Sentiment Analysis on Bengali Text using Lexicon Based Approach

Rajib Chandra Dey
Dept. of Information & Communication Technology
Comilla University
Comilla - 3506, Bangladesh
rajibict14@gmail.com

Orvila Sarker
School of Computer Science
The University of Adelaide
Adelaide, Australia
orvila.sarker@gmail.com

***Abstract*—In this modern era, we daily involve in the internet strongly. We express our opinion about products, services, books, movies, songs, politics, sports, organizations, etc. through the internet in social media, blogs, micro-blogging websites or any media. Public opinion with Bengali text in internet media is increasing very rapidly. Due to a few works in Bengali text sentiment analysis, it has become an important issue of extracting opinions, emotions from Bengali textual data through Sentiment Analysis (SA) for better knowledge extraction. Sentiment Analysis (SA) is effectively used for classifying the opinion expressed in a text according to its polarity (e.g., positive, negative or neutral). This paper represents a lexicon dictionary-based approach for polarity detection of Bengali text data. We compared our proposed model with machine learning classifiers such as Decision Tree (DT), Naive Bayes (NB) and Support Vector Machine (SVM) classifiers and it works as a much better accurate model for Bengali text polarity detection.**

Keywords—Natural Language Processing, Sentiment Analysis, Lexicon Based Approach, Polarity Detection, Bengali Text

I. INTRODUCTION

Internet users around the world are increasing very quickly. In their day to day life, they express their reviews/opinions through the internet. Manual analysis of textual reviews/opinions is required too much time. For a huge amount of textual data analysis, it is quite impossible. For proper knowledge extraction from this huge amount of textual data, we need to build a system that can be done through Sentiment Analysis (SA), which plays a great role as a Natural Language Processing (NLP) tool to detect the polarity of the textual data.

Sentiment Analysis (SA) or Opinion Mining is a way of detecting the polarity that is expressed in written text or through textual data [1]. In another way, Sentiment analysis indicates to analyze the public's real opinions, emotions, sentiments, attitudes, and evaluations about various issues and services [2]. For commercial organizations, consumer's feedback about particular products, and services are very much important. Business organizations can improve their product quality [3] and services depending on customer opinions/reviews about their products and services. In Twitter, Facebook, Blogs or any internet sites where a large number of users use it to give their opinions/reviews to a specific topic. According to opinions/reviews, sentiments can be estimated through analysis.

Two popular approaches that help very much in sentiment analysis. The first process is lexicon-based weighted words of textual data and the second process is based on machine learning approaches. Lexicon based approach uses a dictionary of sentiment words with a sentiment score and for finding the polarity of a text line, token words are matched with this dictionary.

Bengali is a very popular language around the world. Every day more than 261 million people of the world are using Bengali and it is the primary language in Bangladesh [4] where it has huge importance and influence for businesses as well as Government offices. To avoid unwanted circumstances, it is important for the Government to understand the public sentiments as well as be updated about political agendas.

In this paper, we have developed a polarity detection system on Bengali textual opinions/reviews like as products, services, books, movies, songs, sports, politics, organizations using Python and lexicon dictionary-based approach where we have collected 5200 opinions/reviews by crawling and manually from some internet sites and social media. Data is preprocessed for getting clean and in a suitable format. Tokenization, punctuation removing, stop words removing & stemming is used during preprocessing. With the help of the Bengali sentiment words dictionary, boost word check, negation check & normalization of the score, we finally achieved sentiment results. We also have evaluated the system efficiency and performance by comparing with well-known machine learning classifiers. The evaluation of our developed system has provided with the acceptability of 92% accuracy as well as better precision, recall & F1-score.

The remaining of this paper is organized in the following way: related works in Sentiment Analysis in **Section II**. Our methodology for Bengali text Sentiment Analysis is given in **Section III**. In **Section IV,** presents Results and discussion. Finally, **Section V** is for the conclusion of this paper.

II. RELATED WORK

Due to a lack of tools in Sentiment Analysis (SA) on Bengali Text, this field has become a growing issue for research. Most of the sentiment tools developed for English. Few researchers work to develop Sentiment Analysis (SA) tools for Bengali Text polarity detection.

Document-level Sentiment Analysis has been performed in [5]. For polarity detection in Bengali language texts using WordNet [6] and SentiWordNet [7], the authors proposed valency analysis. In [8] authors translate 16000 Amazon watch reviews from English into Bangla using Google Translator for polarity detection on product reviews for both Bengali and English. Polarities are divided into weak, steady

978-1-7281-5842-6/19/$31.00 ©2019 IEEE

and strong where they achieve 85% accuracy for Bengali. Paper [9] proposes a human sentiment analysis model using the lexicon-based approach on any given data set. A lexicon-based approach has been proposed for sentiment analysis of news articles [10]. For sentiment analysis, a dataset from BBC, comprising of news articles between the years 2004 and 2005 has been used. They used the WordNet dictionary for positive, negative or neutral classification. The paper [11] demonstrates a sentiment analysis model trained using TF-IDF and they also used lexicon-based features to analyze the sentiments expressed by students in their textual feedback. A semi-supervised bootstrapping approach for polarity detection is used in [12]. They separate text line polarity as positive and negative. Sentiment analysis was done at the sentence level [13]. A dynamic dictionary with predefined positive and negative words was used to find the sentiment polarity of the sentence. 91% accurate results were achieved for the classification of news articles. In [14] authors proposed for the unigram presence method with negation handling and stemming. This proposed system obtained average accuracy but showed low performance. A polarity detection system is shown in [15] on Bengali movie reviews. They mainly used two machine learning approaches Support Vector Machines and Naive Bayes. Using a lexicon-based approach in [16] they show how cricket fan's sentiments vary over the period of time and they collected the tweets over Cricket tournament for sentiment analysis. Sentiment analysis has been performed in [17] for product review using the lexicon method. They have used product reviews from Twitter using twitter API.

III. Methodology

The steps of our proposed Bengali text polarity detection system are shown in Fig. 1.

We have developed a polarity detection system that can detect the polarity of the Bengali Text. A lexicon Bengali words dictionary is used in our system. Our dictionary contains sentiment words with a sentiment score. We have performed tokenization, punctuation remove, stop words remove and stemming. We calculate each word polarity of the text with the help of our lexicon dictionary and also checked boost word & negation. After that, the total sentiment score of the text is calculated and finally, we got sentiment results through normalization of the total sentiment score. The whole process of our developed system is described in the following way:

A. Data Collection

We have collected opinions/reviews of Bengali Text by web crawling and manually from some sites and social media such as Facebook groups, Twitter, Blogs, online newspaper sites. This data set contains opinions/reviews about products, services, books, songs, organizations, sports, movies and politics. We have collected 5200 sentences where 2600 sentences are positive and 2600 sentences are negative reviews/opinions.

B. Preprocessing of Text

Collected text data from several sites and social media may be contained some unwanted or noisy information. For getting clean and suitable data, we have to process the collected web data. It is a very important step in sentiment analysis for obtaining accurate sentiment results.

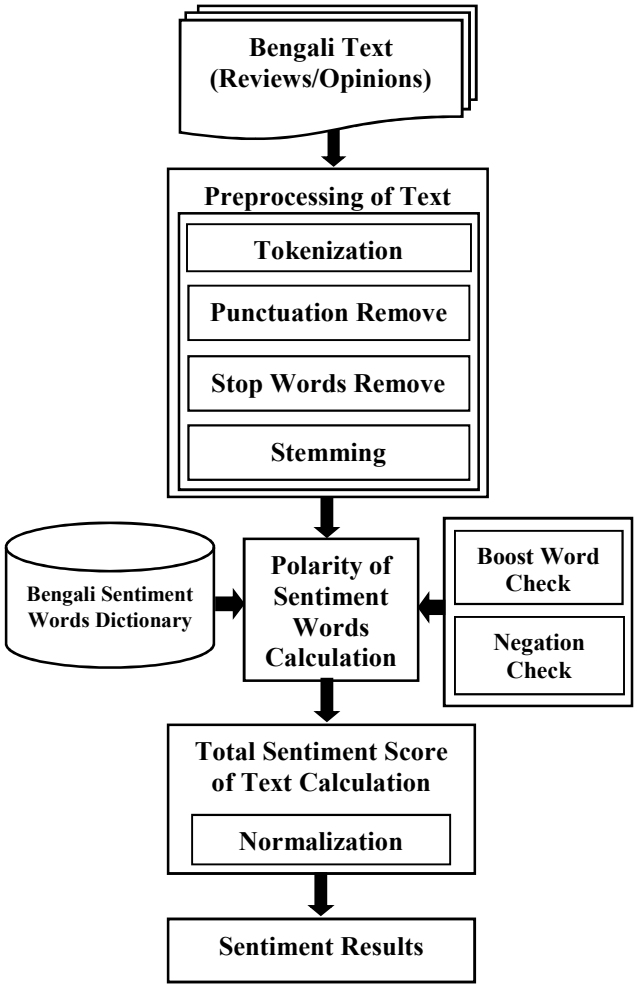

Fig. 1. Step by step process of proposed model.

Our steps of preprocessing are in the following way:

1) Tokenization: Tokenization is the process of splitting a text line into individual components such as words, phrases or symbols. Each component is known as a token. We can see the following example:

A book's review: একাত্তরের দিনগুলি বইটি খুব ভালো এবং অনেক পছন্দের !!! ভালো লাগলে আপনারা অনলাইনে কিনতে পারেন।

After tokenization, this review text appeared as

'একাত্তরের', 'দিনগুলি', 'বইটি', 'খুব', 'ভালো', 'এবং', 'অনেক', 'পছন্দের', '!!!', 'ভালো', 'লাগলে', 'আপনিও', 'অনলাইনে', 'কিনতে', 'পারেন', '।'

2) Punctuation Remove: In Bengali text, it contains several punctuations and it has very little influence in sentiment analysis. To get clear text, we can simply remove those punctuations. After removing punctuations, review text appeared as:

'একাত্তরের', 'দিনগুলি', 'বইটি', 'খুব', 'ভালো', 'এবং', 'অনেক', 'পছন্দের', 'ভালো', 'লাগলে', 'আপনিও', 'অনলাইনে', 'কিনতে', 'পারেন'

3) Stop words Remove: Stop words are used to complete a sentence but it has no importance in sentiment analysis. Normally in sentiment analysis, stop words are removed [18].

In Bengali Text, there are many stop words such as 'আমার', 'আমি', 'আপনি', 'আপনিও', 'হয়', 'হবে', 'পারি', 'পারেন', 'করে', 'করেন', 'থাকে', 'থাকেন', 'এবং' etc. We have created a Bengali stop words list. During the stop words remove process, we have checked our token with the stop words list. If a match is found then we removed it from the token list of review/opinions. After removing stop words, the above token list appeared as
'একাত্তরের', 'দিনগুলি', 'বইটি', 'খুব', 'ভালো', 'অনেক', 'পছন্দের', 'ভালো', 'লাগলে', 'অনলাইনে', 'কিনতে'

4) Stemming: Stemming is used in natural language processing to change a word into its original form or stem. Then each word can easily be recognized from the lexicon dictionary. We have checked each word for stemming. If a word contains such as ('া', 'ি', 'ে', 'ী', 'ই', 'য়', 'টি', 'টা') at the end, then the word is used for stemming. We checked stem word with lexicon dictionary where our lexicon dictionary contains only sentiment words with a sentiment score. If a match occurs, then we change the word from the token list with the stemming word. Otherwise, the word remains unchanged. After stemming, the above token list appeared as
'একাত্তরের', 'দিনগুলি', 'বইটি', 'খুব', 'ভাল', 'অনেক', 'পছন্দ', 'ভাল', 'লাগলে', 'অনলাইনে', 'কিনতে' and we can say it our final token list. Here, ['ভালো', 'পছন্দের'] they are sentiment words. Then, they are changed with stemming words ['ভাল', 'পছন্দ']. But ['একাত্তরের', 'দিনগুলি', 'বইটি', 'লাগলে', 'অনলাইনে', 'কিনতে'] they are not sentiment words and that's why they remain unchanged.

C. Polarity of Sentiment Words Calculation

For calculating polarity of sentiment words, we have to be familiar with the following terms:

1) Bengali Sentiment Words Dictionary: We have developed a lexicon dictionary[1] of Bangla sentiment words. Our lexicon dictionary contains more than 5100 words. We assigned a sentiment score to each word. This sentiment score is given in the range of -4 to +4 where +4 is for strong positivity and -4 is for strong negativity. Sentiment score with zero (0) indicates neutral.

2) Boost Word Check: In our proposed system, we specified some words which are used in Bengali text to boost its next words in a text line. According to the presence of boost word in a text, polarity of text will be increased or decreased.

We have created a list of boost words such as 'অল্প', 'অধিক', 'বেশি', 'সবচেয়ে', 'অনেক', 'খুব' etc.
For example, 'আইফোন অনেক ভালো মোবাইল ।'. Here, 'অনেক' is a boosted word according to our system. It boosts the next word 'ভালো' and 'ভালো' is a positive sentiment word. Then 'অনেক' increases the positivity of the text line.

3) Negation Check: Naturally, in Bengali text, negation words are placed at the end of the text line. We have developed a list of Bengali negation words such as 'নয়', 'নেই', 'না', 'নাই', 'নি'. If any negation word is found in a text line, the score of the sentiment word is multiplied by -1 which means that negative sentiment will be positive or positive sentiment will be negative. For example, 'শাওমি মোবাইল খারাপ না'. Though 'খারাপ' is a negative sentiment word, this sentence is predicted as positive by our system due to the presence of negation word 'না'. This sentence is an actual positive sentence. If negation & boost word both are found in the text line, then the score of sentiment word is increased or decreased according to the boosted word.

After the preprocessing of text, it is a very important task to calculate the polarity of each sentiment word from the final token list. We compared each final token word with our Bengali sentiment words dictionary. For each token word, a separate sentiment score is assigned according to our lexicon dictionary. If any token did not contain any sentiment word and then its token sentiment score value is assigned to zero (0). The final token list is also checked for boost word & negation word. If a match has occurred, the sentiment of the token list will be increased or decreased according to the boost word & negation word. Finally, we got a separate sentiment score for each token of the final token list.

D. Total Sentiment Score of Text Calculation

Total sentiment score of a text is calculated combining all final token sentiment score of the text line. In our lexicon Bengali dictionary, sentiment word score ranges from -4 to 4. It needs to normalize the total sentiment score for a text line. Using normalization[2], we constructed our sentiment results within range -1 to +1 where -1 is for strong negativity and +1 is for strong positivity.

1) Normalization: If the total sentiment score of a text line is denoted as T-Score and normalization parameter is denoted as alpha where alpha = 15, Then normalize score equation can be given as follows:

$$Normalize\ Score = \frac{T-Score}{\sqrt{(T-Score)^2+alpha}} \quad (1)$$

Normalize the total sentiment score between -1 and 1 and alpha approximates the max expected value.

E. Sentiment Results

Our main purpose is to analyze reviews/opinions about such as products, services, sports, books, songs, organizations, and politics. We want to analyze which reviews/opinions are positive and which reviews/opinions are negative. After normalization, we can easily detect which are positive reviews/opinions and which are negative reviews/opinions. For a text line, if normalization score is greater than zero and less than or equal to 1 (**0 < Normalize score <= 1**), then the text line is predicted as **positive.** Else if normalization score is less than zero and greater than or equal to -1 (**-1 <= Normalize score < 0**), then the text line is predicted as **negative.** Otherwise, the text is neutral. Some predicted reviews/opinions by our proposed system are shown in Table I.

[1]https://github.com/Fighter-1/Programming-/tree/master/Dictionary
[2]https://www.nltk.org/_modules/nltk/sentiment/vader.html#normalize

TABLE I. SOME PREDICTED REVIEWS/OPINIONS BY OUR PROPOSED SYSTEM

Reviews/Opinions	Normalized Score	Predicted Result
একাত্তরের দিনগুলি বইটি খুব ভালো এবং অনেক পছন্দের !!! ভালো লাগলে আপনিও অনলাইনে কিনতে পারেন ।	0.92	Positive
শাওমি মোবাইল খুব খারাপ ...!!	-0.73	Negative
শাওমী মোবাইলে ক্ষতিকর বিকিরণ একটু বেশিই হয় ।	-0.53	Negative
বাংলাদেশ খুব ভালো ক্রিকেট খেলেছে	0.60	Positive
জলের গান ব্যান্ড অনেক ভালো গান করে :)	0.59	Positive

IV. RESULTS AND DISCUSSION

The following performance measurement techniques are used to evaluate our proposed polarity detection system: **Accuracy, Precision, Recall, and F1-score.** Our data set contains 2600 positive reviews/opinions and 2600 negative reviews/opinions. We also have used popular machine learning classifiers such as Decision Tree (DT), Naive Bayes (NB) and Support Vector Machine (SVM) classifiers for our data set. We used these machine learning classifiers using Scikit-learn[3]. Scikit-learn is an open-source and very popular machine learning Python library. We compared our proposed system with machine learning classifiers in terms of Accuracy, Precision, Recall, F1-score and achieved better acceptable Accuracy, Precision, Recall & F1-score than machine learning classifiers.

Accuracy indicates the percentage of text lines or sentences in the test set that the classifier correctly labels. True Positive, True Negative, False Positive and False Negative is denoted as TP, TN, FP, and FN.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100\% \quad (2)$$

Precision indicates the number of text lines or sentences in the test set that is correctly labeled by the classifier from the total text lines or sentences in the test set that is classified by the classifier for a particular class.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

Recall indicates the number of text lines or sentences in the test set that is correctly labeled by the classifier from the total lines or sentences in the test set that are actually labeled for a particular class.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

F1-score is the weighted harmonic mean of Precision and Recall for a particular class.

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

[3]https://scikit-learn.org

A scatter plot is shown in Fig. 2. indicates the predicted positive and negative sentences of data set by our proposed system. The blue portion is for positive sentences and the red portion is for negative sentences. All positive sentences are remains in the range 0 to 1 (without including 0). All negative sentences are remains in the range -1 to 0 (without including 0). In the figure, we can see that a large amount of positive data are remains in range (0.30 to 1) and a large amount of negative data are remains in range (-1 to -.33).

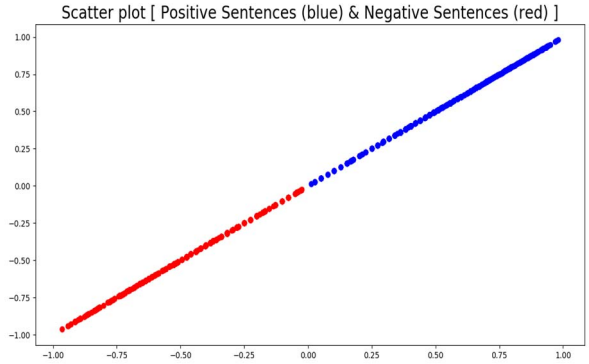

Fig. 2. Scatter plot for positive and negative predicted sentences by our proposed system.

TABLE II. ACCURACY COMPARISON OF OUR PROPOSED SYSTEM AND MACHINE LEARNING CLASSIFIERS.

Method	Accuracy (%)
Proposed System	92%
Decision Tree Classifier	87%
Naive Bayes Classifier	86%
Support Vector Machine Classifier	88%

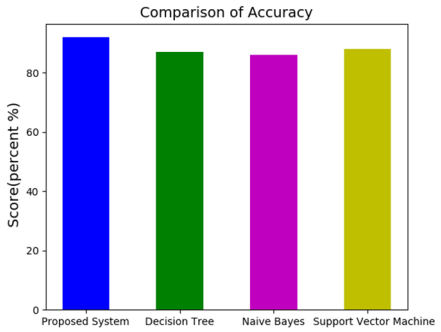

Fig. 3. Visualization of accuracy comparison in bar diagram.

TABLE III. PERFORMANCE COMPARISON OF OUR PROPOSED SYSTEM AND MACHINE LEARNING CLASSIFIERS.

Method	Precision	Recall	F1-score
Proposed System	0.91	0.94	0.92
Decision Tree Classifier	0.83	0.94	0.88
Naive Bayes Classifier	0.82	0.93	0.87
Support Vector Machine Classifier	0.84	0.94	0.89

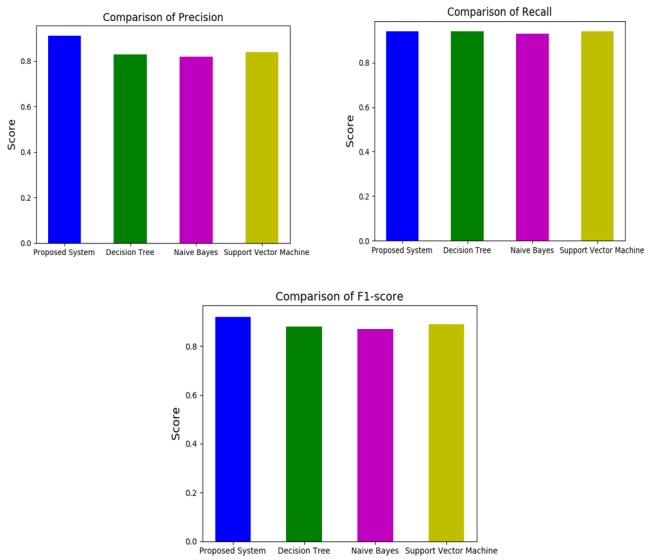

Fig. 4. Visualization of performance comparison in bar diagram.

We observed that our proposed model for polarity detection in Bengali text is better than supervised machine learning classifiers and accuracy, precision, recall & f1-score of our proposed technique is much superior. Our Bengali text polarity detection system is mostly depending on the Bengali lexicon sentiment word dictionary. A lack of sentiment words in the dictionary may change the results. Increasing proper sentiment words in the lexicon dictionary, very much accurate results can be achieved by our polarity detection system.

V. Conclusion

Our main purpose of this work is to develop a Bengali text polarity detection system for sentiment analysis. We have developed a polarity detection system using a lexicon dictionary-based approach. Our system is mostly depending on the Bengali sentiment words dictionary and we developed this dictionary much properly where it contains more than 5100 sentiment words. Due to the lack of standard data set in Bangla, we have collected Bengali reviews/opinions from sites and social media. Data is preprocessed for getting clean and in a suitable format. Tokenization, punctuation removing, stop words removing & stemming is used during preprocessing. With the help of the Bengali sentiment words dictionary, boost word check, negation check & normalization of the score, we finally achieved sentiment results. We also evaluated our system efficiency and performances and also compared with machine learning classifiers. We observed that our Bengali text polarity detection system is much better with 92% accuracy as well as better precision, recall & F1-score. In the future, we will enrich our sentiment Bengali words dictionary with more sentiment words for achieving much better accuracy.

References

[1] J. Reis, P. Olmo, F. Benevenuto, H. Kwak, R. Prates, and J. An, "Breaking the news: first impressions matter on online news," in International AAAI Conference on Web and Social Media(ICWSM), 2015.

[2] O. Kolchyna, T.T. P. Souza and P. C. Treleaven and T. Aste, "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination," arXiv:1507.00955v3 [cs.CL], Sep 2015.

[3] B. Liu, "Sentiment analysis and opinion mining," Synthesis lectures on human language technologies, vol. 5, 2012, pp. 1–167.

[4] N. Tabassum and M. I. Khan, "Design an Empirical Framework for Sentiment Analysis from Bangla Text using Machine Learning," in Proc. IEEE International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019.

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

[6] T. S. Utomo, R. Sarno and Suhariyanto, "Emotion Label from ANEW Dataset for Searching Best Definition from Wordnet," IEEE International Seminar on Application for Technology of Information and Communication (ISEMANTIC), 2018, pp. 249-252.

[7] A. Agarwal, V. Sharma, G. Sikka and R. Dhir, "Opinion mining of news headlines using SentiWordNet," in Proc. IEEE Symposium on Colossal Data Analysis and Networking (CDAN), 2016.

[8] K. M. A. Hasan, M. S. Sabuj and Z. Afrin, "Opinion mining using Naive Bayes," in Proc. IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), 2015, pp. 511–514.

[9] V.Singh, G. Singh, P. Rastogi and D. Deswal, "Sentiment Analysis Using Lexicon Based Approach," IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC), 2018, pp. 13-18.

[10] S. Taj, B. B. Shaikh and A. F. Meghji, "Sentiment Analysis of News Articles: A Lexicon based Approach," International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2019.

[11] Z. Nasim, Q. Rajput and S. Haider, "Sentiment Analysis of student feedback using machine learning and lexicon based approaches," IEEE International Conference on Research and Innovation in Information Systems (ICRIIS), 2017.

[12] S. Chowdhury and W. Chowdhury, "Performing sentiment analysis in Bangla microblog posts," International Conference on Informatics, Electronics & Vision (ICIEV), 2014.

[13] M. U. Islam, F. B. Ashraf, A. I. Abir and M. A. Mottalib, "Polarity detection of online news articles based on sentence structure and dynamic dictionary," in Proc. IEEE 20th International Conference on Computer and Information Technology (ICCIT), Dhaka, 2017.

[14] A. Kaur and V. Gupta, "Proposed algorithm of sentiment analysis for Punjabi text," Journal of Emerging Technologies in Web Intelligence, vol. 6, no. 2, 2014, pp. 180-183.

[15] N. Banik and H. H. Rahman, "Evaluation of Naïve Bayes and Support Vector Machines on Bangla textual movie reviews" IEEE International Conference on Bangla Speech and Language Processing(ICBSLP), 2018.

[16] A. Agarwal and D. Toshniwal, "Application of Lexicon Based Approach in Sentiment Analysis for short Tweets," in Proc. IEEE International Conference on Advances in Computing and Communication Engineering (ICACCE), 2018, pp. 189-193.

[17] P. Ray and A. Chakrabarti, "Twitter sentiment analysis for product review using lexicon method," International Conference on Data Management, Analytics and Innovation (ICDMAI), 2017, pp. 211-216.

[18] J. Leskovec, A. Rajaraman, and J. D. Ullman, "Mining of massive datasets," Cambridge University Press, 2014.