INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-1818)
Available online at: https://www.ijariit.com

Sentiment Analysis of Tweets using Machine Learning

Disha C. Kini
mca18_917ics@met.edu
MET Institute of Computer Science, Mumbai, Maharashtra

Harshali Patil
harshalip_ics@met.edu
MET Institute of Computer Science, Mumbai, Maharashtra

ABSTRACT

The growth of various social networking sites has enabled people to express their opinions, thoughts, and feelings. Social networking platforms such as Twitter, Facebook, etc. are filled with opinions and thoughts. The enormous data gathered using these platforms can help to analyze the sentiment of people. Sentiment analysis is the method to interpret and classify emotions into categories viz positive, and negative.. Data mining is harnessed to uncover relevant information from web pages. The main emphasis of this research is to classify which machine learning algorithm gives better accuracy. Python language is used to implement the proposed system. Machine learning classifiers such as Naïve Bayes, SVM, and Decision Tree are applied to categorize the tweets as positive or negative.

Keywords--- *Twitter, Sentiment Analysis, Machine Learning, Naïve Bayes, SVM, Decision Tree*

1. INTRODUCTION

Twitter has acquired a lot of popularity over the years. It can be used by anybody to post tweets about any event. It is a stage where individuals can state their viewpoints, contemplations, or feelings. It contains an enormous measure of information. The length of any tweet can't be beyond 280 characters, making the data helpful. The number of tweets on Twitter is more than five hundred million every day, which is enormous data for sentiment analysis. Twitter users range from daily users to celebrities, from business executives to political figures. So, twitter reflects the views of all groups.

Sentiment analysis or opinion mining is the analysis of opinions or feelings from the text information. It identifies the opinion or sentiment of each person concerning a specific event.

Sentiment analysis is applied to discover the polarity of text viz positive, negative, or neutral. It helps us in deciding if a particular item or a service is good or bad. Manufacturers or developers of the products can review their products by analyzing the sentiments of the people, that is, whether people are liking their product or not. Marketing personnel can see how people are reacting to their advertising campaign. They can analyze the sentiments related to this. Political parties can see how their political campaign is running. Filmmakers can see how people are reacting to their newly released movie by analyzing the sentiments of the people [5]. Sentiment analysis utilizes Natural Language Processing to assess an individual's state of mind and feelings through the piece of text. A mainstream use for this technology comes from its implementation in the social media field to explore how individuals feel about specific topics, explicitly through the word-of-mouth of social media users in text posts or their tweets with regards to Twitter. Here, we will use different machine learning algorithms to analyze the sentiments of the people. Machine learning methods like Naïve Bayes Classifier, Support Vector Machine method and Decision Tree Classifier are used. We will compare these methods based on their accuracy and see which classifier gives the best result.

2. LITERATURE REVIEW

The author uses the Naive Bayes algorithm to analyze the sentiment of tweets. Six different sentiments are analyzed using the sentiment package which are joy, anger, fear, disgust, sadness, and surprise. The polarity of text is also found. Polarity can be positive, negative, or neutral. New words were identified using word clouds, and then polarity was assigned to them [1].

SVM, Adaboosted Decision Tree, and Decision Tree based composite sentiment categorization model is presented for improving the overall accuracy of the classifier in the classification of tweets. The proposed approach categorizes tweets as positive or negative. For analytical evaluation of the proposed classifier, accuracy and f-measure are used. The comparative observations taken against the SVM, Adaboosted Decision Tree, and Decision Tree proves that the hybrid model improved the overall classification accuracy and f-measure of sentiment prediction as compared to existing techniques for classification [2].

Tweets were collected and passed through machine learning classifiers. A voted classification mechanism was used to obtain the class of tweets. RNN, LSTM, and CNN-RNN models were applied to categorize the tweets. Deep Learning models and their several compositions have shown better performance compared to machine learning algorithms [3].

The author has made a comparison between SVM, Naive Bayes, and Maximum Entropy regarding sentence-level sentiment analysis for estimation of depression. The voting model and feature selection technique are used. The performance was examined on two datasets, viz twitter dataset, and 20newsgroups. The experiment reveals that SVM shows the superior result as compared to Naive Bayes and ME classifiers [4].

The different machine learning techniques of data analysis of Twitter, like Naïve Bayes, SVM, and Maximum Entropy Method are compared. The analysis of Twitter data is being done in various aspects to mine the sentiments. The study defines the concept of opinion in the analysis of sentiment of Twitter. The machine learning method such as Naïve Bayes has the highest accuracy and can be considered as the baseline learning method as well as in some cases Maximum Entropy methods are very effective [5].
Tweets gathered from Twitter are preprocessed utilizing Natural Language Toolkit procedures. The features of the tweets are chosen dependent on the Chi-Square test and the Naïve Bayes classifier is utilized to characterize the tweets as positive and negative. It is observed that if the quantity of features increases, the accuracy of the selected features also increases [6].

3. MACHINE LEARNING METHODS FOR SENTIMENT ANALYSIS

3.1 Naïve Bayes Classifier

Naive Bayes is a machine learning algorithm which utilizes the Bayes theorem for classification. It is an easy and most powerful classification algorithm that assists in making predictions quickly. It utilizes the accompanying Bayes theorem:

$$Prob(A/B) = (Prob(B/A) * Prob(A))/Prob(B)$$

where A and B, are some events, and P(.) is a probability.

Prob(A|B) is known as (posterior probability) of hypothesis A which occurs when some condition is already provided.

Prob(B|A) is also called likelihood probability. It is the probability of evidence E when we presume that the given hypothesis is true [5].

Multinomial Naive Bayes represent the features of the model as frequencies of their occurrences. It can be utilized to calculate maximum likelihood estimates based on the training data to estimate the conditional probability.

3.2 Support Vector Machine

Support Vector Machines employ the technique from computational learning theory, which aims at reducing the structural risk. Finding the decision boundary which increases the distance between two classes is the basic principle of SVM [2]. The vectors that define this decision boundary are known as support vectors. SVM is a machine learning method, utilized for both classifications in addition to problems of regression. Classification and regression both are a subcategory of supervised machine learning. Classification can be defined as predicting a label. So, the principal job of the Support Vector Machine classifier is to perform classification. That is, it classifies the data in different classes by drawing a hyperplane which differentiate between different classes which we plot in n-dimensional space. The fig.1 depicts the support vector machine classification.

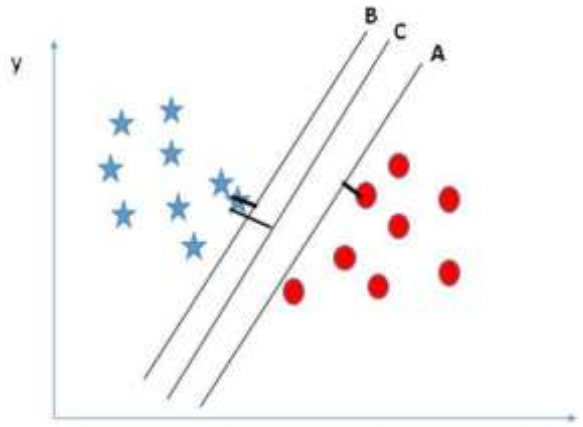

Fig. 1: Support Vector Machine

3.3 Decision Tree Classifier

Decision Tree is an ML method that can be utilized for classification and problems of regression. The structure of this classifier is like a tree. A decision tree is used to represent alternatives and their subsequent results in the form of graphs. The internal nodes represent the rules about decisions and the leaf node addresses the result. Decision nodes are utilized to form any conclusion and have numerous branches. Leaf nodes are the outcome to these decisions and do not consist of any further branches. The decisions are carried out based on the traits of the dataset given. A decision tree can contain categorical data (yes/no) as well as numeric data. Test data can be classified quickly owing to the decision tree algorithm.

4. IMPLEMENTATION

Sentiment Analysis focuses on textual data. The data needs to be filtered, i.e., processed and then analyzed. For performing the research, the dataset available on Kaggle [7] was selected. The dataset consists of 14000+ tweets labeled as positive, negative, and neutral. The dataset consists of more negative data compared to positive data, which makes the dataset imbalanced.

The dataset is preprocessed. First, the punctuation marks are removed, then using regular expression, URLs are removed. These texts are converted to lowercase. Stop words like (a, an, the, etc.) which do not play any significant role in sentiment analysis are removed.

The dataset is split into two sets viz training set and testing set. The training set consists of 80% of the data, whereas the testing set consists of 20%. The dataset is balanced using SMOTE, which creates new synthetic training examples based on the minor class.
For training the classifier, it is needed to change the words to numerical value, because algorithms work just with numbers. The words are vectorized into numbers using CountVectorizer. A vocabulary is made from the given dataset, and every word is allocated 0 or 1 if it is present in the current sentence. Vectorization is performed only on the training set.
The three machine learning algorithms discussed above were applied on the training set. Accuracy of each method was calculated and the confusion matrix was obtained.

5. RESULT

The Accuracy and Confusion Matrix of different Machine Learning methods are:

Accuracy = ((True Positive + True Negative) * 100)/ Total

5.1 Naïve Bayes Classifier

Table 1: Confusion Matrix for Naïve Bayes Classification

	Positive	**Negative**	**Total**
Positive	345	98	443
Negative	123	1743	1866
Total	468	1841	2309

Based on the above table 1, accuracy can be calculated as follows

Accuracy = ((345 + 1743) *100)/2309

= 90.42 %

5.2 Support Vector Machine

Table 2: Confusion Matrix for Support Vector Machine

	Positive	**Negative**	**Total**
Positive	372	168	540
Negative	96	1673	1769
Total	468	1841	2309

Based on the above table 2, accuracy can be calculated as follow

Accuracy = ((372 + 1673) *100)/2309

= 88.57 %

5.3 Decision Tree Classifier

Table 3: Confusion Matrix for Decision Tree Classifier:

	Positive	**Negative**	**Total**
Positive	290	228	518
Negative	178	1613	1791
Total	468	1841	2309

Based on the above table 3, accuracy can be calculated as follows.

Accuracy = ((290 + 1613) *100)/2309

= 82.42 %

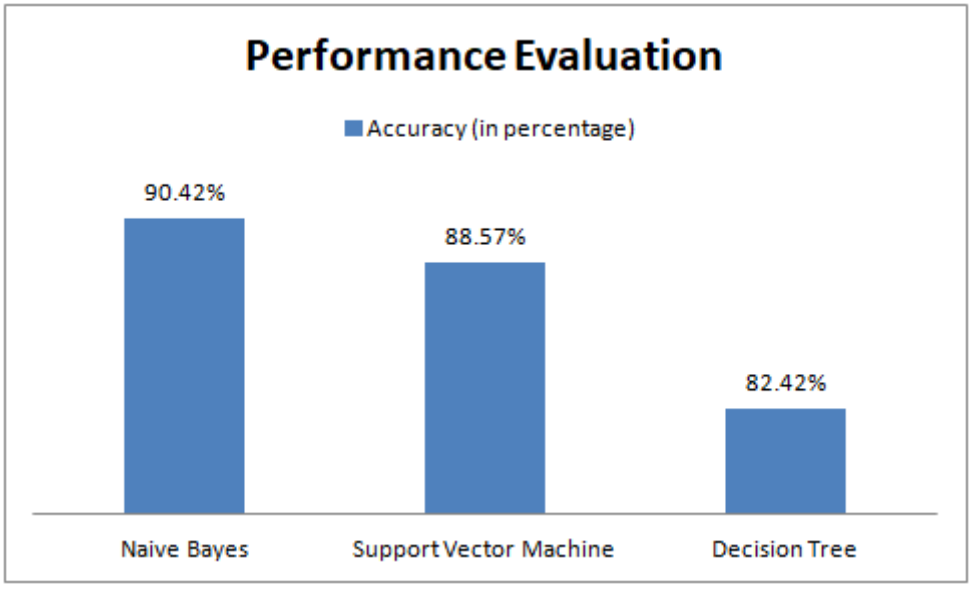

Fig. 2: Performance Evaluation and comparison

After examining the performance on the dataset collected from Kaggle [7] consisting of more than 14000 tweets, experiment suggests that Naïve Bayes had accuracy of 90.42%, SVM had accuracy of 88.57% and Decision Tree classifier had accuracy of 82.42%. The performance evaluation of classifiers used is shown in above fig. 2.

6. CONCLUSION

Looking at past 10 years data, it is observed that the number of users and their interactions on social media are increasing exponentially. People find it easy to express their opinions on such platforms. Twitter is one of the well known social media platforms. Secondary data is used for the experiment and more than 14000+ tweets have been examined. To classify the tweets into classes, namely positive, negative, three popular methods used.

A comparative analysis between three machine learning algorithms viz Naïve Bayes, SVM, and Decision Tree is made. The experiment results suggest that Naïve Bayes gives better results compared to SVM and Decision Tree classifiers Naive Bayes is one of the most widely used methods for classification problems and in sentiment analysis problems, it generates outcome on the basis of data with very high accuracy.

7. REFERENCES

[1] Sentiment Analysis: Machine Learning Approach by Dipak R. Kawade, Dr. Kavita S. Oza, International Journal of Engineering and Technology (IJET) - June 2017.

[2] Sentiment Analysis of Tweets using Machine Learning Approach by Megha Rathi, Aditya Malik, Daksh Varshney, Rachita Sharma, Sarthak Mendiratta Jaypee Institute of Information Technology JIIT Sec-62 Noida, Proceedings of 2018 Eleventh International Conference on Contemporary Computing (IC3), 2-4 August, 2018, Noida, India.

[3] Sentiment Analysis using Machine Learning and Deep Learning by Yogesh Chandra and Antoreep Jana, 7th International Conference on Computing For Sustainable Global Development (INDIACom).

[4] Sentiment Analysis of Social Networking Sites (SNS) Data using Machine Learning Approach for the Measurement of Depression by Anees Ul Hassan, Jamil Hussain, Musarrat Hussain, Muhammad Sadiq, Sungyoung Lee, Department of Computer Science and Engineering Kyung Hee University, Suwon, South Korea.

[5] Twitter Sentiments Analysis Using Machine Learning Methods by Lokesh Mandloi and Ruchi Patel, Medicaps University Indore, 2020 International Conference for Emerging Technology (INCET) Belgaum, India. Jun 5-7, 2020 .

[6] Sentimental Analysis of Tweets Using Naive Bayes Algorithm by M. Vadivukarassi, N. Puviarasan and P. Aruna, Annamalai University, Tamil Nadu, India, World Applied Sciences Journal 35 (1): 54-59, 2017 .

[7] https://www.kaggle.com/crowdflower/twitter-airline-sentiment

