An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques

Conference Paper · February 2019

DOI: 10.1109/CCOMS.2019.8821658

5 authors, including:

Rashedul Amin Tuhin
East West University (Bangladesh)
28 PUBLICATIONS **219** CITATIONS
SEE PROFILE

Bechitra Kumar Paul
Human Resocia Co., ltd.
1 PUBLICATION **83** CITATIONS
SEE PROFILE

Faria Nawrine
University of Dhaka
2 PUBLICATIONS **170** CITATIONS
SEE PROFILE

Amit Kumar Das
East West University (Bangladesh)
56 PUBLICATIONS **1,419** CITATIONS
SEE PROFILE

2019 IEEE 4th International Conference on Computer and Communication Systems

An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques

Rashedul Amin Tuhin, Bechitra Kumar Paul, Faria Nawrine, Mahbuba Akter, Amit Kumar Das
Department of Computer Science and Engineering
East-West University
Dhaka, Bangladesh
e-mail: mcctuhin@ewubd.edu, bechitra@outlook.com, nawrinefaria@gmail.com, mahbubaakter74@gmail.com, amit.csedu@gmail.com

Abstract— **Sentiment analysis has become a leading context for scientific and commercial market research in the field of machine learning. Currently, it's a more prominent research field of Bangla language processing system as there are few research works regarding sentiment analysis for this language. In essence, sentiment analysis is an automated process of text mining to determine the emotion from a given text. By using sentiment analysis, a given text can be categorized into several emotions. This paper deals with six individual emotion classes- happy, sad, tender, excited, angry and scared. Here, we proposed two methods of machine learning techniques- Naïve Bayes Classification Algorithm and Topical approach to extract the emotion from any Bangla text. Proposed methods have been applied for both article and sentence level of scope. A comparative analysis of the performance between these two methods has been done, and the topical approach achieved the best performance for both levels of magnitude.**

Keywords- Sentiment analysis, Supervised learning.

I. INTRODUCTION

Sentiment Analysis which is also known as opinion mining has become a topic of much interest and development for research area as it has many practical and empirical applications. It explores people's sentiments, opinions, behavior, attitudes, and emotions towards entities such as individuals, organizations, products, services, issues and their attributes from written or spoken language [1]. Since publicly and privately accessible information over the internet is always thriving, a massive number of texts revealing opinions are available in different blogs and other articles in online journal, social media, and product review sites. Besides this, with the current progress of machine learning, the competency of algorithms to analyze the text has progressed numerously. Because of that, the conduct of sentiment analysis is developing day by day in the field of product analysis, social media monitoring, and market analysis and so on.

Bangla is the fourth most popular language in the world and it has approximately 250 million native speakers all over the world. Numerous research works have been done on sentiment analysis for English and other languages such as Chinese, Hindi, Urdu, and Arabic ([2]- [3]) while sentiment analysis for Bangla is still at a constructive stage. There exists a few research works for Bangla on sentiment analysis due to its lack of resources and its complexity. Therefore this research paper has been done to detect sentiment from Bangla text document.

The principal obligation of this sentiment analysis process is to classifying the polarity of a given text whether the expressed emotion of a text document on Bangla is happy/sad/angry/tender/excited/scared concerning training data. We use these six basic emotions as our emotion class because each of this emotion class can represent many emotions. For example joy, smile, optimistic, laugh, pleased these five emotions can be merged by using happy class and so on.

Sentiment analysis can be applied at different level of scope. In this research paper sentence and document level of magnitude for determining emotions has been implemented. This research work extracts emotion by using Automatic systems which are opposite of the rule-based systems and don't confide in the manually crafted rules. This system relies on machine learning techniques to learn from data.

In this paper, two approaches are proposed to extract the emotion from Bangla text. Those approaches are- Naïve Bayes classification algorithm and Topical approach. We used manually created data corpus with 7,500+ sentences as learning materials. We operate a comparative study between these two methods by experimenting with a large number of test dataset with a different combination of texts and the topical approach achieved the best performance.

The rest of the part of this paper is organized by the following sections. Section 2 provides a short description of some previous works related to our topic. Section 3 gives a simple view of the overall working procedure and provides the information regarding dataset preparation. In Section 4, we briefly describe the methodology for both proposed methods. In Section 5, to compare the result different experiments have been done between these two proposed methods. In Section 6, we conclude the paper and provide some directions for future works.

II. RELATED WORKS

As far as the author know none of the research work on sentiment Analysis from Bangla text has been done to determine the emotion for six individual classes by using

978-1-7281-1322-7/19/$31.00 ©2019 IEEE

these two methods. All the research works on sentiment analysis for Bangla text have been done to assess emotion for a maximum of three categories (Negative, Positive and Neutral) and most of them are done by using NLP techniques.

In paper [4] the author proposed a hybrid mechanism to extract the opinion from both Bangla and English text based on news corpus by combining both rule-based and automated system and implemented this by using both NLP and SVM. In paper [5] the author used the concept of term frequency and inverse document frequency (Tf and IDF) value to get a better solution, and they achieve a more accurate result by extracting the different features of positive, negative or neutral words of Bangla text. The researchers in [6] proposed an emotion tracking system on topic or event was carried out by employing sense based affect scoring techniques by using SentiWordNeT for both Bangla and English text. The researchers of paper [7] developed their methodology to detect the sentiment from Bangla text using contextual valency analysis and to implement this they used WordNet and SentiWordNet to get the sense and polarity of each word in the text. In paper [8] they used Support Vector Machine (SVM) Algorithm and Maximum Entropy (MaxEnt) to automatically extract the sentiments whether the polarity of text is either positive or negative from Bangla Microblog (Twitter) posts. Lots of research work on sentiment analysis has been done for the English Language by using both NLP and machine learning techniques. In the paper [2]four approaches were proposed (Topical approach, Emotional approach, Retrieval approach, and Lexicon approach) to calculate the emotional score of English words in messenger logs for six individual classes. And compare to other procedures Topical approach gives the best performance. In paper [9] the researchers used three machine learning techniques- SVM, MaxEnt and Naïve Bayes to classify the sentiments of Twitter messages with emoticons. In paper [10], [11] the authors used LDA (Latent Dirichlet Allocation) model to extract emotion from the particular text document. In paper [10] the researchers introduced an LDA based model for interpreting sentiment, and they used it for giving rank to the tweets concerning their popularity. In paper [11] both LDA and SVM have been used to specify the opinions from IMDB movie review dataset. The authors in the paper [3] proposed a system for sentiment analysis of Hindi movie review by using HindiSentiWordNet (HSWN) and Synset replacement algorithm.

III. PROPOSED MODEL

To make an automated system for emotion detection, we used machine learning techniques due to its proven accuracy level. Among three broad categories of machine learning classification approaches the supervised learning approach has been used because of its compatibility in modeling and regulating dynamic systems. Here two techniques (Naïve Bayes Algorithm, Topical approach) have been proposed for detecting emotion from Bangla text for six individual emotion classes. Figure1 represent the simple working procedure for this study-

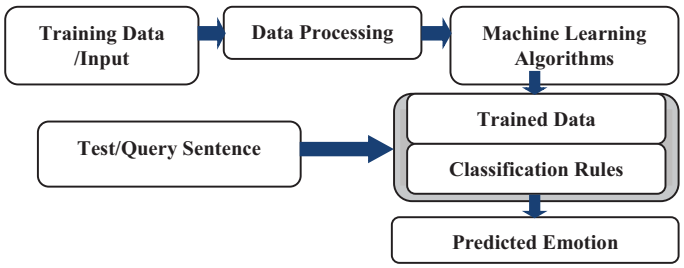

Figure 1. Model of working mechanism for both proposed approaches

A. *Dataset Preparation*

To conduct the procedure with a better performance the first and the most important part is to prepare a suitable dataset. For better execution, we used manually created data corpus consisting of 7,500 Bangla sentences. Total dataset split into two parts- training set and test set.

B. *Training Dataset Preparation*

We took 7,400 Bangla sentences as our training dataset. To make it compatible for the execution, we manipulate this dataset based on the needs. It is elected to keep the dataset set quite specific and straightforward to ignore groundless complexity. We ignored mixed sentences, special characters and punctuation marks. As mentioned before, in this paper six basic and conventional emotion classes have been used to detect emotion from Bangla text. For that reason, it is needed to set an emotion for each sentence in the training dataset. Manually for each sentence, a suitable (based on the meaning) emotion has been set (Table 1). For better performance, we try to provide at least one sentence correspond to each emotion class for each unique word. For example, we consider a sentence "আমি ভালো আছি ", here exist 3 individual words. For each individual word of this sentence, we will try to provide at least one sentence corresponds to each emotion class (Table 2).

TABLE I. MAPPING OF MOTION CLASS

Happy	*Tender*	*Excited*	*Sad*	*Angry*	*Scared*
Fulfilled	Intimate	Ecstatic	Down	Irritated	Tense
Contented	Loving	Energetic	Mopey	Miffed	Nervous
Glad	Warm-Hearted	Aroused	Grieved	Upset	Anxious
Complete		Bouncy	Dejected	Mad	Jittery
Satisfied	Touched	Nervous	Depressed	Furious	Frightened
Pleased	Kind	Perky	Heartbroken	Raging	Terrified

TABLE II. PROVIDING EMOTION CLASS FOR EACH WORD

Emotions	আমি	ভালো	আছি
Happy	আমি সুখে আছি	আমি ভালো আছি	আমি আনন্দে আছি
Tender	আমি তাকে ভালোবাসি	সে অনেক ভালো মানুষ	আমি তার সাথে সুখে আছি

Excited	আমি আজ অনেক আনন্দিত	আমি ভালো ফলাফল করায় সবাই আনন্দিত	আমি অনেক আনন্দে আছি
Sad	আমি সুখে নেই	আমি ভালো নেই	আমি কষ্টে আছি
Angry	আমি তার সাথে রাগ করেছি	সে ভালো কাজ করেনি	আমি তার সাথে রেগে আছি
Scared	আমি তাকে ভয় পাই	সে ভালো না তাই তাকে ভয় হয়	একা আছি তাই ভয় হচ্ছে

IV. THE METHODOLOGY OF PROPOSED APPROACHES

A. *Naïve Bayes Classification Algorithm*

Naïve Bayes classification is among the most successful known algorithms for learning to classify text documents. In this approach, based on probability it provides an emotional score for each word w_i correspond to each emotional class c_j. By using this emotional score, it tries to detect emotion for query sentence. For better understanding, all the steps of this approach are described here with equation-

1. Emotion Detection in Sentence level

Step1: Firstly Probability of each emotion class c_j is calculated by using equation (1), where **NS(c_j)** denotes the number of sentences containing $\mathbf{c_i}$ class and N denotes the total number of sentences in the dataset.

$$\mathrm{P(cj)} = \frac{\mathrm{NS(cj)}}{\mathrm{N}} \quad (1)$$

Step2: The probability of each word $\mathbf{w_i}$ in the dataset corresponding to each emotion class $\mathbf{c_j}$ is calculated by using equation (2). Then for each word $\mathbf{w_i}$ we get probability value for six individual class c_j.

$$P(w_i, c_j) = \frac{NS(w_i, c_j)}{NS(c_j)} \quad (2)$$

Here, **NS (w_i,c_j)** means number of sentences that contains w_i word for c_j emotion class.

Step3: To detect emotion of a query sentence firstly we try to find out whether each word of this sentence is available in the training dataset or not. If those words exist in the training data, then we determine their probability correspond to six individual classes by using equation (2). Then by using the probability of each word, we calculated the emotional score of this query sentence for each emotion class by applying equation (3). Finally, we detect emotion **(e_d)** for the query sentence by using the highest value from six individual emotion classes. To detect this we use equation (4).

$$EQ(c_j) = \prod p(w_i, c_j) \quad (3)$$

$$\mathrm{Emotion(e_d)} = \mathrm{Maximum[EQ(c_j)]} \quad (4)$$

But the problem will arise if any word from the query sentence is not available in the dataset. Because the probability for this word will be 0 and when we will multiply this with other words probability the product term will be 0. It will affect the accuracy of the result. To solve this problem, Binning technique is implemented here.

Binning Technique For better performance, we use the binning technique to provide a value for missing words in the dataset. We supposed a query sentence "আজকের দিনটা সুন্দর ছিলো" and consider 'ছিলো' this word is missing in the training dataset. Firstly, we will calculate the probability of other words on this sentence by using equation (2). Then we will calculate the mean value among those words probability. Let consider the probability of these three words are- P(আজকের) = 0.3, P(দিনটা) = 0.35, P(সুন্দর) = 0.07. Now applying equation (5) and (6) we will calculate the probability of a missing word, Where TW denotes total word in the sentence and MW denotes the number of missing words.

$$Mean\ value = \frac{\sum p(other\ words)}{N(other\ words)} \quad (5)$$

$$P\ (\text{missing word}) = \text{Mean value}*(\text{TW-MW}) \quad (6)$$

$$\mathrm{P(ছিলো)} = \frac{0.3+0.35+0.07}{3}*(4-1)=0.72$$

2. Emotion Detection in Article/Document level

In article level of scope firstly we differentiate all the sentences in the article. Then we applied sentenced level emotion detection technique (explained above equation 1-3) for each sentence in the article. We get the probability for 6 individual classes for each sentence. We used this probability of each sentence in the article to calculate the emotional score of the whole article **(E_{ar})** for six individual classes by applying equation (7). Finally, from the maximum value among the 6 emotion classes, we detect emotion **(E_a)** for the article

$$E_{ar}(Cj) = \prod EQ(c_j) * P(c_j) \quad (7)$$

$$Emotion(E_a) = \mathrm{Maximum}[E_{ar}(\mathrm{c_j})] \quad (8)$$

B. *Topical Approach*

In this study, we use the generic topical approach to extract the emotion from an article. In this approach the concept of ***tf, idf*** is used to calculate the emotional value of each word in the sentence. Term Frequency-Inverse document frequency in short ***tf- idf*** is a numerical statistic that is used to find out the importance of a word in a document or a corpus. To detect emotion from a text article by using this approach firstly, for each sentence in the dataset the ***idf*** value of each word is calculated by using the equation (9), where $\mathbf{w_i}$ denotes current word, ***NS*** is the total number of sentences in the dataset and $\boldsymbol{NS_{wi}}$ denotes the number of sentences containing this word $\mathbf{w_i}$.

$$idf(w_i) = \log(\frac{NS}{NS_{wi}}) \quad (9)$$

Then the probability of each word $\mathbf{w_i}$ in the dataset corresponding to each emotion class $\mathbf{c_j}$ is calculated by using equation (10)

$$P(w_i, c_j) = \frac{NS(w_i, c_j)}{NS} \quad (10)$$

Where **NS($\mathbf{w_i}$,$\mathbf{c_j}$)** means the number of sentences that contain $\mathbf{w_i}$ word for $\mathbf{c_j}$ emotion class. Then, the ***idf*** value of each word is distributed to the six emotion classes by observing the probability of each sentence that contains word $\mathbf{c_j}$. This is done by using equation (11)

$$(w_i, c_j) = P(w_i, c_j) * idf(w_i) \quad (11)$$

To detect emotion from a query sentence firstly, we need to find out whether all words in the sentence are available in the dataset or not just like Naïve Bayes. If available then, the emotional value for each word corresponds to six emotion class will be calculated by using equation (11). After that, for

each emotion class, the emotional value of the query sentence will be calculated by using equation (12). And finally, the emotion of the query sentence Emotion (QS) will be detected by observing the maximum value among the emotion classes by using equation (13).

$$EQ(c_j) = \sum Emotion(w_i, c_j) \quad (12)$$

$$Emotion(QS) = Maximum[EQ(c_j)] \quad (13)$$

To detect emotion for an article firstly for each sentence in the article we will calculate the emotion value for six emotion classes. By using this value, the emotion value for an entire article will be calculated to six individual categories by using equation (14). After that, the emotion of the given article (QR) will be detected by using equation (15).

$$Emotion(E_{ar}, c_j) = \sum EQ(c_j) \quad (14)$$

$$Emotion(QR) = Maximum[Emotion(E_{ar}, c_j)] \quad (15)$$

V. Experimental Results and Discussion

In this paper, to evaluate the performance of the proposed methods four different experiments were performed. In our first experiment, we have determined the performance for sentence level of scope by using both proposed methods for six individual classes. 100 Bangla sentences were used as our query sentence. Then we calculated the accuracy for both proposed methods. The same thing has been done to measure the performance in article level of scope. One hundred articles from different online news portal have been used as our test data. After that, a comparative analysis has been done between these two methods and a comparison graph is plotted by showing the accuracy for both proposed methods (Figure 2). Among these two methods, the accuracy of topical approach is much higher than Naïve Bayes classification algorithm for both sentence and article level of scope.

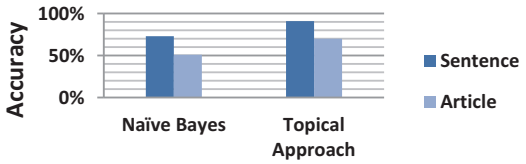

Figure 2. Performance of Naive Bayes and Topical Approach for both article and sentence level of scope (for 6 emotion class)

In our second experiment, we mapped these six emotion classes into two higher emotion classes-Positive and Negative (Table 3).

TABLE III. Mapping into higher emotion class

Higher Emotion class	Basic Emotion Class
Positive	Happy, Excited, Tender
Negative	Sad, Angry, scared

After that, the same procedure of the first experiment has been followed to evaluate the performance of proposed approaches for two higher emotion classes. We get another comparison graph between these proposed methods (Figure 3). From Figure 3, it is clear that the accuracy level of Topical Approach is again much higher than the Naïve Bayes Classification Algorithm.

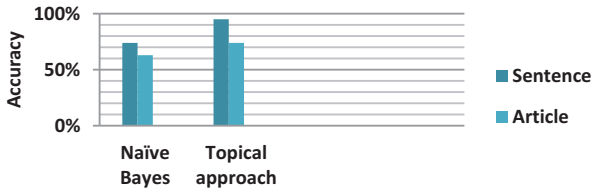

Figure 3. Performance of Naive Bayes and Topical Approach for both article and sentence level of scope (for 2 emotion class)

Another critical factor is that the most frequent words in the datasets and their higher probability for any particular emotion class affect the performance (Figure 4).

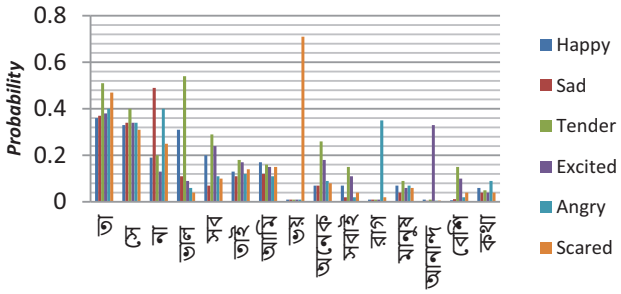

Figure 4. Most Frequent Words in the dataset and their probability for each emotion class

In our third experiment, we try to figure out whether the number of emotion classes affects the performance or not and it is explicit that the number of emotion classes is inversely proportional to the accuracy level. For two emotion classes, the accuracy level for both methods is much higher than six emotion classes. (Figure 5).

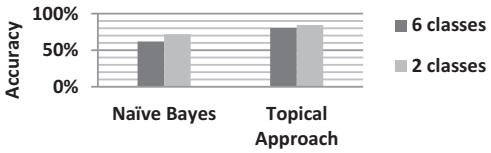

Figure 5. Comparison between these two approaches in respect to the number of classes

From this three experiment, it is seen that the Bayesian Classifier gives less accuracy in the Article level, whereas the topical model still works better. The reason for this less accuracy for the Naive Bayes is pretty straightforward. The main reason is the multiplication operation in between the items of an item set. However, it is not the case for the Topical approach as it takes the decision operation by a summation operation and comparison.

In our final experiment, we have compared our works with two other papers regarding sentiment analysis from Bangla text. We select these two papers based on their working methods and performance level. From Table 4, it is evident that none of the paper used more than three emotion

classes to detect emotion. Beside this, it is visible that our training dataset is much more vibrant than the other research papers even this dataset can be used for future work. Here, another important factor is that in paper one the authors used SVM and they achieved higher accuracy for that. But we didn't use SVM. The main reason is that SVM is much efficient for separating two classes only, but it becomes more complicated when it needs to classify several classes with noisy data. On the other hand, for Topical Approach, it is proven that it can work with several classes with reasonable accuracy. Therefore, considering the overall situation, it can be said that in some respect the performance of this research paper is much better compared to others for Bangla text.

TABLE IV. COMPARATIVE ANALYSIS WITH OTHER PAPERS REGARDING SENTIMENT ANALYSIS FOR BANGLA TEXT

Paper Title	*Classificati-on Methods*	*Number Of Emotion Classes*	*Learning Materials (Data Set)*	*Highest Accuracy*
Our Paper: Sentiment Analysis from Bangla Text Using ML Techniques	1. Naive Bayes Classification Algorithm. 2. Topical Approach	6 emotion classes (Happy, Sad, Tender, Angry, Excited, Scared)	Manually created data corpus with 75000 sentences. 7400 sentences used as training dataset and 100 sentences used as test data	Highest accuracy acquired by Topical Approach (above 90%)
Paper 1: Performing a sentiment Analysis in Bangla Microblog Posts. [8]	1. SVM 2.Maximum Entropy	2 emotion class (Positive or Negative)	1300 Bangla Tweet used as learning material. 1000 sentences used as training data and 300 used as text data	Best accuracy attained by SVM (93%)
Paper 2: Detecting Sentiment from Bangla Text using Machine Learning Technique and Feature Analysis **[5]**	1. Used the concept of term frequency and inverted document frequency (Tf-Idf)	3 Emotion Class (Positive , Negative or Neutral)	1500 short Bangla comment used as learning material, 1400 used as training data and 100 used as test data	Accuracy was 83%

VI. CONCLUSION

Based on today's perspective text has become a treasure trove of revealing useful information and people's opinions regarding anything. So uncover the views from the text is an important task now for so many fields like product analysis, social media monitoring, market research and analysis and so on. Based on these needs our paper works on detecting emotion from Bangla text by using two proposed methods. We achieved a satisfying accuracy of above 90% for Topical approach on the sentence level. Its accuracy for article level is also gratifying. But still there exist some limitations. Due to the lack of smooth Data and Complexity of Bangla Language using supervised method is not an efficient way to mine information from a huge dataset. To make it more realistic in the entire research the primary goal is to find out the expected data from a random data through the data set include lots of missing value and noisy data, and it is clear here why the traditional algorithms like Naïve Bayes fail to extract information. Moreover, there still exist some way to improve performance for sentiment analysis from Bangla text. But for that, we need more labeled data which is time-consuming. So, in the future, we will try to improve its performance by developing a hybrid mechanism using these two approaches and will also implement other approaches like the LDA model, Decision Tree as well. Beside this, we will try to make a useful application for users by implementing these concepts.

REFERENCES

[1] Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, pp. 7-8, May 2012.

[2] Lun-Wei Ku, Cheng-Wei Sun," Calculating emotional score of words for user emotion detection in messenger logs", 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI),8-10 Aug. 2012.

[3] Pooja Pandey, Sharvari Govilkar, "A Framework for Sentiment Analysis in Hindi using HSWN", International Journal of Computer Applications (0975 – 8887) Volume 119 – No.19, June 2015.

[4] Amitava Das, Bandyopadhyay, S. "Phrase level polarity identification for Bengali", International Journal of Computational Linguistics and Applications, 1(2), pp. 169–181, 2010.

[5] Muhammad Mahmudun Nabi, Md. Tanzir Altaf, Sabir Ismail, "Detecting Sentiment from Bangla Text using Machine Learning Technique and Feature Analysis", International Journal of Computer Applications (0975 – 8887) Volume 153 – No 11, November 2016.

[6] Dipankar Das, "Analysis and Tracking of Emotions in English and Bengali Texts: A Computational Approach", WWW 2011 – Ph. D. Symposium, March 28–April 1, 2011, Hyderabad, India.

[7] K. M. Azharul Hasan, Mosiur Rahman, Badiuzzaman, "Sentiment detection from Bangla text using contextual valency analysis", 2014 17th International Conference onComputer and Information Technology (ICCIT), 22-23 Dec. 2014.

[8] Shaika Chowdhury, Wasifa Chowdhury, "Performing sentiment analysis in Bangla microblog posts", 2014 International Conference on Informatics, Electronics & Vision(ICIEV), 23-24 May 2014.

[9] Alec Go, Richa Bhayani, Lei Huang, "Twitter Sentiment Classification using Distant Supervision".

[10] Suvarna D.Tembhurnikar, Nitin N.Patil, "Sentiment Analysis using LDA on Product Reviews: A Survey", International Journal of Computer Applications (0975 – 8887) National Conference on Advances in Communication and Computing (NCACC 2015).

[11] Raja Mohana S.P, Umamaheswari K, Ph.D., Karthiga R, "Sentiment Classification based on Latent Dirichlet Allocation", International Journal of Computer Applications (0975 – 8887) International Conference on Innovations in Computing Techniques (ICICT 2015).