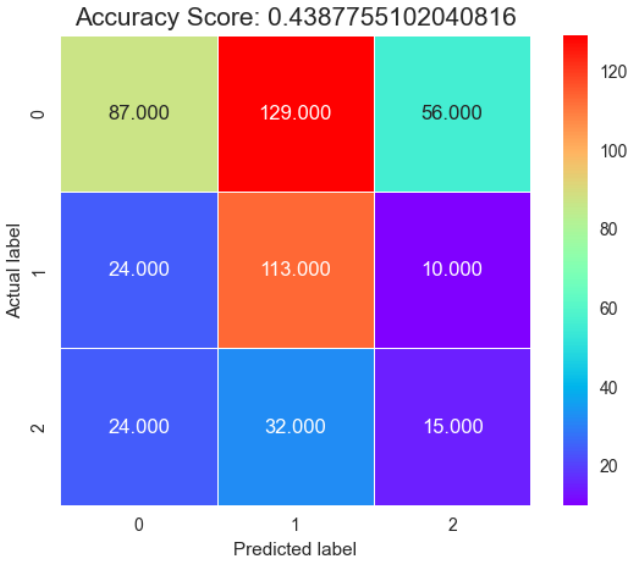

Figure 5: KNN accuracy and confusion matrix

```
Logistic Regression accuracy :  0.7346938775510204
confusion_matrix:
 [[254  14   4]
 [ 46  99   2]
 [ 52  12   7]]
```

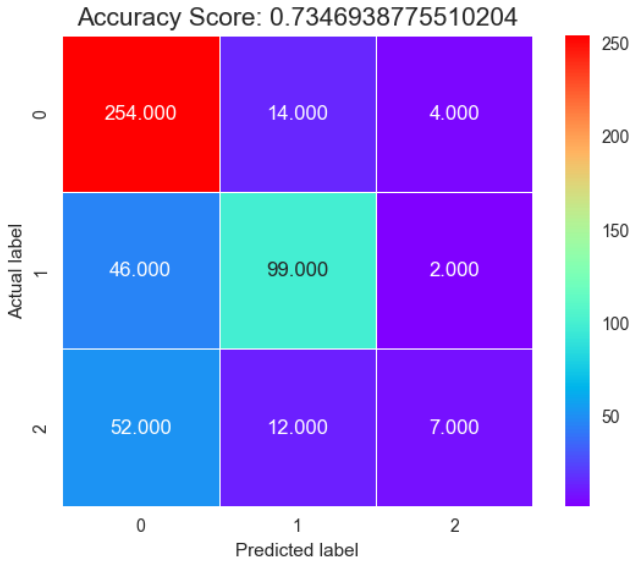

Figure 6: Logistic regression accuracy and confusion matrix

```
SVM accuracy :  0.6959183673469388
confusion_matrix:
 [[232  20  20]
 [ 41  98   8]
 [ 46  14  11]]
```

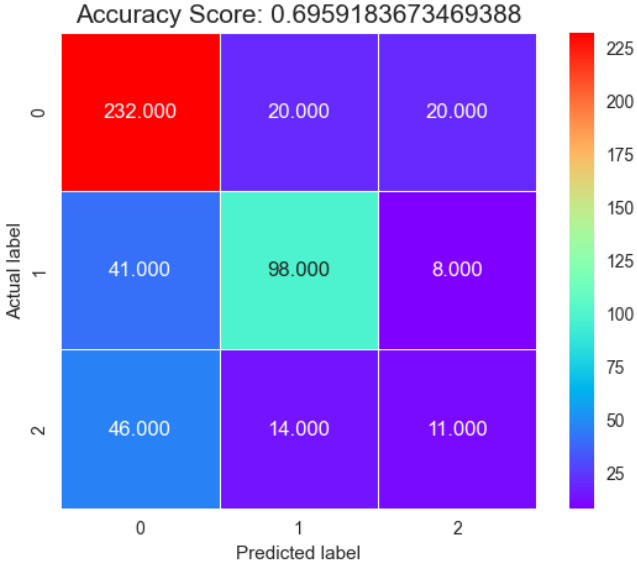

Figure 7: SVM accuracy and confusion matrix

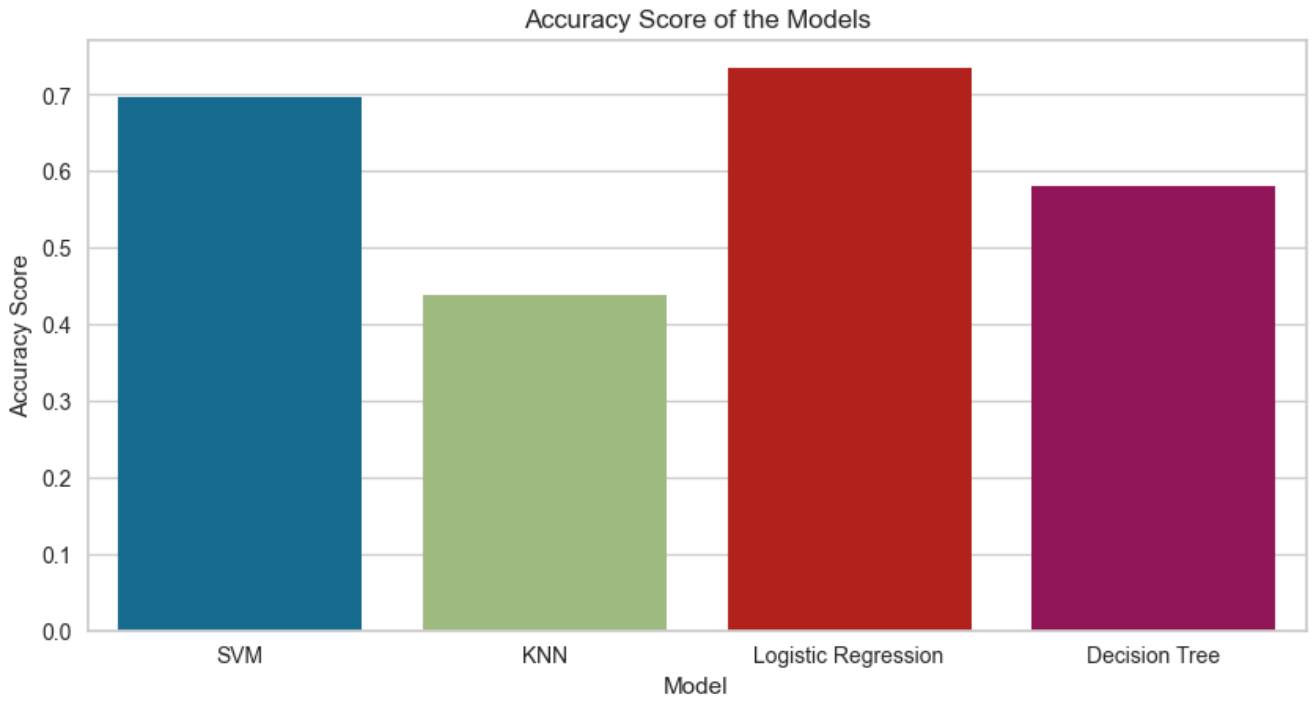

Figure 8: Accuracy Comparison amongst models

VIII. Conclusion

With the number of Internet users increasing quickly, SA is dependent on the dataset of specific content. The manual creation of positive and negative dictionaries with weighted values is still a challenging undertaking when mining data from the Bangla dataset for techniques relying on lexicon-based dictionaries. However, careful examination of this data will produce more specific classifications of polarity. In this study, we used machine learning classifiers such as logistic regression, support vector machine (SVM), K-nearest neighbor, and decision tree to identify the three types of polarity from the texts. Any rule-based technique is necessary for categorical particular domain-based data to identify text category and categorization since a document might belong to more than one category. On the Bigram feature matrix, we had the greatest accuracy of 73% in cricket. Compared to the other two polarities in CM, neutral data identification has received less attention. Every dataset varies in unique ways. Our results will predict more accurately than those produced with the present dataset if we utilize, say, fifty (50) thousand datasets in our machine learning method. Parts-of-speech (POS) taggers and the elimination of more distinct stop words will help us further enhance our outcomes. More datasets will be added to our technique in the future. Approximately 2500 data have been employed as sources in our methodology.

IX. References

[1] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Clay-pool Publishers LLC, 2012.

[2] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in Proceedings of the International Conference on Language Resources and Evaluation (LREc), Valletta, Malta,
2010, vol. 10, pp. 1320–1326.

[3] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37
(2011), 267–307.

[4] E. Boiy, M.-F. Moens, A machine learning approach to sentiment analysis in multilingual web texts, Inf. Retr. 12 (2009), 526–558.

[5] X. Liu, W.B. Croft, Statistical language modeling for information retrieval, Ann. Rev. Inf. Sci. Technol. 39 (2005), 1–31.

[6] Wikipedia, Bengali language. https://en.wikipedia.org/wiki/Bengali_language

[7] S. Buddeewong, W. Kreesuradej, A new association rule-basedtext classifier algorithm, in 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05), IEEE, Hong
Kong, China, 2005, p. 2.

[8] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, N. EightStreet, Stroudsburg, PA, 18360, United States, 2002, vol. 10, pp.
79–86.

[9] K. Dave, S. Lawrence, D.M. Pennock, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, in Proceedings of the 12th International Conference on World Wide Web, Association for Computing Machinery, New York, NY, United States, 2003, pp. 519–528.

[10] M.A. Karim, Technical Challenges and Design Issues in Bangla Language Processing, IGI Global, 2013.

[11] G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng, X. Wu, Chinese text sentiment analysis based on extended sentiment dictionary, IEEE Access. 7 (2019), 43749–43762.

[12] N.A. Abdulla, N.A. Ahmed, M.A. Shehab, M. Al-Ayyoub, Arabic sentiment analysis: lexicon-based and corpus-based, in 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), IEEE, Amman, Jordan, 2013, pp. 1–6.

[13] E.M. Alshari, A. Azman, S. Doraisamy, N. Mustapha, M. Alkeshr, Effective method for sentiment lexical dictionary enrichment based on word2vec for sentiment analysis, in 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), IEEE, Kota Kinabalu, Malaysia, 2018, pp. 1–5.

[14] S.A. Mahtab, N. Islam, M.M. Rahaman, Sentiment analysis on Bangladesh cricket with support vector machine, in 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), IEEE, Sylhet, Bangladesh, 2018, pp. 1–4.

[15] C.J. Hutto, E. Gilbert, Vader: a parsimonious rule-based model for sentiment analysis of social media text, in Eighth International AAAI Conference on Weblogs and Social Media, Ann Arbor, MI,
USA, 2014.

[16] S. Chowdhury, W. Chowdhury, Performing sentiment analysis in Bangla microblog posts, in 2014 International Conference on Informatics, Electronics & Vision (ICIEV), IEEE, Dhaka,
Bangladesh, 2014, p. 1–6.

[17] K.M.A. Hasan, M. Rahman, Sentiment detection from Bangla text using contextual valency analysis, in 2014 17th International Conference on Computer and Information Technology (ICCIT), IEEE, Dhaka, Bangladesh, 2014, pp. 292–295.

[18] S. Islam, A. Islam, A. Hossain, J.J. Dey, Supervised approach of sentimentality extraction from Bengali Facebook status, in 2016 19th International Conference on Computer and Information Technology (ICCIT), IEEE, Dhaka, Bangladesh, 2016,
pp. 383–387.

[19] R.A. Tuhin, B.K. Paul, F. Nawrine, M. Akter, A.K. Das, An automated system of sentiment analysis from Bangla text using supervised learning techniques, in 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), IEEE, Singapore, 2019, pp. 360–364.

[20] N. Tabassum, 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), IEEE, Cox'sBazar, Bangladesh, 2019, pp. 1–5.

[21] Z. Shunxiang, Z. Wei, Y. Wang, T. Liao, Sentiment analysis of chinese micro-blog text based on extended sentiment dictionary, Future Gener. Comput. Syst. 81 (2018), 395–403.

[22] S. Akter, M.T. Aziz, Sentiment analysis on Facebook group using lexicon based approach, in 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), IEEE, Dhaka, Bangladesh, 2016, p. 1–4.

[23] A. Rahman, Bangla absa datasets for sentiment analysis. 2018.https://github.com/AtikRahman/Bangla_ABSA_Datasets

[24] M. Rahman, E.K. Dey, Datasets for aspect-based sentiment analysis in Bangla and its baseline evaluation, Data. 3 (2018), 15.

[25] G. Hub, Adjective, নাম বিশেষণ
http://www.grammarbd.com/engrammar/adjective

[26] G. Hub, Adverb, ক্রিয়া বিশেষণ
http://www.grammarbd.com/engrammar/adverb

[27] S. Hossain, Bltk, the bengali natural language processing toolkit. 2020. https://pypi.org/project/bltk/

[28] N.L. Ranks, Bengali stopwords - ranks nl. https://www.ranks.nl/stopwords/bengali

[29] T. Traylor, J. Straub, N. Snell, Classifying fake news articles usingnatural language processing to identify in-article attribution as a supervised learning estimator, in 2019 IEEE 13th International Conference on Semantic Computing (ICSC), IEEE, Newport Beach, CA, USA, 2019, pp. 445–449.

[30] A.G. Vural, B.B. Cambazoglu, P. Senkul, Z.O. Tokgoz, A framework for sentiment analysis in Turkish: application to polarity detection of movie reviews in Turkish, in: E. Gelenbe, R. Lent (Eds.), Computer and Information Sciences III, Springer, London, England, 2013, pp. 437–445.

[31] 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), IEEE, Dhaka, Bangladesh, 2019, pp. 1–6.