Detecting Sentiment from Bangla Text using Machine Learning Technique and Feature Analysis

Muhammad Mahmudun Nabi
Department of Computer Science and Engineering, Sylhet Engineering College, Sylhet, Bangladesh.

Md. Tanzir Altaf
Department of Computer Science and Engineering, Sylhet Engineering College, Sylhet, Bangladesh.

Sabir Ismail
Assistant Professor, Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet, Bangladesh.

ABSTRACT
Sentiment Analysis is an ongoing field of research in text mining, especially in the area of Bangla language as there are few research works done in this particular sector. In general, sentiment classification means the analysis to determine the expression of a speaker whether he or she holds a positive or negative opinion to a specific subject on a given text. It is consider the information in a text at the document, and extract unique feature/aspect level whether the given query holds an expression of positive, negative or neutral. In this process of retrieving information of a document the author use Tf.Idf (term frequency–inverse document frequency) to come out a better solution and give more accurate result by extracting different feature of a positive, negative or neutral word of sentiment analysis in particular view of Bangla text. The author calculate the total positivity, negativity of sentence or document with respect to total sense. Sufficient example and experiment are presented to describe the feature extraction of sentiment that it's found in this methodology.

General Terms
Machine learning, Bangla text analysis

Keywords
Sentiment detection, Text mining, Bangla words, Supervised learning, Feature extraction.

1. INTRODUCTION
Sentiment analysis is a very large area of Natural Language Processing system. It is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. A common use case for this technology is to discover how people feel about a particular topic. It is an important application of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in source materials [1]. The advantages of sentiment analysis in the modern era is to come out a straight decision of users review on a product, comments, blogs and other articles in online journal, facebook, twitter, blogs and websites.

Bangla (one of the important Indo-Iranian languages) is the sixth-most popular in the world and spoken by a population that now exceeds 250 million. It is the primary language in Bangladesh and second language in India [2]-[4]. Lots of research on sentiment analysis has been done on different languages such as English (Picard et al., 2011; Sebastiani, 2002; Lu, Liu, & Zhang, 2006), Chinese (Zhanget al. 2011), Urdu (Sayed et al. 2011) etc. But in contrary, there are many few research work done on Bangla language over the the last few years due to its lack of recourses and complexity of Bangla language.

Detecting sentiment is a very difficult task in Bangla language. In this process the main task of sentiment analysis is to determine the positivity/negativity/neutral expression on a given Bangla text with respect to some document of training data. In Bangla the possible example of sentiment analysis is-

Positive sentence: আমি ভালো আছি। in this sentence the positive word is ভালো. Position of word 2nd and feature of this word is 'ভ' 'াা' 'াো' 'ল' (starting letter- ভ, ending letter- ল)

Negative sentence: সে খারাপ ছাত্র। in this sentence the positive word is খারাপ. Position of word 2nd and feature of word is- 'খ' 'াা' 'র' 'াা' 'প' (starting letter খ, ending letter- প)

Neutral Word: সে ভাল ছাত্র কিন্তু সে পরিশ্রমী না in this sentence the word কিন্তু neutralizes the sentence, so it do not get any exact information to differentiate the sentence whether it is positive or negative.

The fundamental task in sentiment analysis is classifying the sentimental state of a Bangla text in a given document whether the expressed opinion in a document is positive, negative, or neutral (Shaikh et al., 2007).Here it is suggested TF.IDF (Term frequency-inverse document frequency) searching model in the area of text mining process. It is a way to score the importance of words (or "terms") in a document based on how frequently they appear across multiple documents. In this research on Bangla language they use some document of training data to search the possibility of given query in a document and extract a unique feature to give an output of positive, negative or neutral review of the text.

2. RELATED WORKS
The method that it use and extract a feature of positive, negative, neutral words have not been suggested in the last few research of Bangla language processing, so far the authors know. Although there are some Researchers have expressed their interest in Bengali text and there are many publications based on sentiment analysis with data resources from various Bengali corpuses have been published. Das & Bandyopadhyay (2010a) presented two different approaches for identifying emotion holders from Bengali sentences. In this work, first approach, the baseline model, is developed based on the combinations of various part-of-speech (POS) features extracted from the phrase-based similarities and the

second approach, syntactic model, is based on the argument structure of the sentences with respect to the verbs [5].

Most research on opinion analysis has focused on sentiment analysis [6], Another research paper describes a method for ranking a large list of adjectives according to a subjectivity score without resorting to any knowledge-intensive external resources (such as lexical databases, parsers or manual annotation). The method only requires a list of adjectives to be ranked and a small set of "seeds" (manually selected subjective adjectives) [7], A novel trainable method that statistically combines two indicators of grad ability is presented and evaluated, complementing existing automatic techniques for assigning orientation labels [8], To determine this sentiment polarity, another research paper propose a novel machine-learning method that applies text-categorization techniques to just the subjective portions of the document [9], Another method for obtaining opinion-bearing words can be used effectively to identify opinion-bearing sentences [10]. Review mining [11], customer feedback [12] and strength of document orientation [13]. Methods on the extraction of opinionated sentences in a structured form can be found in [14]. Some machine learning text labeling algorithms like Conditional Random Field (CRF) ([15], Some research paper authors address the overfitting problem from a different perspective, by factoring the CRF distribution into a weighted product of individual "expert" CRF distributions. The call of this model a logarithmic opinion pool (LOP) of CRFs (LOP-CRFs) [16]. Support Vector Machine (SVM) [17] have been used to cluster same type of opinions. 'Reference [18]' proposed an automatic sentiment detection technique using machine learning formula for Bangla text. They also discover some important challenges for Bangla language processing. Recognizes the emotion of Bangla language using a pre-processing technique to retrieve and store the bloggers comments on specific topics [3].

A semi-supervised approach to sentiment classification is proposed in [19] to detect the ambiguous and unambiguous review via a combination of active learning, transudative learning, and ensemble learning.

3. PROCEDURES AND FEATURE ANALYSIS

Symbolic techniques are currently considered to be outdated, and machine learning techniques have proven to be more accurate and user-friendly. In this study, the authors will employ supervised learning methods. The feature sets and supervised classifier employed in the experiments are discussed in this process step by step.

3.1 Collecting Data Set

The data set used to conduct this paper is the manually collected corpus. It is opted to keep the data set quite small, consisting of around 1500 short Bangla comment from various social sites, since this research is only working with Bangla text, it is tried to keep the data set quite simple. Here it is ignored mixed sentences, that happens quite a lot with Bangla, because the main theme of this experiment to gain information from Bangla text. The dataset splited into training set and test set, comprising 1400 and 100 sentences respectively.

3.2 Bangla Word Experiments

Words are the elements that build up any conversation and express one's sentiment. Some words are key to find the main theme of any statement. On this segment of this study, it will

discuss about some features of words that helped this experiment to achieve its goal.

3.2.1 Most Frequent Words

Dividing this collected training dataset in two parts, it's separated positive comments and negative comments. In the first stage, the most frequent words of the corpus are gathered by using the frequency distribution of the words. The feature set consisting of the n-most frequent words in the corpus is a simple bag-of-words. Table 1 and following bar chart shows 20 most frequent words in this data set and the number of occurrence of that word in positive or negative sentences.

Table 1. Total occurrences of experiment in positive negative sentence

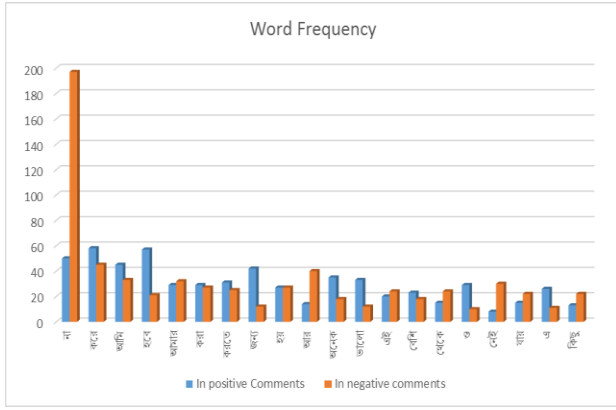| Word	Total occurrence	In positive comments	In negative comments
না	247	50	197
করে	103	58	45
আমি	78	45	33
হবে	78	57	21
আমার	61	29	32
করা	56	29	27
করতে	56	31	25
জন্য	54	42	12
হয়	54	27	27
আর	54	14	40
অনেক	53	35	18
ভালো	45	33	12
এই	44	20	24
বেশি	41	23	18
থেকে	39	15	24
ও	39	29	10
নেই	38	8	30
যায়	37	15	22
এ	37	26	11
কিছু	35	13	22

Figure 1: Top 20 frequently used words from training dataset

From the table and bar chart, it is clear that "না" word is used rapidly in the negative comments which is the Bangla equivalent of no. The other most frequent words are pronouns (i.e. 'আমি' and 'আমার'), verbs (i.e. 'করে', 'হবে', 'করা', 'করতে', 'হয়'), adjectives (i.e. 'অনেক', ' ভালো', ' বেশি'), conjunctions (i.e. 'আর', 'ও') etc. There are too many variations of words in this experimental data set, and therefore this experiment move to find out the most informative words among the corpus.

3.2.2   Starting and Ending Letters
From collected 750 positive and negative Bangla word this research is extracted the starting letter and ending letter and they are given below in table 2.

Table 2. Most frequently used letters in sentiment analysis of Bangla language processing in this experiment.

Bangla Letters	Positive starting letter	Negative starting letter	Positive ending letter	Negative ending letter
অ	15	98		
আ	33	18		
ই			10	3
উ	14	6		
এ				1
ঐ	1			
ও			1	1
ঔ		2		
ক	9	43	11	31
থ	8	10	1	3
গ	4	8	7	4
ঘ	2	5		
চ	1	8		1
ছ				1
জ	4	11	1	1
ঝ		4		
ঞ			1	
ট			3	15
ঠ	1		4	2
ণ			11	11
ত	7	6	15	56
থ			3	5
দ	5	51	7	13
ধ	3		4	5
ন	12	25	17	35
প	30	15	1	9
ফ	2	2	1	1
ব	5	57	3	2
ভ	31	9	4	1
ম	27	13	5	9
য	3	9	12	14
র		8	38	39
ল	3	6	8	13
শ	10	8	6	11
ষ			5	5
স	71	13	1	5
হ	1	10	2	4
া			40	57
ি			15	23
ী			14	15
ু			7	3
ে			26	20
ো			3	5
ং			1	1
য়			15	21

3.2.3   Most Informative Words

Since it know that a frequent word does not necessarily make a relevant word for the learners, here it is used Tf-Idf classifier to find out which words are more informative. Tf-Idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification [20]. The Term Frequency (TF) counts how many frequent each term appears in a document, as following equation-

$$tf\,(t,d) = count\,(t \in d) \dots\dots\dots\dots\dots\dots.(1)$$

And, Document Frequency (DF) counts how many frequent each term appears in documents, as following equation-

$$df\,(t,D) = |\{d\,|\,t \in d, d \in D\}|$$
$$\dots\dots\dots\dots\dots\dots(2)$$

Inverse Document Frequent (IDF) is the inverse of the DF. IDF checks how many documents have a specific term, and assumes that the term is less useful when appears in many documents, as following equation-

$$idf\,(t,D) = log\,\frac{|D|}{df\,(t,D)}$$
$$\dots\dots\dots\dots\dots\dots\dots\dots(3)$$

Finally, The TF.IDF is computed by the combination of the TF and the IDF as follows-

$$tfidf\,(t,d,D) = tf\,(t,d) * idf\,(t,D)\dots\dots\dots(4)$$

Where TF is counted number in a document, IDF is log of $\frac{|D|}{df\,(t,D)}$ and DF is already explained. After the TF.IDF algorithm, the authors can obtain meaningful words from the documents and sets of the documents. Table 3 and 4 shows words that hold top TF.IDF weight in positive and negative sentences.

Table 3. Top 10 Tf-Idf weight value of positive data set

No.	Word	Tf-Idf weight value
1	করে	0.008722615854301763
2	হবে	0.008572225925779319
3	না	0.007519496426122209
4	জন্য	0.006316376997942656
5	অনেক	0.005263647498285546
6	ভালো	0.004962867641240659
7	করতে	0.0046620877841957705
8	করা	0.004361307927150881
9	ধন্যবাদ	0.004060528070105993
10	হয়েছে	0.0037597482130611046

Table 4. Top 10 Tf-Idf weight value of negative data set

No.	Word	Tf-Idf weight value
1	না	0.03228895591636539
2	করে	0.007375649828611383
3	আর	0.006556133180987896
4	নেই	0.004917099885740922
5	করা	0.00442538989716683
6	হয়	0.00442538989716683
7	করতে	0.004097583238117435
8	থেকে	0.003933679908592737
9	কোন	0.003769776579066804
10	কিছু	0.0036058732495433425

3.3 Bangla Sentence Analysis

In this section, the authors will analysis the feature of sentence. Here the author discuss about their procedures in different situation of sentiment analysis.

3.3.1   Sentence Length

Sometimes sentence length plays an important role in text classification. Now it is need to count the length of positive and negative sentences of this training data set to find out some information. Figure 2 shows different lengths of sentences and their occurrences.
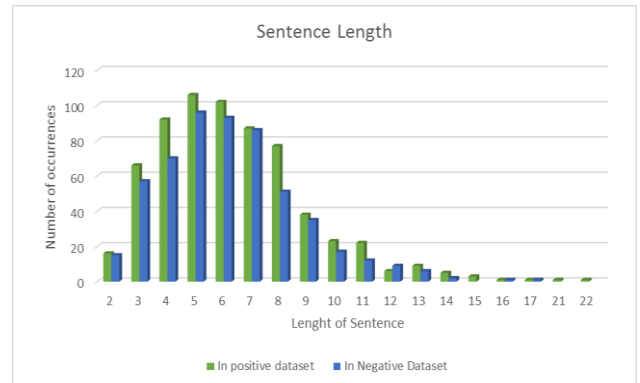

Figure 2: Number of occurrence vs. Sentence length

3.3.2   Pattern Analysis

In this experiment the authors wanted to figure out some patterns to understand positive and negative sentences so that they can categorize them. Every sentence has unique words, but they follow some interesting way of positioning them in sentences. After analyzing some data, the authors realized that if they take last 4 or in some case more word backward form a sentence, they can sort if the sentence is positive or negative. Although, it needs proper pos-tagging of words specially verbs and having them lemmatized. It also needs a rich Bangla positive and negative word collection. This could be the future scope of this study. Table 5 and 6 has some examples of that realization, the bold words the authors think that has an effect of being the sentence is positive or negative.

Table 5. Bangla positive pattern analysis

Bangla Positive Sentence	Pattern
আমি আমার মত **ভাল আছি।**	Positive word + Verb

Bangla Sentence	Pattern
আমি আমার জীবনে অনেক **অর্জন করেছি।**	Positive word + Verb
আমি জীবনের প্রতিটি মুহূর্ত **ভালবাসি।**	Positive word
আমি অনেক **শক্তিশালী।**	Positive word
আমি ভাগ্যে **বিশ্বাস করি।**	Positive word + Verb
সবকিছু **ঠিক** সময় **হয়েছে।**	Positive word + Verb
আশা মানুষকে **জাগ্রত করে।**	Positive word + Verb
আমার পরিবার আমার সব চাওয়া **পূর্ণ করে।**	Positive word + Verb
তারা একটি সুইমিং পুলে **সাঁতার কাটছে**	Verb
আমার আপেল সবচেয়ে **ভাল লাগে**	Positive word + Verb
রতন ক্রিকেট খেলা **পছন্দ করে**	Positive word + Verb
সূর্য পূর্ব দিকে **উঠে**	Verb

Table 6. Bangla negative pattern analysis

Bangla Negative Sentence	Pattern
সে **ভাল নেই।**	Positive Word + Negation
আমার সাইন্স ফিকশন **পড়তে ভাল লাগে না**	Verb + Positive Word + Negation
সে **খারাপ** ছাত্র।	Negative Word
কুকুর বিড়ালকে **তাড়া করে না**	Verb + Negation
ইন্টারনেট স্পীড খুব কম **কম**	Negative Word
আমি এখন থেকে **চাই না**	Verb + Negation
তানজির এই কোম্পানিতে **কাজ করে না**	Verb + Negation
আমি তোমার সাথে হাটতে **চাই না**	Verb + Negation
বাচ্চাদের সাথে খেলতে তার **ভাল লাগে না**	Positive Word + Negation
সে পরিশ্রমী **না**	Verb + Negation
ফোনে বেশিক্ষন কথা বলা আমার **পছন্দ না**	Positive Word + Negation
সবাই **ন্যায় বিচার পায় না**	Positive Word + Negation

4. EVALUATION AND RESULTS

In order to evaluate the performance of the learning algorithm that is used, first use the standard precision and recall to measure the positive and negative sentiments for each classifier using sets of features. Then it is need to use the accuracy metric to compare the overall performance of the classifier. Sentiment analysis task can be interpreted as a classification task where each classification label represents a sentiment. Hence, it define and calculate the four metrics for each label (positive and negative) the same way as in general classification task.

Now calculate the percentage of positivity of a test sentence using this equation,

Positivity= (Σ tfidf weight from positive data set) / (Σ tfidf weight from positive data set × Σ tfidf weight from negative data set) × 100 ……………………………………………...(5)

Similarly, calculate the percentage of negativity of a test sentence using this equation,

Negativity= (Σ tfidf weight from negative data set) / (Σ tfidf weight from positive data set × Σ tfidf weight from negative data set) ×100 …………………………………………....(6)

If total positivity is greater than total negativity of that sentence, it is assumed the sentence is positive. Otherwise, it can consider the sentence is negative.

This research took 50 positive sentences and 50 negative sentences for testing purpose and through their training data. Then it's evaluate the precision, recall and accuracy.

In a classification task, precision, recall and accuracy are explained using four terms - true positive, true negative, false positive and false negative.

True Positive (tp) is defined as the number of sentences, from the test set, correctly labeled by the classifier as belonging to a particular class or label.

True Negative (tn) is defined as the number of sentences, from the test set, correctly labeled by the classifier as not belonging to a particular class or label.

False Positive (fp) is defined as the number of sentences, from the test set, incorrectly labeled by the classifier as belonging to a particular class or label.

False Negative (fn) is defined as the number of sentences, from the test set, that are not labeled by the classifier as belonging to a particular class or label but should have been.

Now define the evaluation metrics using these four terms as follows:

Precision is the number of sentences in the test set that is correctly labeled by the classifier from the total tweets in the test set that are classified by the classifier for a particular class. That is,

$$Precision\ (P) = \frac{tp}{tp+fp}$$ …………………………………(7)

Recall is the number of sentences in the test set that is correctly labeled by the classifier from the total tweets in the test set that are actually labeled for a particular class. That is,

$$Recall(R) = \frac{tp}{tp+fn}$$ …………………………………...(8)

Accuracy is the percentage of tweets in the test set that the classifier correctly labels. That is-

$$Accuracy(ACC) = \frac{tp+tn}{tp+tn+fp+fn} \times 100$$
……………………(9)

The experimental results are given below in 'Table 7'. This result is based on the total collection of document or corpus of different sentiment in Bangla language.

Table 7. Result representation using Confusion Matrix

		Predicted Condition			
	Total Sentences (100)	Predicted Condition positive	Predicted Condition negative	Prevalence= $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total Semtemces}}$	
True Condition	Condition Positive (50)	True positive= 43	False negative= 7	True positive rate (TPR) = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$ = 0.86	False negative rate (FNR) = $\frac{\Sigma \text{ Flse negative}}{\Sigma \text{ Condition positive}}$ = 0.14
	Condition Negative (50)	False positive= 10	True negative= 40	False positive rate (FPR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$ = 0.2	True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$ = 0.8
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive}+ \Sigma \text{ True negative}}{\Sigma \text{ Total Sentences}}$ 100 = 83%	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$ = 0.8113	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$ = 0.1489	Positive likelihood ratio (LR+) = $\frac{TPR}{FPR}$ = 4.3	Diagnostic odds ratio (DOR) = $\frac{LR+}{LR-}$ = 24.57
		False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$ = 0.1887	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$ = 0.8511	Negative likelihood ratio (LR−) = $\frac{FNR}{TNR}$ = 0.175	

5. CONCLUSION

On limitation of Supervised methods cannot always be used, because labelled corpora are not always available. Besides there are some noise in the data. Though this research is select a pure labelled data for their experiment. The system described in this paper proposes a method to recognize the sentiment or opinion and extract a unique feature to come out a better approach to understanding sentiment from Bangla text using. In future, the initial task is that it can evaluate this feature and implement a new method to utilize it and give them a better accuracy to understand the sentiment more precisely by using Bangla text.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] K. M. Azharul Hasan, Md Sajidul Islam, G. M. Mashrur-E-Elahi, Mohammad Navid Izhar, "Sentiment Recognition from Bangla Text", Technical Challenges and Design Issues in Bangla Language Processing, 2013.

[2] K. M. Azharul Hasan, Al-Mahmud, Amit Mondal, Amit Saha, "Recognizing Bangla Grammar using Predictive Parser", International Journal of Computer Science & Information Technology (IJCSIT), 3: 6. pp. 61-73, 2011.

[3] K.M.A Hasan, A.Mondal, A.Saha "A context free grammar and its predictive parser for bangla grammar recognition" 13th International Conference Computer and Information Technology (ICCIT), pp. 87-91, 2010.

[4] Md. Asfaqul Islam, K. M. Azharul Hasan, Md. Mizanur Rahman, "Basic HPSG Structure for Bangla Grammar", Proceedings of the 15th ICCIT, pp. 185-189, 2012.

[5] Das D. and Bandyopadhyay S. (2010a). Finding emotion holder from Bengali blog texts -An unsupervised

syntactic approach. PACLIC 24 - Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation.

[6] M Tomohiro Fukuhara, Hiroshi Nakagawa and Toyoaki Nishida.Understanding sentiment of people from news articles:Temporal sentiment analysis of social events. Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2007.

[7] Baroni M and Vegnaduzzo S. Identifying subjective adjectives through web-based mutual information. Proceedings of Konvens, pages 17-24, 2004.

[8] Vasileios Hatzivassiloglou and Janyce Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the International Conference on Computational Linguistics (COLING), 2000.

[9] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the Association for Computational Linguistics (ACL), pages 271-278, 2004.

[10] Soo-Min Kim and Eduard Hovy. Automatic detection of opinion bearing words and sentences. In Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), 2005.

[11] Hu and Liu. Mining and summarizing product reviews. Proceedings of 10th ACM SigKDD, 2004

[12] Michael Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of the International Conference on Computational Linguistics (COLING), 2004.

[13] Esuli A and Sebastini F. Determining the semantic orientation of terms through gloss analysis. Proceedings of CIKM, 2005.

[14] Nozomi Kobayashi, Kentaro Inui and Yuji Matsumoto. Opinion Mining from Web documents: Extraction and Structurization. Journal of Japanese society for artificial intelligence, Vol.22 No.2, special issue on data mining and statistical science, pages 227-238, 2007.

[15] Yejin Choi, Clarie Cardie, Ellen Riloff and Siddharth Patwardhan. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns Proceeding of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 355-362, 2005.

[16] Andrew Smith, Trevor Cohn and Miles Osborne. Logarithmic Opinion Pools for Conditional Random Fields. In Proceeding of the 43rd Annual Meeting of the ACL, pages 18-25, 2005.

[17] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 412-418, 2004.

[18] Das, D., Roy, S., and Bandyopadhyay, S., "Emotion tracking on blogs - A case study for Bengali", In Jiang, H. et al. (Eds.), IEA/AIE 2012, (LNAI) ol. 7345, pp. 447–456). Springer, 2010.

[19] Dasgupta, S., & Ng, V., "Topic-wise, sentiment-wise, or otherwise? Identifying the hidden dimension for unsupervised text classification" In proceeding of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 580—589, EMNLP, 2009.

[20] http://www.tfidf.com/   Date:10/2/2016 Time: 8:35 AM