ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/332582518

Analyzing Performance of Different Machine Learning Approaches With Doc2vec for Classifying Sentiment of Bengali Natural Language

Conference Paper · February 2019

DOI: 10.1109/ECACE.2019.8679272

CITATION
1

READS
59

5 authors, including:

Md. Tazimul Hoque

2 PUBLICATIONS **7** CITATIONS

SEE PROFILE

Ashraful Islam

University of Louisiana at Lafayette

21 PUBLICATIONS **8** CITATIONS

SEE PROFILE

Eshtiak Ahmed

Ahsanullah University of Science & Tech

20 PUBLICATIONS **21** CITATIONS

SEE PROFILE

Khondaker A. Mamun

University of Toronto

69 PUBLICATIONS **293** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Speech Signal Processing View project

Bengali to English Machine Translation (Undergraduate Thesis Project) View project

All content following this page was uploaded by Ashraful Islam on 09 August 2020.

The user has requested enhancement of the downloaded file.

Analyzing Performance of Different Machine Learning Approaches With Doc2vec for Classifying Sentiment of Bengali Natural Language

Md. Tazimul Hoque*, Ashraful Islam†‡, Eshtiak Ahmed†‡, Khondaker A. Mamun* and Mohammad Nurul Huda*
*Department of Computer Science and Engineering
United International University, Dhaka-1212, Bangladesh
Email: tazim.ndc@gmail.com, {mamun, mnh}@cse.uiu.ac.bd
†Department of Computer Science and Engineering
Daffodil International University, Dhaka-1207, Bangladesh
Email: {ashraful, eshtiak}.cse@diu.edu.bd
‡Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh

***Abstract*—Vector or numeric representation of text documents has been a revolution in natural language processing as it represents similar parts of text in such a way that they are very close to each other, making it very easy to classify or find similarities among them. These vectors also represent the way we use the words or parts of documents as well which helps finding similarity even between pair of words. While *word2vec* is such a technique that represents each word as a vector, *doc2vec* takes it to another level by representing a whole sentence or document as a vector. Being able to represent an entire document as a vector allows comparing a substantial number of words or sentences at a time which can save computational power as well as bandwidth. This relatively newer *doc2vec* technology has not yet been implemented for Bengali sentiment analysis and its feasibility is also unknown. In this study, we have trained a *doc2vec* model using a corpus constructed with 7,000 Bengali sentences. The model consists of two types of data differentiated by their polarity i.e. *positive* and *negative*. Later, we have employed several machine learning algorithms for comparing the accuracy of classification among which Bi-Directional Long Short-Term Memory (BLSTM) has obtained the highest accuracy of 77.85% along with precision, recall and F-1 score of 78.06%, 77.39% and 77.72% respectively.**

***Keywords*—Sentiment Analysis (SA), Machine Learning (ML), Natural Language Processing (NLP), Bi-directional Long Short-Term Memory (BLSTM), Sequential Model (SM), doc2vec.**

I. Introduction

In recent years, various social media platforms e.g. Facebook, Twitter, Youtube, Google+ play a vital role in day to day life due to their ease-of-access, portability, and affordability [1], [2]. According to Statista, around 2.46 billion people are using social media worldwide as of 2017 and it is expected to reach 3.02 billion in 2021 where Facebook has remained the most popular one as of April, 2018 [3]. Another survey conducted in September 2018 by StatCounter says that 89.04% of social media users interact using Facebook in Bangladesh [4]. A very large number of data has been comprised over the Internet as a result of enormous dealing with social media platforms which conveys a significant contribution in Sentiment analysis (SA) [1]. To be specific, analyzing the reactions by users accumulated from social media contents and posts lead to categorize them into several labels i.e. sad, angry, love.

SA is also known as opinion mining, mood extraction or emotion analysis which is an application of Natural Language processing (NLP). The year 2001 or around can be marked as the beginning of the research awareness in the field of SA and opinion mining [5]. Research papers mentioning sentiment analysis focus specifically on the application of text classification according to their polarity positive (good), negative (bad) or neutral. But now-a-days SA expresses broadly to mean the computational treatment of opinion or review in text, processing natural language, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information [6]. In addition, recent advents in machine learning research, particularly deep learning based methods e.g. recurrent neural network (RNN), avail the opportunities to infer decisions by training a model in SA. Moreover, the latest key technique titled as *doc2vec* developed by Google Inc. [7] in which usually a document is represented by a vector, can be an emerging tactics for classifying emotions or opinions from social media reactions and posts. Although a lot of research has been conducted in the area of SA and they are mainly based on the social media posts written in English, still these areas are yet to be explored for the social media posts in Bengali language.

This paper aims to analyze public sentiments composed in Bengali on any topic and then categorize them into two particular classes i.e. *positive sentiment*, *negative sentiment*. For this we are considering Facebook post reactions *Love, Wow, Sad, Angry,* and *Haha* which represent different states of emotion. Here, *Love* and *Wow* reactions are considered as *positive sentiment* whilst *Sad* and *Angry* reactions are considered as *negative sentiment*. Facebook added these new reactions feature allowing users to react along with *Like* in a post. We have employed nine machine learning methods i.e. Logistic Regression (LR), Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), Decision Tree (DT), K-Neighbors, Linear Discriminant Analysis (LDA), Gaussian

978-1-5386-9111-3/19/$31.00 ©2019 IEEE

Naive Bayes (GaussianNB), Sequential Model (SM), and Bidirectional Long Short-term Memory (BLSTM) to build classification models so that they can classify the sentiments from users' reactions in different posts published in Bengali. Among this methods, BLSTM has performed best as it provides an accuracy of 77.85% along with precision, recall and F-1 score of 78.06%, 77.39% and 77.72% respectively. Therefore, this results are very promising for further investigations in this ground.

The layout of this paper is organized as follows- a brief description of related works are narrated in section II, followed by the essential indications for building the corpus and model experimented in this study in section III. Thereafter, conducted methods are being plotted sequentially in section IV. Then section V explains the results established in this study. Finally, section VI concludes the findings of this study with possible future implications.

II. Related Works

Many research works are accomplished by measuring the overall polarity of a document or sentence to determine if it is a positive or negative review [8]–[10]. Turney et al. used simple unsupervised learning algorithm which finds average semantic orientation of the phrases form the review containing adjectives or adverbs [8]. In system [9] Dave et al. trained a classifier using a self-tagged corpus of reviews form web sites. Pang et al. applied machine-learning method for text categorization to just the subjective portions of the document [10]. Phrase-level sentiment analysis is discussed in [11] which identifies the contextual polarity for a large subset of sentiment expressions. In their work they explained that contextual polarity of a phrase may be different from the polarities of the words appear in that phrase. Some popular approaches of sentiment analysis subjective lexicon, using N-Gram modeling, machine learning are discussed in [12]. Using deep learning model, Ouyang et al. proposed a framework *word2vec* + Convolutional Neural Network (CNN) [13] for classifying sentiment of movie reviews into fives labels: negative, somewhat negative, neural, somewhat positive and positive. They achieved 45.4% accuracy.

Though a lot of works have been explored considering the research works for Bengali in this ground, very few experiments have been investigated in recent years. Chowdhury et al. worked on sentiment analysis in Bengali microblog posts using SVM and Maximum Entropy (MaxEnt) classification techniques [14]. They collected 1,300 tweets using Twitter API and split the dataset as 1,000 tweets for training and 300 tweets for testing. They identified the overall polarity of a sentence as either negative or positive. Their achieved accuracy is 93% for SVM using unigrams with emoticons as features. Das et al. developed a phrase level polarity classification system using SVM [15]. They constructed a Bengali News corpus containing 3,435 distinct word-forms. It can categorize opinion phrase as either positive or negative. Their evaluated result have a precision of 70.04% and a recall of 63.02%. Amin et al. used "*word2vec*" model for vector representation of Bengali words [16]. They achieved 75.5% of accuracy using "*word2vec*" word co-occurrence score with the words sentiment polarity score. They collected 16,000 Bengali single line and multiline comments from blog posts and tagged them as positive or negative comment by a survey. Hassan et al. used deep recurrent model Long Short Term Memory (LSTM), with two loss functions binary cross-entropy and categorical cross-entropy for Bengali sentiment analysis [17]. They used 10,000 Bengali and Romanized Bengali text samples which were divided into three categories - Positive, Negative and Ambiguous. They achieved 70% accuracy with Bengali dataset and using Bengali and Romanized Bengali dataset the accuracy score was 55%.

III. Corpus and Model Preparation

1) Corpus Collection: The aim of this study is to analyze public sentiment on any topic from Bengali text and then categorize it based on sentiment polarity. We have considered positive and negative sentiment polarity in this work. To construct a corpus for Bengali sentiment analysis, different sources have been considered, among which Facebook post data seems most promising for SA as they represent the most natural form of language. In Facebook posts, people react with different reactions i.e. "Like", "Love", "Wow", "Sad", "Angry", and "Haha", each of which represent different states of emotion. Our aim is to classify these emotions into either positive or negative class. Users react with "Like" more than other reactions as it is easy to perform although it does not represent a specific sentiment polarity that can be classified as positive or negative [18]. Correlation among "Like" and other reactions can be expressed as-

- Strongly positive correlation with "Love" and "Wow".
- Weakly positive correlation with "Sad" and "Angry".

Although "Like" reaction is the most common, we have considered this as low-effort data from users and ignored it while classifying the sentiment polarity of a post. Furthermore, we have observed that people reacted Wow reaction in any funny, sarcastic posts more than any other reactions. Therefore, we cant polarize post sentiment in either positive or negative category based on "Wow" reaction.

We have used **Facebook Graph API** [19] implemented by our own Python script to collect data regularly from some popular Bengali Facebook public pages. We have collected 6244 Facebook posts which were pre-processed afterwards to validate them as proper text data. The pre-processing stage includes the filtering of any kind of hyperlink, special characters, duplicate post and non Bengali phonetics. This filtering shrunk the volume of our data set to 4317 posts. We stored this data into database which contains following columns - page type, page post text, and reaction counts of - "like", "love", "wow", "sad", "angry" and "haha". Fig. 1 demonstrates the total flow of data collection and corpus preparation from Facebook posts.

To prepare positive and negative post documents from this database we had to categorize multiple reactions into either positive or negative. Here, "Love" and "Wow" reactions represent positive polarity and on the other hand "Sad" and "Angry" reactions represent negative polarity. We considered total count of "Love" and "Wow" reactions as summation of total positive reactions, and total count of "Sad" and "Angry" reactions as summation of total negative reactions. Comparing the total numbers of positive and negative reactions of a post, we categorized it accordingly. This process is summarized in

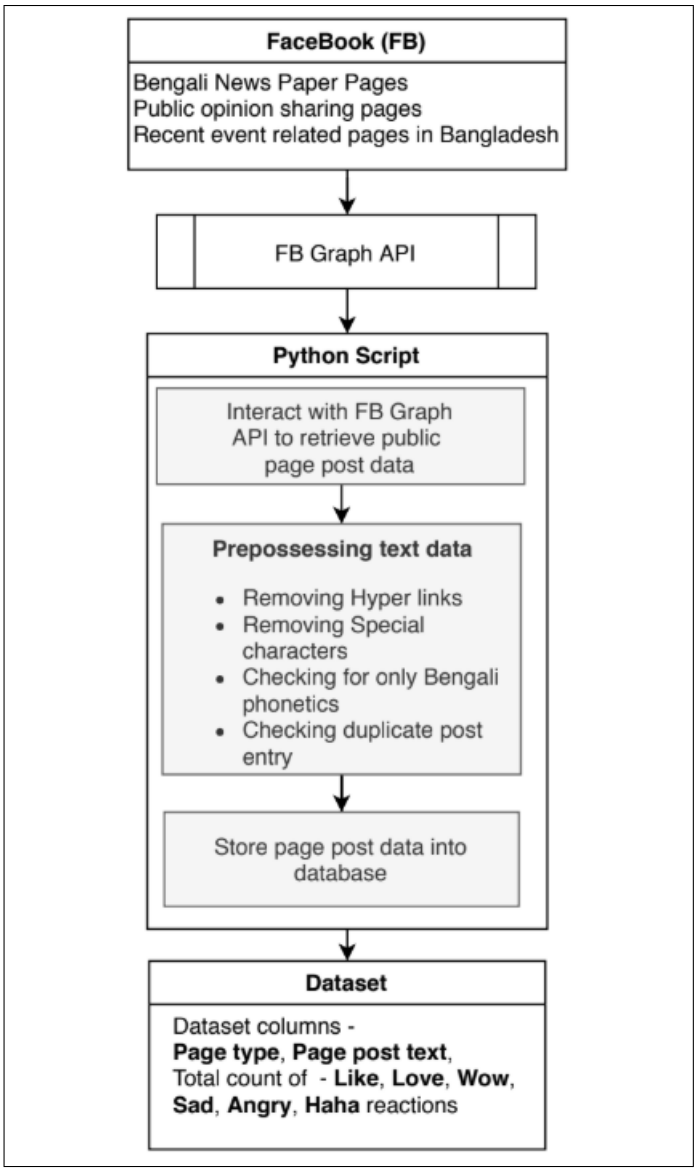

Fig. 1: Flow of data collection and corpus preparation from Facebook post.

Algorithm 1. Here, we have not categorize a post's sentiment if-

- the total number of "Wow" reaction is greater than positive or negative reactions.
- the total number of positive and negative reactions are same or both are zero.

The procedure to determine whether a post is categorized or not is shown in Algorithm 2. After this procedure we have 3,193 posts where majority reaction counts are -

- Love: 1,162
- Wow: 529
- Sad: 1,007
- Angry: 495.

Algorithm 1 Preparing Positive/Negative Documents from Facebook Page Posts

procedure PARSEPOSTS($posts$)
  for `each post` **do**
    $positive \leftarrow count(Love) + count(Wow)$
    $negative \leftarrow count(Sad) + count(Angry)$
    if $Categorizable() = false$ **then**
      skip to the next post
    else if $positive > negative$ **then**
      save post text into *positive.txt*
    else
      save post text into *negative.txt*
    end if
  end for
end procedure=0

So, finally we have 1,691 posts with positive polarity and 1,502 posts with negative polarity. To keep equal polarity data, we finally stored 1,500 posts per sentiment polarity (positive and negative).

Socian Ltd. [20] provided with a public corpus containing 4,000 labeled Bengali sentences according their sentiment polarity, either positive or negative which contains equal distribution of labeled data. They have collected this corpus from different social media platforms, news paper sites and blogs. We included this data set with our prepared corpus. This way finally we managed to prepare a corpus of 7,000 posts (3,500 for each sentiment polarity).

2) Model Preparation: Creating numerical representation of any document is the goal of *doc2vec* [21]. Here each document or sentence is represented as a vector where similar documents have closer values. We used our corpus, prepared using the process described in the previous subsection to train *doc2vec* model. All the labeled sentences from our corpus were fed into *doc2vec* model to build its vocabulary. Here each labeled sentence contains a list of Bengali words and a label either Positive or Negative based on its sentiment polarity. An example of labeled sentences used to train *doc2vec* is -

[['word1', 'word2', 'word3',..., 'last word'], ['label']]

To configure the *doc2vec* model we have considered window size 50 and vector size 120. Here window size represents maximum distance considered between current and predicted word in a sentence [21]. Output feature vectors dimensionality is represented by vector size . We trained the *doc2vec* model for 40 epochs and stored it for further use. Our final corpus and *doc2vec* model are uploaded in Kaggle [22].

IV. SENTIMENT CLASSIFICATION

For sentiment classification using our prepared *doc2vec* model, we used machine learning approaches i.e. LR, SVM, SGD, DT, K-Neighbors Classifier, LDA, GaussianNB, SM, and BLSTM. Our trained *doc2vec* model contains vector representation of 7,000 labeled sentences with 120 dimension of feature vectors. We split 80% data for training and 20% for testing randomly. All the classifiers are trained and tested using this data accordingly.

We observed different classifiers performance on *doc2vec* model. Among all classifiers, deep learning approach of

Algorithm 2 Checking a post is either categorizable or not

```
procedure CATEGORIZABLE()
    if count(Wow) > positive or negative then
        return false
    else if positive = negative then
        return false
    else if positive = negative then
        return false
    else if positive = negative = 0 then
        return false
    else
        return true
    end if
end procedure=0
```

BLSTM provided the best performance. Hence we are describing about BLSTM and the configurations we used to train this model. BLSTM an extension of LSTM, can improve performance of a model in sequence classification problems. We configured LSTM with 32 hidden nodes and the dropout value was set to 0.5 to reduce overfitting. LSTM hidden layers were wrapped using a Bidirectional layer which created two copies of hidden layers. This fits those two hidden layers for both the original and the reverse sequence of input. We used 10% data for validation from training data set in time of training the model. BLSTM model was trained using 200 epochs with a batch size of 200. This configuration provides an accuracy of 77.85% which is the best among all the classifiers employed.

V. Result Analysis

In this study, we have applied the most common machine learning performance metrics i.e. accuracy, precision, recall, F-1 score for the evaluation of engaged classifiers and the obtained results are represented in TABLE I. This results are sorted decreasingly based on the classification accuracy achieved by the employed classifiers. According to the data available in TABLE I, BLSTM has the best performance as it has gained an accuracy of 77.85% whilst GaussianNB has attained lowest accuracy which is 59.21% for the corpus we have built in this study. In addition, TABLE II illustrates the confusion matrix for BLSTM.

Fig. 2 and Fig. 3 convey the history of accuracy and loss respectively on the training and validation datasets over the epochs for model training. It is reported in the plot available in Fig. 2 that the model could achieve more accuracy as the rates of accuracy in both training and validation datasets are increasing significantly over model training epochs. On the other hand, from the plot of loss on both the training and the validation datasets represented in Fig. 3 indicates the sign for stopping model training in earlier epoch if depart process is found consistently in the parallel plots.

VI. Conclusion and Future Implications

While the *doc2vec* technology has been employed in numerous research based studies for sentiment analysis in the English language, its use in Bengali sentiment analysis has not been seen so far. However, our classification accuracy

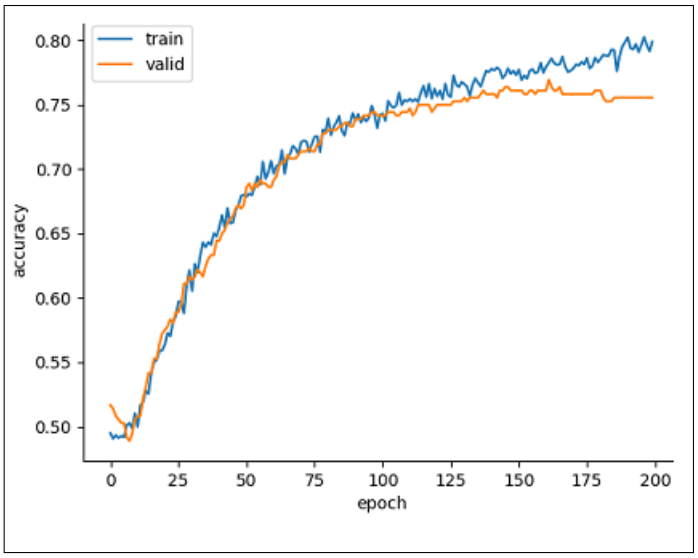

Fig. 2: A plot of accuracy on the training (train) and validation (valid) datasets over training epochs for BLSTM

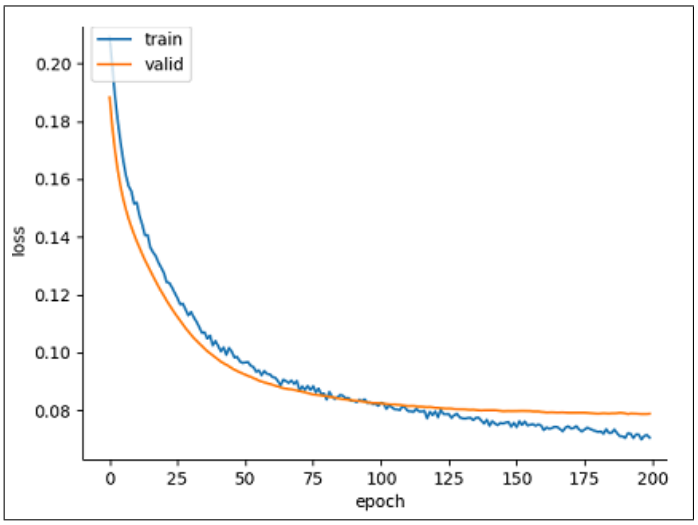

Fig. 3: A plot of loss on the training (train) and validation (valid) datasets over training epochs for BLSTM

for different classifiers shows that this technology has enough potential if implemented properly. The primary contribution of this study is that it presents the very first *doc2vec* model for Bengali sentiment analysis while the achieved classification accuracy is significantly better than that of other implementations using *word2vec*.

TABLE I. Obtained results for employed classifiers

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-1 Score (%)
BLSTM	77.85	78.06	77.39	77.72
SM	74.35	72.94	77.42	75.12
SGD	74	73.46	75.14	74.29
LR	73.14	71.89	76	73.88
LDA	72.42	72.49	72.28	72.38
SVM	67.21	67.04	67.71	67.37
K-Neighbors	63.28	60.59	76	67.42
DT	61.07	61.05	61.14	61.09
GaussianNB	59.21	57.32	72.14	63.88

TABLE II. CONFUSION MATRIX FOR BLSTM

		Predicted	
		Negative	**Positive**
Actual	**Negative**	549	152
	Positive	158	541

Although the model is currently constructed with the polarity of sentiment, it is a definite possibility that multi-class model can be prepared given enough time and larger volumes of data. In future, we can work with multiple class classification instead of the polarities being just positive and negative. Detecting non-polarized textual data can be another improvement of our current system. Additionally, we pre-processed dataset to get the text containing only Bengali phonetics which filtered out Romanized Bengali texts. This narrowed down our dataset and also the scope to work with Latin letters used to write Bengali sentences (Romanized Bengali text). Filtering special characters removed any kind of emoticons used in the textual post, but emoticon plays a vital role in sentiment expression. We are intending to work with Romanized text and emoticons in our next research work involving sentiment analysis. To summarize, this study represents the great potential of Bengali *doc2vec* technique and opens the door to more significant contributions in this aspect.

REFERENCES

[1] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[2] R. Gaspar, C. Pedro, P. Panagiotopoulos, and B. Seibt, "Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events," *Computers in Human Behavior*, vol. 56, pp. 179–191, 2016.

[3] "Social Media Statistics Facts," https://www.statista.com/topics/1164/social-networks/, (Visited on 10/27/2018).

[4] "Social Media Stats Bangladesh," http://gs.statcounter.com/social-media-stats/all/bangladesh, (Visited on 10/27/2018).

[5] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, 2008.

[6] "Sentiment analysis," https://en.wikipedia.org/wiki/Sentiment_analysis, (Visited on 10/27/2018).

[7] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, 2014, pp. 1188–1196.

[8] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pp. 417–424, 2002.

[9] K. Dave, S. Lawrence, and 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 (WWW '03)*, 2003, pp. 519–528.

[10] B. Pang, L. Lee, Z. A. Bán, B. Pang, L. Lee, and S. Vaithyanathan, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, vol. 48, no. 1, pp. 49–55, 2002.

[11] T. Wilson, J. Wiebe, and P. Hoffman, "Recognizing contextual polarity in phrase level sentiment analysis," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 2005, pp. 347–354.

[12] A. Kaur and V. Gupta, "A Survey on Sentiment Analysis and Opinion Mining Techniques," *Journal of Emerging Technologies in Web Intelligence*, vol. 5, no. 4, pp. 367–371, 2013.

[13] X. Ouyang, P. Zhou, C. H. Li, and L. Liu, "Sentiment analysis using convolutional neural network," in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, 2015, pp. 2359–2364.

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

[15] A. Das and S. Bandyopadhyay, "Opinion-polarity identification in bengali," in *International Conference on Computer Processing of Oriental Languages*, 2010, pp. 169–182.

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

[17] A. Hassan, M. R. Amin, A. K. A. Azad, and N. Mohammed, "Sentiment analysis on bangla and romanized bangla text using deep recurrent models," in *IWCI 2016 - 2016 International Workshop on Computational Intelligence*, 2017, pp. 51–56.

[18] "Facebook Reactions," http://minimaxir.com/2016/06/interactive-reactions/, (Visited on 10/27/2018).

[19] "Facebook Graph API," https://developers.facebook.com/docs/graph-api/, (Visited on 10/27/2018).

[20] "Socian Bangla Sentiment Dataset," https://github.com/socianltd/socian-bangla-sentiment-dataset-labeled/, (Visited on 10/27/2018).

[21] "Doc2vec paragraph embeddings," https://radimrehurek.com/gensim/models/doc2vec.html, (Visited on 10/27/2018).

[22] "Sentence corpus and Doc2Vec file for Bengali Sentiment Analysis," https://www.kaggle.com/tazimhoque/bengali-sentiment-text, (Visited on 10/29/2018).

View publication stats