2019 International Conference on Electrical, Computer and Communication Engineering (ECCE),
7-9 February, 2019

Design an Empirical Framework for Sentiment Analysis from Bangla Text using Machine Learning

Nusrath Tabassum
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
Chittagong-4349, Bangladesh
nusrathtabassum13@gmail.com

Muhammad Ibrahim Khan
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
Chittagong-4349, Bangladesh
muhammad_ikhancuet@yahoo.com

***Abstract* – Natural Language Processing (NLP) lends a helping hand for programming the computers to inspect a huge amount of data. Sentiment analysis is an application of NLP which deals with data to examine the sentiment or opinion that can be either positive or negative. Using Bangla text, sentiment analysis has become a challenge as there were only few works on it. As a decision maker, sentiment extrication not only capturing consumer attitudes but also helps in social behavior observance, politics and policy making. This paper quantifies total positivity and negativity against a document or sentence using Random Forest Classifier to classify sentiments. We contemplate the use of unigram, POS tagging, negation handling and classifier.**

Index Terms – Bangla language, Feature extraction, Sentiment inspection.

I. INTRODUCTION

Sentiment analysis examines the behavior or attitude of a person, product and provides review based on positivity or negativity. It uncovers the feelings of a person about a specific topic. It provides a straight decision of end users review on comments, products, business decisions etc. Now-a-days people express their feelings through comments in facebook, twitter and many other websites. In this paper, data is driven using facebook and twitter.

Bangla is the eighth most spoken language in the world. Across the world, there are about 261 million Bengali speaker. Bangla is the native language in Bangladesh and second language in India. So sentiment analysis in Bangla text has become a demanding issue in this modern era. There are lots of research work on sentiment analysis on English that is completely opposite from Bangla. Hardly few research works were done on Bangla because of resource lacking and Bangla language complexity.

In this paper, expression such as positive or negative determination is the pivotal task. For example-
"আশা মানুষকে জাগ্রত করে ।" This is a positive sentence. In this sentence, there are two positive words such as "আশা" and "জাগ্রত" .
"বৃষ্টিতে বাড়ল দুর্ভোগ।" This is a negative sentence. Here negative word is " দুর্ভোগ " .

In our research work, we count the number of such positive and negative words as a determinator for sentiment analysis. Because such kind of words extract a distinctive feature which determine whether the comment is positive or negative. The proposed system is tested against trained data. Main contribution of our research –

- Developing an empirical framework for sentiment analysis
- Reducing structural complexity as our proposed system doesn't work with Bangla sentence structure
- Developing Bangla feature words list
- Making a simple but efficient model
- Inspecting users attitude on particular issue
- Making accuracy level higher

This paper is assembled as follows. Section II discusses about correlated work. Section III illustrates system elucidation. In section IV, we discuss our proposed methodology. Section V represents evaluation and experimental results. In section VI, we compare our proposed system. The conclusion is drained in section VII.

II. RELATED WORKS

With the tremendous usage of website such as facebook, twitter etc sentiment analysis has become one of the most interesting field for for many researches.

(Nabi et al., 2016) proposed a method of Tf-Idf to recognize the sentiment or opinion that resulted in a better approach because the success rate was 83%. But there was some lackings in their proposed system because of ignored mixed sentences and some distorted noisy data [1].

(Hasan et al., 2013) erected phrase patterns to match with the predefined phrase pattern in Bangla text. The sentiment orientation for each sentence was cumulated for recognizing sentiment from Bangla text [2].

(Pak et al., 2010) suggested an approach of multinomial Naïve Bayes with n-gram and POS-tag as features. Their system detected positive, negative and neutral opinions at document level [3].

(Alam et al., 2015) designed an intelligent agent for generating basic seven facial expressions such as sad, surprise, joy, fear, disgust, anger, contempt. These expressions were classified in two labels ranging from -5 to +5. Their accuracy was 94%. [4].

SentiWordNet was used for valence setting to obtain the sense of each word in Bangla text. Total positivity, negativity

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

and neutrality was calculated with respect to the total sense [5].

Subjectivity classifier was used to mark opinionated words bearing sentences by using sentiment lexicon, POS tagging and theme clusters. SVM was used to identify phrase level polarity for Bangla [6].

(Kaur et al., 2014) proposed for unigram presence method as well as simple scoring method with negation handling and stemming. This proposed system gained finer accuracy but showed low performance [7].

Ekman's six basic emotions were assigned to the blogger's comments using the Bengali WordNetAffectLists. This research was done at word, sentence and paragraph level [8].

(Chowdhury et al., 2014) used semi-supervised bootstrapping approach for automatic sentiments extraction conveyed by users and identifying the polarity of positive or negative text [9].

(Balahur et al., 2012) carried out research on three different languages- French, German and Spanish to deal with the problem of sentiment detection. They had used three distinct Machine Translation (MT) systems- Bing, Google and Moses. SVM was used for hyperplane identification that segregated positive and negative examples in the training phase [10].

(Hassan et al., 2016) used Romanized Bangla and Bangla texts consisting of 9337 samples collected from Facebook, Twitter, YouTube, online news portals and product review pages. Deep recurrent model specially Long Short Term Memory (LSTM) was applied on text corpus [11].

III. System Elucidation

In this research, we have utilized supervised learning method. The system is interpreted step by step here.

A. Data Set Collection

Posts and comments in different websites such as facebook, twitter etc aren't only in English but also in other languages. Bangla data set collection is a compelling need in the present time as bangla is one of the most popular spoken languages ranked eighth in the world. To regulate this research we have collected different comments and opinions manually using facebook and twitter. 1050 Bangla texts indicating positive and negative sentiments are taken in this experiment.

B. Data Set Distribution

A collection of 1050 texts are split into training and testing data. we have trained 730 feature words from 850 texts manually for the purpose of training. The remaining 200 texts are used for testing. There are 100 positive texts and 100 negative texts that we have tested.

C. Used Classifier

For both classification and regression problems Random Forest algorithm can be used. Here we have used it as a classifier. A set of decision trees are created in the random forest classifier. The number of trees in the forest and the results it gets have a undeviating relationship between them. The votes from different decision trees are then clumped to opt the final result of the tested comments. Random forest creation and to make a prognostication using the created random forest are two stages in the random forest classifier.

IV. Procedures and Feature Analysis

Sentiment detection prognosticates the sentiment class that is either positive or negative. We indicate the following task by using machine learning classifier. Our proposed system is given below -

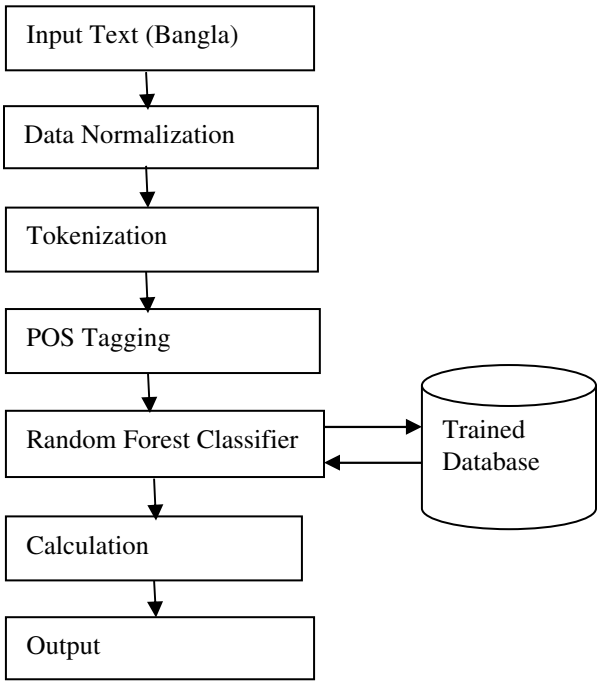

Figure 1. Proposed system

Following steps will give explanation of the process that we have carried out for our proposed system.

A. Preprocessing Steps

For the preprocessing, the following strides are needed-

1. Data Normalization : Data normalizer takes the selected comment that we want to test as input and putrefies the whole comment as a list of sentences. For each sentence remove punctuation mark at the end of the sentence ("।", "?") .

2. Data Tokenizer: After getting the normalized sentences, tokenizer is used. It is the program module that accepts the given normalized sentences to be parsed into unbroken, individual words called " tokens" .

3. POS Tagging : For the text analysis task, Part-of-Speech is one of the most significant process to differentiate tokens into their parts of speech such as noun, adjective, adverb etc. Every token was tagged with their parts-of-speech.

B. FEATURE SET CONSTRUCTOR

Feature extractor generates all unigrams that express not many information about the linguistic pattern of sentiments. Parts of speech features are also generated by first tokenizing the normalized sentences.

C. TRAINED DATABASE

For the purpose of training, we have trained feature words. A specific tag such as positive tags are assigned for positive words and negative tags are assigned for negative words respectively. Negation words are also tagged as negation.

Table 1. Sample training data

S_id	Word	Tag
1	ভাল	Positive
2	খুশি	Positive
3	খারাপ	Negative
4	অপরাধ	Negative
5	না	Negation
6	নাই	Negation

D. FEATURE CLASSIFICATION

Random forest classifier provides the desired result of the feature words that is either positive or negative. The working principle is given below-

1. Select p sentences from the total q sentences randomly where p << q.

2. Among the p sentences, node r is calculated.

3. Break the node into child nodes using the best split.

4. Only the child nodes that are noun, adjective, verb and adverb are extracted.

5. Matches the extracted child nodes with trained database and recognizes the featured words.

6. Assigns a specific tag to the featured words of the child nodes.

7. The negation words are" নাই, নই, নেই, নহে, নয়, না ". For the purpose of negation handling, When one negative word and one negation word occur at the same simple sentence then consider them as a positive tag. For example: "সে খারাপ না". Here "খারাপ" and "না" together is considered as positive tag.

8. The sum of positive words and the sum of negative + negation words are calculated separately.

E. SENTIMENT DETECTION

If the difference between the number of positive words and the number of negative + negation words is greater than zero then the comment is positive. If the difference is equal or less than zero then the comment is negative. Example of some tested comments-

Table 2. Some tested comments

Comments	Detected Sentiment
তিনি সৎ লোক	Positive
জিনিসটি ভাল চলছে	Positive
বৃষ্টিতে বাড়ল দুর্ভোগ	Negative
নষ্ট হয়ে গিয়েছে ভাই	Negative

V. EVALUATION AND EXPERIMENTAL RESULTS

The effectiveness of our system is evaluated using the following performance measurement techniques –

- Confusion Matrix
- Precision & Recall
- F1-score
- Accuracy & Error-rate

These techniques use four terms: true positive, true negative, false positive and false negative.

	True Value: Positive	True Value: Negative
Predicted value: Positive	81	11
Predicted value: Negative	19	89

Figure 2. Confusion Matrix using Unigram + POS

	True Value: Positive	True Value: Negative
Predicted value: Positive	85	11
Predicted value: Negative	15	89

Figure 3. Confusion Matrix using Unigram + POS + Negation Handling

Table 3. Precision and Recall

Feature	Precision		Recall	
	Positive	Negative	Positive	Negative
Unigram +POS	0.88	0.82	0.81	0.89
Unigram + POS+Negation	0.88	0.85	0.85	0.89

Table 4. F1-score, Accuracy and Error-rate

Feature	F1-score		Accuracy	Error-rate
	Positive	Negative		
Unigram+POS	0.84	0.85	85%	15%
Unigram + POS+Negation	0.86	0.86	87%	13%

VI. COMPARISON WITH EXISTING SYSTEMS

We have compared our proposed system with the existing systems regarding reference [1] and [9]. We have compared the dataset distribution and the performance.

Table 5. Comparison of Training & Testing dataset between existing and proposed system

	Training Dataset	Testing Dataset
[1]	1400 (sentence)	100 (sentence)
[9]	1000 (tweet)	300 (tweet)
Proposed System	850 (text)	200 (text)

Table 6. Comparison of Precision and Recall between existing and proposed system

Method		Precision		Recall	
		Positive	Negative	Positive	Negative
Tf-Idf [1]		0.81	0.85	0.86	0.8
SVM [9]	Unigram	0.68	0.66	0.63	0.70
	Unigram+ Negation	0.68	0.66	0.64	0.70
Maximum Entropy [9]	Unigram	0.69	0.67	0.65	0.72
	Unigram+ Negation	0.69	0.67	0.65	0.72
Proposed System	Unigram +POS	0.88	0.82	0.81	0.89
	Unigram+ POS+ Negation	0.88	0.85	0.85	0.89

Table 7. Comparison of F1-score, Accuracy and Error rate between existing and proposed system

Method		F1-score		Accuracy	Error rate
		Positive	Negative		
Tf-Idf [1]		0.83	0.82	83%	17%
SVM [9]	Unigram	0.65	0.68	67%	33%
	Unigram+ Negation	0.66	0.68	67%	33%
Maximum Entropy [9]	Unigram	0.67	0.69	68%	32%
	Unigram+ Negation	0.67	0.69	68%	32%
Proposed System	Unigram +POS	0.84	0.85	85%	15%
	Unigram+ POS+ Negation	0.86	0.86	87%	13%

VII. CONCLUSION

In the modern era, people involved in corporate world are running out of time. So they always want the gist of the documents. Our research faces this challenge by fulfilling the demand of exhibiting the outcome that is either positive or negative from a comment or opinion. We propose combination of unigram, POS tagging, negation handling and random forest classifier to provide more accurate result. Emoticons has become a medium of expressing sentiment. As a future work, we plan to deal with emoticons and vast data set and also to demonstrate the neutral sentiment and compound sentences.

VIII. REFERENCES

[1] Nabi, M. M., Altaf, M. T., & Ismail, S. (2016). Detecting sentiment from Bangla text using machine learning technique and feature analysis. *International Journal of Computer Applications*, *153*(11).

[2] Azharul, K. M., Sajidul, H., Mashrur-E-Elahi, I., & Izhar, M. N. (2013). Sentiment Recognition from Bangla Text. *Technical Challenges and Design Issues in Bangla Language Processing, IGI Global*.

[3] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010, pp. 1320-1326).

[4] Alam, L., & Hoque, M. M. (2015, May). The design of expressive intelligent agent for human-computer interaction. In *Electrical Engineering and Information Communication Technology (ICEEICT), 2015 International Conference on* (pp. 1-6). IEEE.

[5] Hasan, K. A., & Rahman, M. (2014, December). Sentiment detection from Bangla text using contextual valency analysis. In *Computer and Information Technology (ICCIT), 2014 17th International Conference on* (pp. 292-295). IEEE.

[6] Das, A., & Bandyopadhyay, S. (2010). Phrase-level polarity identification for Bangla. *Int. J. Comput. Linguist. Appl.(IJCLA)*, *1*(1-2), 169-182.

[7] Kaur, A., & Gupta, V. (2014). Proposed algorithm of sentiment analysis for punjabi text. *Journal of Emerging Technologies in Web Intelligence*, *6*(2), 180-183.

[8] Das, D., Roy, S., & Bandyopadhyay, S. (2012, June). Emotion tracking on blogs-A case study for Bengali. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 447-456). Springer, Berlin, Heidelberg.

[9] Chowdhury, S., & Chowdhury, W. (2014, May). Performing sentiment analysis in Bangla microblog posts. In *2014 International Conference on Informatics, Electronics & Vision (ICIEV)* (pp. 1-6). IEEE.

[10] Balahur, A., & Turchi, M. (2012, July). Multilingual sentiment analysis using machine translation?. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 52-60). Association for Computational Linguistics.

[11] Hassan, A., Amin, M. R., Mohammed, N., & Azad, A. K. A. (2016). Sentiment Analysis on Bangla and Romanized Bangla Text (BRBT) using Deep Recurrent models. *arXiv preprint arXiv:1610.00369*.