IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SENTIMENT ANALYSIS ON TWITTER TWEETS USING MACHINE LEARNING ALGORITHMS

N. Anuhya[1], D. *Lalitha Bhaskari*[2]

N. *Anuhya, M.Tech, Department of CSSE, AU College of Engineering (Autonomous), Andhra University, Visakhapatnam, India*
D. Lalitha Bhaskari, Professor, Department of CSSE, AU College of Engineering (Autonomous), Andhra University, Visakhapatnam, India.

***Abstract*---** Growth in the area of sentiment analysis has been rapid and aims to explore the opinions or text present on different platforms of social media through machine-learning techniques with sentiment. on daily basis nearly 326 millions of active users are there on twitter platform. This data can belong to any domain. Currently semantic approach is followed, which is based on Lexicons and bag of words technique are used. Later came Machine learning which had supervised, and unsupervised approaches followed by Deep Learning with Artificial Neural Networks. This work focused on development and differentiation between Naive Bayes and Random Forest by Machine learning on a large set of tweets each from the fields of Movies, Sports and Politics respectively. Both Positive and Negative (did not consider neutral case)are considered as class labels and a conclusion is drawn to finalize which algorithm works well for Sentiment Analysis.

***Keywords* ---** Twitter, Naive Bayes, Machine Learning, Random Forest, Ensemble Learning.

I. **Introduction**

In recent years, a huge number of people have been attracted to social-networking platforms like Facebook, Twitter and Instagram. Most of the people are using social sites to express their emotions, beliefs or opinions about things, places or personalities. Based on Omnicare agency survey nearly 500 million tweets have been generated on a single day in 2020. It has the majority of data is in the form of textual representation unlike Facebook and YouTube where Images and Videos are present respectively. The Methods of sentiment analysis can be categorized predominantly [1] as machine-learning [2], Lexicon-based [3] and hybrid [4,5]. Similarly, another categorization has been presented [6] with the categories of statistical, knowledge-based and hybrid approaches. There is a space for performing challenging research in broad areas by computationally analyzing opinions and sentiments [7]. Therefore, a gradual practice has grown to extract the information from data available on social networks for the prediction of Twitter #tags, to know the accuracy of sentiment analysis and predictions can be obtained by behavioral analysis based on social networks [8].
Data was collected from the public accounts of Twitter. Opinion Lexicon [9] was used to find a total number of positive and negative tweets. Firstly, the classifier is trained with already classified data, and then tested on data that is to be classified. This dataset was tested by supervised machine-learning algorithms for naive bayes and random forest. It also uses Bayes theorem in the background and there has been many variations provided nowadays.

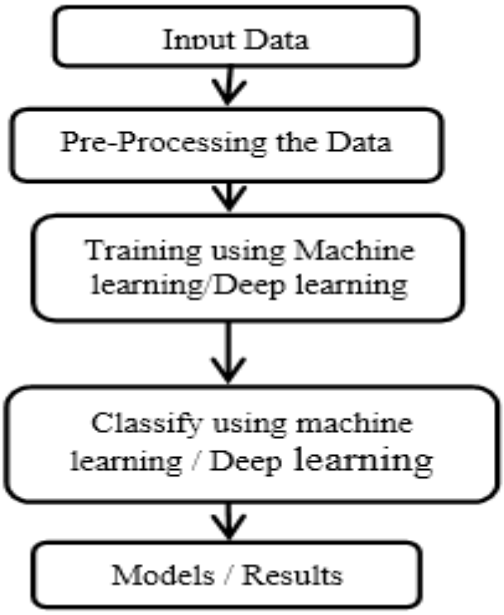

Figure 1: Process of Sentiment Analysis

Random Forests are part of Ensemble Learning which operate mainly by constructing and using Decision Trees which play a vital role in Machine Learning. The reason that sometimes Ensemble Learning is preferred over traditional Machine Learning is due to the possibility of obtaining better predictive performance than that of a single algorithm. Due to many advantages this algorithm is often used in many Scientific Works [10].

II. Literature Review

Opinion Mining and Sentiment Analysis on a Twitter Data Stream[11] has divided the work into two parts. At first the classification was done on the preprocessed tweets as neutral, polar and irrelevant. Then the polar tweets are subdivided into either positive or negative. This two-step division has provided great results as the classifiers used needn't worry about neutral data and the accuracy was also improved. Here the employed Machine learning classifiers were Naive Bayes and Random Forest. The highest accuracy was found for Bayesian models followed by Random Forest.

In recent years, a considerable amount of research has been carried out in the area of sentimental analysis. In[12], authors have proposed a technique for classifying student data generated on Twitter into different categories to face the different problems of students. In[13], the authors presented the logical approach to analyzing the feelings posted on various social media platforms. They analyzed the feelings of the text using combined categorical grammar, annotation, the acquisition of lexicons and the semantic network. Simple sentiment classification techniques and data collection methods are discussed in[14]. For the domain of electronic products, the accuracy of the classification procedure with selected feature vectors is verified with different classifiers, such as Naive Bayes, Support Vector Machine and so on[15]. In[16] authors presented a hybrid approach that combines the use of lexicons with a learning classifier of machines to detect the polarity of subjective texts within the context of consumer products. In[17] the authors proposed an array of machine-learning methods with semantic analysis to categorize the sentence and reviews of different products using WordNet for better accuracy based on twitter info. In[18]concentrates on implementation and comparison between Naive Bayes, Random Forest from Machine learning and Convolutional Neural Networks from Deep Learning applied on a large set of tweets each from the fields of Movies, Sports and Politics, Technology and Stock exchange respectively. The class labels considered are Positive and Negative and a conclusion is drawn to decide which algorithm works well for Sentiment Analysis.

III.Methodology

A suitable model is preferred to evaluate the choosen type of data. In practice, it is always recommended to compare different classification models on the particular dataset and consider the prediction performances as well as computational efficiency.

3.1. Naive Bayes

Naive Bayes classifiers is recapitulate the concept of Bayes' rule. The probability model that was formulated by Thomas Bayes (1701-1761) is quite simple yet powerful, it can be written down in simple words as follows:

*posterior probability = ((conditional probability * prior probability) / evidence)*

Bayes' theorem forms the core of the whole concept of naive Bayes classification. The posterior probability, in the context of a classification problem, can be interpreted as: "What is the probability that a particular object belongs to class I given its observed feature values?"

Let, xi be the feature vector of sample i, i∈ {1,2,...,n}, ωj be the notation of class j,j∈ {1,2,...,m}, and P(xi| ωj) be the probability of observing sample xi given that is belongs to class ωj. The general notation of the posterior probability can be written as

$$P(\omega j|\ xi) = P(xi|\ \omega j).P(\omega j)P(xi)$$

The objective function in the naive Bayes probability is to maximize the posterior probability given the training data in order to formulate the decision rule. Tokenization describes the general process of breaking down a text corpus into individual elements that serve as input for various natural language processing algorithms. Usually, tokenization is accompanied by other optional processing steps, such as the removal of stop words and punctuation characters, stemming or lemmatizing and the construction of n-grams.

Algorithm

Training Phase:

Input: Set of training tweets along with their labels
Output: A Naive Bayes Model for prediction

Step 1: Clean the tweets, i.e. remove unnecessary stop words, retweets, special characters etc.
Step 2: Construct the Vocabulary from all the cleaned tweets.
Step 3: Vectorize the tweets based on the Vocabulary constructed in step-2.
Step 4: Calculate the Maximum Likelihood Hypothesis for the given vocabulary and the tweets.
Step 5 : Adjust the model according to the training labels.

Testing Phase:

Input: Set of testing tweets, Naive Bayes Model
Output: Predicted class labels for the testing tweets

Step 1: Clean the tweets, i.e. remove unnecessary stop words, netweets, special characters, etc.
Step 2: Use the Vocabulary built in the "Training Phase" to vectorize the tweets.
Step 3: Feed the tweets to the constructed Naive Bayes model to obtain the predicted class label
Step 4: Test for the accuracy of the system.

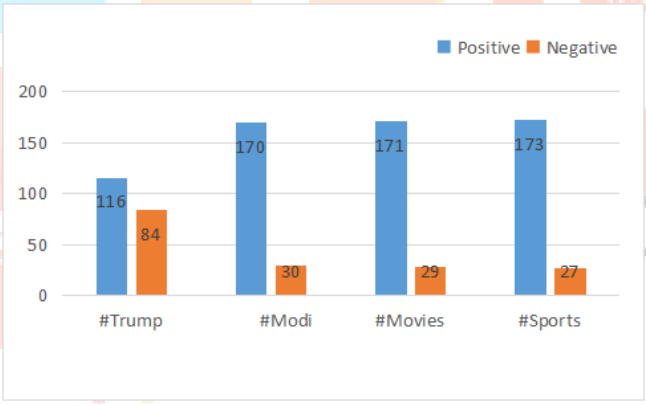

Figure 2: Bar Chart for Naïve Bayes Classification on Twitter Data of #Trump, #Modi,#Movies, #Sports

Hash Tag	Positive	Negative
#Trump	116	84
#Modi	170	30
#Movies	171	29
#Sports	173	27

Table 1: Positive and Negative score for Naïve Bayes classification on Twitter data of #Trump, #Modi, #Movies, #Sports

3.2. Random Forest

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set *X* = *xi*, ..., *xn* with responses *Y* = *yi*, ..., *yn*, bagging repeatedly (*B* times) selects a random sample with replacement of the training set and fits trees to these samples:
For *b* = 1, ..., *B*:

1. Sample, with replacement, *n* training examples from *X*, *Y*; call these *Xb*, *Yb*.
2. Train a classification or regression tree *fb* on *Xb*, *Yb*.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}.$$

or by taking the majority vote in the case of classification trees.

This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single training set would give strongly correlated trees (or even the same tree many times, if the training algorithm is deterministic). Bootstrap sampling is a way of de-correlating the trees by showing them different training sets.

Algorithm

Training Phase:

Input: Set of training tweets along with their labels
Output: A Random Forest Model for prediction

Step 1: Clean the tweets, i.e. remove unnecessary stop words, retweets, special characters, etc.
Step 2: Construct the Vocabulary from all the cleaned tweets.
Step 3: Vectorize the tweets based on the Vocabulary constructed in step-2.
Step 4: Randomly select "k" features from total "m" features, where k << m
Step 5: Among the "k" features, calculate the node "d" using the best split point.
Step 6: Split the node into daughter nodes using the best split.
Step 7: Repeat 4 to 6 steps until "l" number of nodes has been reached.
Step 8: Build forest by repeating steps 4 to 7 for "n" number times to create "n" number of trees.
Step 9: Adjust the model according to the training labels.

Testing Phase:

Input: Set of testing tweets, Random Forest Model
Output: Predicted class labels for the testing tweets

Step 1: Clean the tweets, i.e. remove unnecessary stop words, retweets, special characters, etc.
Step 2: Use the Vocabulary built in the "Training Phase" to vectorize the tweets.
Step 3: Feed the tweets to the constructed Random Forest model to obtain the predicted class labels.
Step 4: Test for the accuracy of the system.

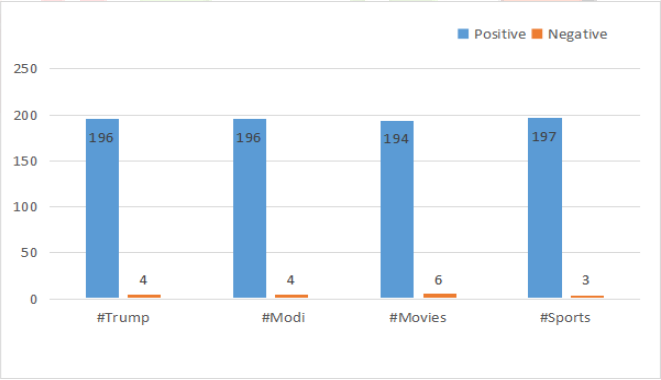

Figure 3: Bar Chart for Random Forest on Twitter Data of #Trump, #Modi, #Movies, #Sports

Hash Tag	Positive	Negative
#Trump	196	4
#Modi	196	4
#Movies	194	6
#Sports	197	3

Table 2: Positive and Negative score for Random Forest classification on Twitter data of #Trump, #Modi, #Movies, Sports

IV.Experimental Results

Here we choose 200 tweets from the domains of Trump, Modi, Movies, Sports to differentiate the results. The class labels we have divided into are 'Positive' and 'Negative'. The classifiers employed were Naive Bayes and Random Forest from Machine learning

Classifier used	Naive Bayes		Random Forest	
HashTags	Positive(%)	Negative(%)	Positive(%)	Negative(%)
#Trump	58	42	98	2
#Modi	85	15	98	2
#Movies	85.5	14.5	97	3
#Sports	86.5	13.5	98.5	1.5

Table 3: Positive and Negative score for Naïve Bayes and Random Forest classification on Twitter data of #Trump, #Modi,#Movies, #Sports

From the above table, we can observe that the variation between the performance of Naive Bayes and Random Forest.Random Forest is having high accuracy compare to Naive Bayes.

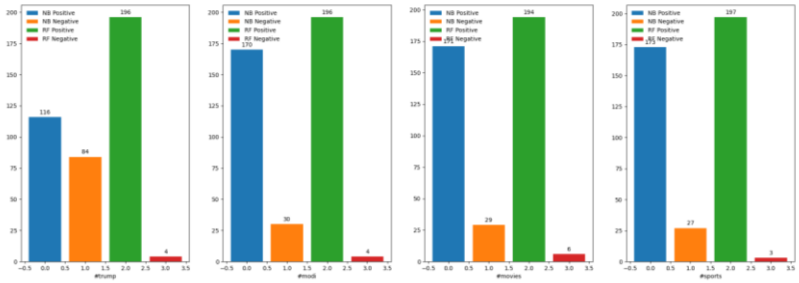

Figure 4: Bar Chart for Naive Bayes and Random Forest on Twitter Data of # Trump, # Modi, #Movies and #Sports

V. Conclusion

In the first case, we presented in this paper the comprehensive procedure for performing sentiment analyzes to classify highly unstructured Twitter data as positive or negative. Secondly, we discussed the Naïve Bayes & Randam Forest algorithms that can be used to perform Twitter data sentimental analysis. The future opportunities in the field of sentiment analysis, therefore, include the development of a methodology for the classification of sentiment that can be applied to any data irrespective of the domain.

References

1. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. Ain Shams Eng. J. **2014**, 5, 1093–1113.
2. Sebastiani, F. Machine learning in automated text categorization. ACM Comput. Surv. **2002**, 34, 1–47.
3. Taboada, M.; Brooke, J.; Tofifiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis. Comput. Linguist. **2011**, 37, 267–307.
4. Prabowo, R.; Thelwall, M. Sentiment analysis: A combined approach. J. Informetr. **2009**, 3, 143–157.
5. Dang, Y.; Zhang, Y.; Chen, H. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. IEEE Intell. Syst. **2010**, 25, 46–53.
6. Cambria, E. Affective computing and sentiment analysis. IEEE Intell. Syst. **2016**, 31, 102–107.
7. Jagdale, O.; Harmalkar, V.; Chavan, S.; Sharma, N. Twitter mining using R. Int. J. Eng. Res. Adv. Tech. **2017**, 3, 252–256.
8. Anjaria, M.; Guddeti, R.M.R. Inflfluence factor based opinion mining of twitter data using supervised learning. In Proceedings of the 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), Bangalore, India, 6–10 January 2014; pp. 1–8.
9. Liu, B.; Hu, M.; Cheng, J. Opinion observer: Analyzing and comparing opinions on the web. In Proceedings of the 14th International Conference on World Wide Web, Chiba, Japan, 10–14 May 2005; pp. 342–351
10. Hasan, Mehedi et al. "A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories." Journal of biomedical informatics 62 (2016): 21- 31.
11. Gokulakrishnan, Balakrishnan, et al. "Opinion mining and sentiment analysis on a twitter data stream." Advances in ICT for emerging regions (ICTer), 2012 International Conference on. IEEE, 2012.
12. Gupta, Ankita, JyotikaPruthi, and NehaSahu. "Sentiment Analysis of Tweets using Machine Learning Approach." International journal of computer science and mobile computing (2017).
13. N. Kasture and P. Bhilare, "An Approach for Sentiment analysis on social networking sites", Computing Communication Control and Automation (ICCUBEA), 2015, pp. 390-395.
14. S. Bhuta, A. Doshi, U. Doshi and M. Narvekar, "A review of techniques for sentiment analysis Of Twitter data", Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014, pp. 583-591.
15. Goel, Ankur, Jyoti Gautam, and Sitesh Kumar. "Real time sentiment analysis of tweets using Naive Bayes." Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on. IEEE, 2016.

16. S. Bahrainian and A. Dangel, “Sentiment Analysis using Sentiment Features”, in Int. joint Conf. of Web Intelligence and Intelligent Agent Technologies, 2013, pp. 26-29.
17. G. Gautam and D. Yadav, “Sentiment analysis of twitter data using machine learning approaches and semantic analysis”, in 7th Int. Conf. on Contemporary Computing, 2014, pp. 437-442.
18. Rayala vinodkumar,D.Lalitha Bhaskar,P.Srinivasarao “Comparison of Sentiment Analysis on Various Twitter #Tags Using Machine Learning and Deep Learning Techniques” Jour of Adv Research in Dynamical & Control Systems, Vol. 11, 04-Special Issue, 2019