Sentimental Analysis on Multi-domain Sentiment Dataset Using SVM and Naive Bayes Algorithm

P. Kiran Kumar(✉), N. Jahna Tejaswi, M. L. Vasanthi, L. L. Srihitha, and B. Phanindra Kumar

Department of Computer Science and Engineering, SASI Institute of Technology and Engineering (Affiliated To JNTU, Kakinada), Tadepalligudem, India
{kiran, tejaswi5a2, vasanthi588, srihitha576, phanindra509}@sasi.ac.in

Abstract. With the advent of the significant data era, people are confronted with the vast amount of information they receive each day. The quantity of information accrued and processed by Facebook, Twitter, and other significant social networks (such as Instagram) is vast. The Twitter platform encourages users to use 280 characters each to tweet their thoughts. Because tweets can use a limited number of characters, sentiment analysis becomes more accurate. Sentiment analysis is a technique for determining whether a text is positively, negatively, or neutral. Some experiments are conducted using Natural Language Processing Toolkit (NLTK) to determine whether a tweet has a neutral, positive, or negative polarity with accuracy. Moreover, by using Naïve Bayes and SVM, the accuracy of the tweets is compared. Finally, the ROC curve will decide the efficiency of both algorithms.

Keywords: Twitter · Social networks · Sentiment analysis · Machine Learning · Natural Language Processing Toolkit (NLTK) · Naïve Bayes · SVM

1 Introduction

Research on sentiment analysis has become an interesting topic in the current world. It is the study of people's emotions, feelings and opinions with in written documents. Many businesses and organizations rely on the idea of their customers to make better decisions regarding their products, services. Platforms, such as Twitter, Facebook Etc. Allow users to take and analyze the opinions and reviews they make. The abundance of this data makes it difficult for users to interpret it correctly. As a result, sentiment analysis methods would prove helpful.

By analyzing Twitter data, we can obtain helpful information in many different fields. An example of the new information we can receive is people's sentiments on a topic. This information can be valuable in evaluating a project for improvement. By observing tweets, public sentiment information can be obtained. However, it may not be feasible to collect and analyze millions of tweets. Therefore, the existence of an application that automatically crawls and analyzes tweet sentiment will be beneficial.

D. Garg et al. (Eds.): IACC 2021, CCIS 1528, pp. 201–213, 2022.
https://doi.org/10.1007/978-3-030-95502-1_16

Several social networks and microblogging platforms allow users to express their opinions on many features of their lives in as little as 280 characters. Twitter is the most famous microblogging and social media platforms. Many of these tweets are challenging to analyze because of misspellings, emojis, and slang words. They all should go through some preprocessing steps before they are used for polarity detection (positive, negative, neutral).This is where sentimental analysis raised.

The following are the steps used in sentimental analysis.

Step1: Tokenization.

It is nothing but dividing a paragraph into a set of statements or dividing a comment into different words.

Step2: cleaning the data.

It is to remove all the special characters or any other word that does not add value to the analytics part.

Step3: removing the stop words.

Stop phrases do now no longer upload tons cost to the analytics part.

Step4: classification.

Our primary task is to classify them as positive words, negative words, or neutral words.

This paper aims to examine and understand the people sentiments using different machine learning techniques like Naïve Bayes and SVM. In this section accuracy and precision of both algorithms were compared to determine the best outcome of the two. All the above mentioned techniques are supervised learning techniques, which means that the desired data must first be trained in all of these instances. Finally, roc curve decides the best efficient algorithm from above mentioned techniques.

2 Related Works

A. Pak et al. developed a sentiment classifier using machine learning techniques to determine the polarity of tweets. The experimental assessment of the proposed strategies suggests development over the preceding methods. The collected corpus was used to train a sentiment classifier capable of detecting positive, negative, and neutral texts. The classifier is based on a multinomial Nave Bayes classifier using N-grams and POS-tags as features [1].

Sahar A. El Rahman et al. proposed a field of study known as sentiment analysis studies' opinions on several social media sites. The proposed model used several algorithms to increase the accuracy of identifying positive and negative tweets. We used an unsupervised machine learning algorithm where previously labelled data did not exist as a first step. Then the lexicon-based algorithm was used to feed the data into several supervised models [2].

Amrita Shelar et al. conducted an exploratory analysis of data from Twitter. We applied techniques in sentiment analysis and discovered people's sentiments in the form of polarity. As a future roadmap, we plan to gather more information about the users and businesses to research potential donors for non-profits [3].

Ritu S. Karan et al. proposed a slang improvement system to improve product popularity on Twitter by scaling the location of tweets. As a result, system performance

is enhanced. An additional feature like location recognition of tweets helps in the product, marketing, political issue and event decision making. Thus, the performance of the system is improved. An additional feature such as location recognition of tweets is helpful when making decisions regarding events, campaigns, and political events [4].

Ike Pertiwi Windasari et al. described those thousands of Twitter users who post their opinions on their tweets every day. Businesses can take advantage of this information, but it takes a long time to do so. This is why there has to be a sentiment analysis to predict tweet sentiment. In this study, we focused our search on keywords related to online transportation, particularly GoJek [5].

Huma Parveen et al. discussed how to preprocess tweeting data to remove noise. These different types of tweeting data cannot be used directly; they need to be converted first before being used. The analysis of Twitter data is done from various perspectives such as Positive, Negative and Neutral sentiments on tweets. This type of analysis will surely help any organization to improve its business productivity [6].

Using three manually annotated datasets collected from Twitter, this study analyzes the sentiment of tweets [7].

A. Agarwal et al. was proposed that a unigram model be used as a baseline, with a gain of 4% in binary classification and 3-way classification. Then, it was examined the tree kernel model and feature-based models, which surpassed the unigram model [8].

Lokesh Mandloi et al. presented these different machine learning techniques of data analysis of tweets, including Naive Bayes, SVM, and the Maximum Entropy Method. Twitter data is analyzed from various viewpoints to learn more about the sentiment analysis of Twitter. It is essential to know that sentiment analysis involves opinions that are classified into positive, negative, and neutral. Most studies have shown that Naive Bayes is the best machine learning method for predicting emotions [9].

Sanjeev Dhawan et al. proposed a sentiment polarity model for sentiments analysis in online social networks using tweet datasets. In the proposed methodology, tweet datasets are obtained from Twitter APIs to analyze sentiments emotions from different users. In this section, we check the polarity of sentiment within every tweet. Polarity is defined as the emotions of users like joy, happiness, sadness, and anger. If the polarity is equal to zero, then the tweet is neutral, and the polarity is more significant than zero, then the tweet is positive. Otherwise, the tweet is negative. By identifying tweets based on their sentiment polarity, the proposed algorithm can identify tweets of various users in this manner [10].

Hao Wang et al. examined a real-time political sentiment analysis problem, using nave Bayes to classify tweets into four categories (positive, negative, neutral, or unsure) and assess whether the system applied to the analysis of tweets during elections. Several other domains could be accessed by using this method, including movie events [11].

Neetu M et al. studied the sentiment analysis problem of tweets related to the domain of the electronic product. The machine learning approach performs better than symbolic techniques that are based on sentiments identification. The performance of the proposed enhanced vector feature was evaluated using several classifiers, including the nave Bayes, Maximum Entropy, SVMs, and Ensembles, and the results were almost similar. In the domain of electronic products opinion, the proposed feature vector showed improved performance [12].

Mohd Naim Mohd Ibrahim et al. proposed the Twitter Sentiment Classification Using Naïve Bayes Based on Trainer Perception. In this paper, researchers identified 50 tweets containing the keywords 'Malaysia' and 'Maybank' from Twitter to train perception processes on feedback from trainers. Using Naive Bayes techniques, 36 tweets were classified as positive, 42 as unfavourable or 57 as neutral. The study involved 27 trainers who were asked to analyze 25 tweets at random for their sentiment and then apply Nave Bayes training to the remaining 25 tweets based on their performance. According to the study, accuracy was 90% * 14% based on the total number of correct tweets [13].

Anis Zarrad et al. proposed that a demo of real-time Twitter sentiment analysis was developed for Bing maps as part of Microsoft Azure. The demo displayed opinions as positive, negative, or neutral tweets, and the statuses were displayed as different colours on the map [14].

The study by ParisaLak et al. showed that sentiment analysis does better at capturing general sentiment in star ratings (negative, neutral, or positive) [15].

Geeta R et al. provides a method by which people's opinions from different locations can be found regarding a particular product. The number of tweets needs to be significant to arrive at accurate results. As a result, even if a device or product does not collect an astronomically enormous amount of data, we can still collect tweets for several months from the data centres. Future work is expected to produce sentiment resources that were not available at this time [16].

Srinidhi Bhat et al. described that With the rapid increase of microblogging sites, there are many opportunities for extracting public opinion and analyzing it in a predictive manner, like sentiment analysis. The experimental assessment of the proposed strategies suggests development over the preceding methods. Rather than adding the value of these adverbs with the whole tweet sentiment, the proposed model calculated the score by multiplying its deal with the adverb like very and much [17].

Ahdi Ramadani et al. classified a relationship between the amount of training data and the classification performance. The more training data used, the more accurate and reliable the classification. Following preprocessing of 8000 raw datasets, 4845 training data points were selected for the study. As the collection of data is improved, the amount of gathered data should also be increased so that classification performance can improve [18].

Abhijit Janardan Patankar et al. have attempted to investigate the performance of states using big data Hadoop to retrieve the online tweets and assign scores for particular tweets that can be used for analyzing the performance of various communities on a large scale. Sentiment analysis has been implemented successfully with Twitter. We also studied NLP and machine learning approaches to sentiment analysis. We have learned many applications for sentimental analysis, and it is an important one to check [19].

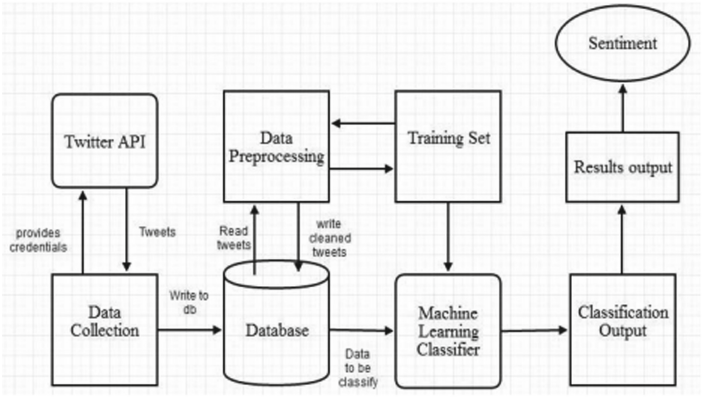
In this paper, an algorithm to analyze Twitter sentiments based on sentiment polarity is proposed in an online social network. Twitter datasets are obtained from Twitter API for analysis of Twitter sentiments emotions of different users. Using the sentiment polarity of each tweet, we determine whether it is positive or negative. If comments are equal to zero, then the tweet is neutral; otherwise, notifications are negative [10].

3 Proposed Methodolgy

Sentiment Analysis
As one of the maxima extensively used strategies in text mining, sentiment analysis refers to analyzing how the text (here, Twitter tweets) is framed in positive, negative, and neutral sentiments. This method investigates tweets, conversations, opinions, and views (to decide business strategy, political analysis, and assess public action).

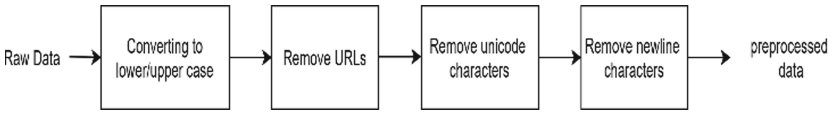
Twitter
Twitter is an American microblogging and social media platforms that permits it's users to post, like, and retweet tweets.

Flow Diagram



Data Collection
Data extracted from Twitter API can be used to generate large datasets of tweets that are not publicly available. To access Twitter data, one needs credentials from the developer site, so one may enter a search query to gain access to it.

Data Processing



NLTK
By using NLTK, we can employ the process of sentiment analysis to analyze linguistic data to gain insights from linguistic data using powerful built-in machine learning operations.

These are the steps involved in NLTK.

- It is generating the listing of phrases withinside the tweet.
- Removing stopwords and words with unusual symbols
- Normalizing the comments in tweets (Fig. 1).

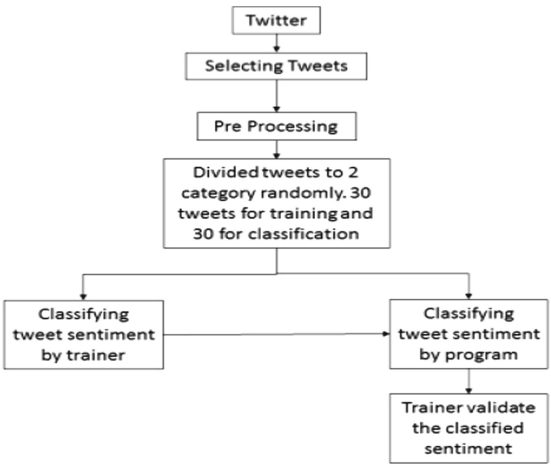


Fig. 1. Overall process

3.1 Naive Bayes Approach

1. Naive Bayes Approach

In Natural Language Processing, Naive Bayes analyses text data and therefore gives us a good result. This classification algorithm is the result of applying the Bayes Theorem to classification algorithms. The Naive Bayes algorithm uses mixture models to determine the probability of the results of the data analysis. The classifier used the concepts in mixture models to determine the effects. Combining a mixture model with the Bayes theorem can perform as a probabilistic classifier. Naive Bayes is likewise called simple Bayes or independence Bayes.

$$P(a/b) = P(b/a)P(a)/P(b)$$

Above, P(a) is the probability of class. P (b) is the prior probability of predictor. P (a/b) is the probability of class m. Where m is the target and predictory is the attribute. P (b/a) is the probability of predictor of the given type.

Input:

Training dataset T,

F= (f_1, f_2, f_3,.., f_n) // value of the predictor variable in testing dataset.

Output:

A class of testing dataset.

Step:

1. Read the training dataset T;
2. Calculate the mean and standard deviation of the predictor variables in each class;
3. Repeat

   Calculate the probability of f_i using the gauss density equation in each class;

   Until the probability of all predictor variables (f_1, f_2, f_3,.., f_n) has been calculated.
4. Calculate the likelihood for each class;
5. Get the greatest likelihood;

3.2 SVM

SVM stands for Support Vector Machine which attempts to learn universally. Support Vector Machine has equal input and output, offering output to be positive or negative and information in vector space. Text documents are not helpful as input to Support Vector Machine. Texts are transformed into structured file formats, which can be employed as inputs for machine learning algorithms. The text scores are calculated and then used as inputs for Support Vector Machines. Text categorizations are compared to finalize the best one between texts. The performance estimation is used to determine which one is the most powerful. For text categorization SVM has been proven as the most important learning algorithms (Fig. 2).

Inputs:Determine the various training and test data.
Outputs:Determine the calculated accuracy.
Select the optimal value of cost and gamma for SVM.
while (stopping condition is not met) **do**
    Implement SVM train step for each data point.
    Implement SVM classify for testing data points.
end while
Return *accuracy*

Fig. 2. Algorithm for SVM

4 Results and Discussion

4.1 Accuracy Classification

Accuracy refers to how close something is to a known value. It is defined as the number of successful recognized entities by total number of entities in the data.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

where TP denotes True Positive (case prediction of correct positive), TN denotes True Negative (case prediction of correct negative), FP denotes False Positive (case prediction of incorrect positive), and FN is False Negative(case prediction of incorrect negative).

When compared to NLTK method, Stanford NER Tagger method is more accurate [10]. The accuracy between Stanford NER Tagger and NLTK method can be plotted:

4.2 Confusion Matrix

Although the confusion matrix is not a performance indicator in and of itself, its components are critical for algorithm evaluation. It gives a matrix-like output with TP, TN, FP, and FN values, just like the accuracy metric.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

a. Precision

Precision metric is computed by dividing the true positives by the sum of predicated positives, the accurate a classifier out of the predicated positives. Precision is known as the number case predictions of correct positive (True Positive) by the total number of case predictions of correct positive (True Positive) and case predictions of incorrect positive(False Postive). As a result, a high precision value represents that the algorithm produced a relevant result.

$$\text{Precision} = \text{TP}/(\text{FP} + \text{TP}) \tag{2}$$

b. Recall

Recall is known as the number case predictions of correct positive (True Positive) by the total number of case predictions of correct positive (True Positive) and case predictions of incorrect negative(False Negative). This means that when a model predicts a favorable outcome, the precision assures that the objects are classified as such. As a result, a high precision value indicates that the algorithm produced a meaningful result.

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{3}$$

c. F1 Score

It is computed by using the following formula.

$$\text{F1 Score} = \text{TP}/(\text{TP} + 1/2(\text{FP} + \text{FN})) \tag{4}$$

From the above formulas and discussion we have calculated the precision, recall and f1 value and Accuracy.

Here for result prediction, we use SVM and Naïve Bayes Algorithms. For Naïve Bayes, we use multinomial Naive Bayes, and for SVM, we use SVC linear kernel. Finally, the result of SVM is more accurate compared to the development of Naïve Bayes.

In the end, we test both algorithms by reviewing Twitter to see which is the best performer. The results are shown in Table.

	SVM	Naïve Bayes
Accuracy	96.27	93.69
Precision	98	99
Recall	96	94
F1 Measure	97	96

The above Table shows the comparison of SVM and Naïve Bayes algorithms. The accuracy, precision and recall of SVM are 96.27, 98 and 96 respectively, where as the accuracy, precision and recall of Naïve Bayes are 93.69, 99 and 94 respectively.

The below bar charts displays the accuracy, precision, recall and F1 score of above discussed algorithms.

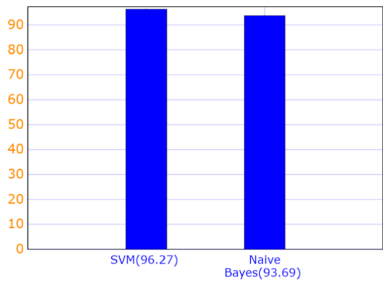


Fig. 3. Comparison of Accuracy on Both Algorithms.

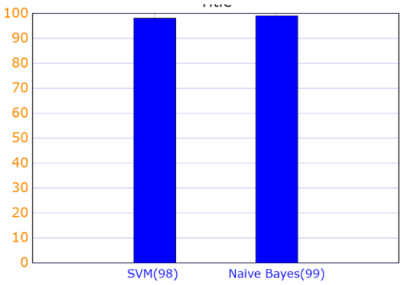


Fig. 4. Comparison of Precision on Both Algorithms

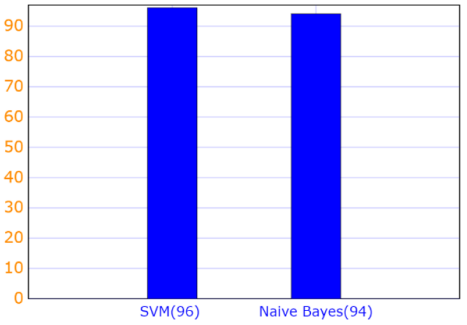


Fig. 5. Comparison of Recall on Both Algorithms.

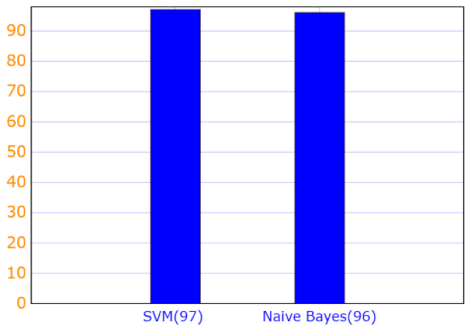


Fig. 6. Comparison of F1 Score on Both Algorithms.

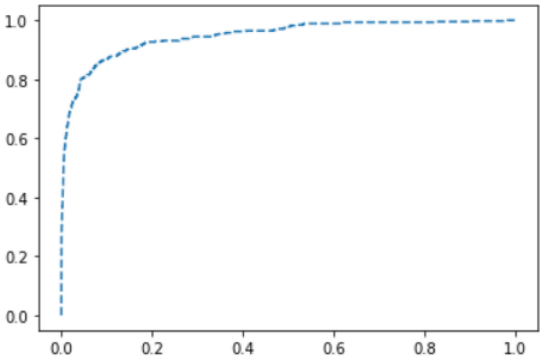


Fig. 7. ROC curve of SVM

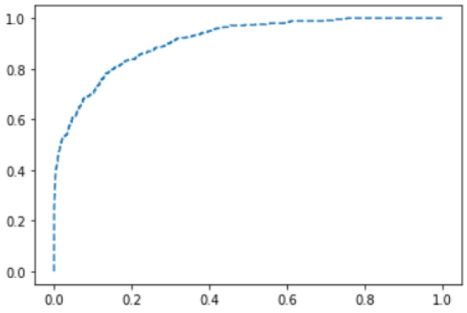


Fig. 8. ROC curve of Naïve Bayes

The area under the Fig. 3 curve is more than the area under the Fig. 4 curve from the above two ROC curves. Figure 3 is discussed about SVM, whereas Fig. 4 is discussed about Naïve Bayes. Finally, SVM is more efficient than Naïve Bayes for sentiment analysis (Figs. 5, 6, 7 and 8).

5 Conclusion

Some Machine Learning Algorithms like Naïve Bayes and SVM algorithms are used in Twitter Sentimental Analysis. Machine Learning techniques are more straightforward and efficient than Symbolic techniques. These two techniques are used in finding the accuracy, F1 measure, Precision, Recall. By comparing the accuracy of both the algorithms, SVM is more accurate than Naïve Bayes. By using the ROC curve, SVM is more efficient when compared to Naïve Bayes.

References

1. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS, Batiment 508, F-91405 Orsay Cedex, France
2. Sentiment Analysis of Twitter Data by Sahar A. El_Rahman Computer and Information Sciences College Princess Nourah Bint Abdulrahman University R
3. Shelar, A., Huang, C.: Sentiment analysis of twitter data. School of Computer Science, Kean University
4. Karan, R.S.: Sentiment analysis on twitter data: a new approach. Computer Engineering, Viva Institute of Technology, Mumbai, India
5. Windasari, I.P., Uzzi, F.N., Satoto, K.I.: Sentiment analysis on twitter posts: an analysis of positive or negative opinion. Department of Computer Engineering Faculty of Engineering – Diponegoro University Semarang, Indonesia
6. Parveen, H.: Sentiment analysis on twitter dataset using Naive Bayes algorithm. Department of Computer Science and Engineering Rungta College of Engineering and Technology Bhilai, India
7. Çeliktuğ, M.F.: Twitter sentiment analysis, 3-way classification: positive, negative or neutral? Computer Engineering Bilkent University, Gazi University Ankara, Turkey
8. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of Twitter data. In: Workshop on Languages in social media, Portland, Oregon, pp. 30–38 (2011)
9. Mandloi, L.: Twitter sentiments analysis using machine learning methods. Tech Computer Science and Engineering Meidcaps University
10. Dhawan, S.: Sentiment analysis of twitter data in online social network. Department of Computer Science and Engineering University Institute of Engineering and Technology (UIET), Kurukshetra University, Kurukshetra-136119 Kurukshetra, India
11. Wang, H., Can, D., Kazemzadeh, A., Bar, F., Narayanan, S.: A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In: ACL 2012 System Demonstrations, Jeju Island, Korea, pp. 115–120 (2012)

12. Neethu, M., Rajasree, R.: Sentiment analysis in Twitter using machine learning techniques. In: 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (2013)
13. Ibrahim, M.N.M.: Twitter sentiment classification using Naïve Bayes based on trainer perception. College of Information Technology University Tenaga Nasional Putrajaya, Malaysia
14. Analyze real-time Twitter sentiment with HBase. Docs.microsoft.com (2017). https://docs.microsoft.com/enus/azure/hdinsight/hdinsight-hbaseanalyze-twitter-sentiment. Accessed 20 Jan 2018
15. Lak, P., Turetken, O.: Star ratings versus sentiment analysis - a comparison of explicit and implicit measures of opinions. In: 7th Hawaii International Conference on System Science, ParisaLak (2014)
16. Geetha, R., Rekha, P., Karthika, S.: Twitter opinion mining and boosting using sentiment analysis. Department of Information Technology, SSN College of Engineering, Kalavakkam, Chennai
17. Bhat, S., Garg, S., Poornalatha, G.: Assigning sentiment score for twitter tweets. Department of Information and Communication Technology Manipal Institute of Technology, Manipal Academy of Higher Education Manipal, Karnataka
18. Comparison of Naive Bayes Smoothing Methods for Twitter Sentiment Analysis. Faculty of Mathematics and Natural Science Universitas Lambung Mangkurat Banjarbaru, Indonesia
19. Patankar, A.J.: Emotweet: a sentiment analysis tool for Twitter. Computer Sci. and Eng. Visvesvaraya Technological University, Belagavi, Karnataka