# Twitter Sentiment Analysis using Machine Learning Techniques

**K. Sentamilselvan, D. Aneri, A. C. Athithiya, P. Kani Kumar**

***Abstract:*** ***Nowadays people share their views and opinions in twitter and other social media platforms, the way of recognizing sentiments and speculation in tweets is Twitter Sentiment Analysis. Determining the contradiction or sentiment of the tweets and then listing them into positive, negative and neutral tweets is the main classifying step in this process. The issue related to sentiment analysis is the naming of the correct congruous sentiment classifier algorithm to list the tweets. The foundation classifier techniques like Logistic regression, Naive Bayes classifier, Random Forest and SVMs are normally used. In this paper, the Naïve Bayes classifier and Logistic Regression has been used to perform sentiment analysis and classify based on the better accuracy of catagorizing Technique. The outcome shows that Naive Bayes classifier works better for this approach. Data pre-processing and feature extraction is realized as a portion of task.***

***Keywords*****:** ***Feature extraction, Logistic regression, multinomial Naïve Bayes, Sentiment Analysis.***

## I. INTRODUCTION

One of the fastest appearing platform to send or receive a message or tweet to a wide range of people is Twitter. Users normally communicate or express their views and thoughts about a concerned subject through tweets. Blogging websites nowadays are becoming common place to express one's opinion which helps the marketing campaigns to share consumer's opinions on different topics concerning brands and products. Researchers also make use of online data in order to Sentiment Analyze the opinion of public on an item or subject.

Normally, Sentiment Analysis is a text which identifies and extract the subjective information. Sentiment analysis is the most common text classification tool that analyses concept and tells whether the underlying sentiment is 0, 1 or -1.

Challenges in Twitter sentiment analyses is (i) some of the tweets are generally noted in an unceremonious languages, and also some message that are short, show limited sign about sentiment. (ii) Sometimes hashtags, URL, abbreviations, emoji's and acronyms are widely used on twitter.

Primarily, foundation classification techniques like Naive Bayes, SVMs and Logistic Regression is used to resolve the Twitter classification problem. Concepts including pre-processing tweets, feature extraction methods, and also developing table using the accuracy of different machine leaning algorithmic techniques.

Firstly, studies related to the investigation of twitter data were discussed. Secondly, methods to read dataset, pre-processing of data and feature extraction, and procedure to balance and finding accuracy has been described. Thirdly, the results for this study are presented. Fourthly, conclusion and further task is provided.

## II. RELATED WORK

According to the author, an ensembles classification has been used to upgrade the correctness of tweet sentimental classification. Companies detect the consumer's interest to choose products and brands based on their opinions. It is a type of blog where the user can create, post, and update as well as read the short messages. [1]

Deep convolution algorithm is used for better performance in twitter analysis. The survey is taken on public reviews about the related product and events. Word embedding method institute individual learning based on twitter. Twitter sentiment is well performed using pre-trained word vector. [3]

[4] Deals with Tweets to polls where sentiments are measured from text based on public opinion. Advanced NLP techniques is useful to enhance opinion estimation. Textual sentiment in tweet message is measured through time, comparing to contemporaneous polling data.

Author proposed Naive Bayes classification algorithm for both sarcastic and non-sarcastic tweets is the form of expressing negative feeling using positive words. In the existing system, logistic regression technique is used to detect sarcasm in tweets, it has some drawbacks which cannot predict continuous variables. [6]

[8] In this paper collection of written texts and dictionary based methods are used to determine both positive and negative words in tweets. The work presented in paper specifies a new techniques for sentiment analysis on twitter data. The overall tweet sentiment was calculated using a linear equation.

Several steps were taken for sentimental analysis on twitter data using machine learning algorithm. Tweets are pre-processed using NLP based techniques. They build the model using Support Vector Machine (SVM), Naive Bayes classifiers, and machine learning classifiers. The outcome shows that decision tree performs effectively showing 100% accuracy. [11] Moreover, substantial work put forward a solution for sentimental analysis based.

**Revised Manuscript Received on February 29, 2020.**

***** Correspondence Author

**Mr.K.Sentamilselvan***, Assistant Professor, Department of Information Technology, Kongu Engineering College, Erode. Email: ksentamilselvan@gmail.com

**Ms.D.Aneri**, Final Year Student, Department of Information Technology, Kongu Engineering College, Erode. Email: aneriamin1109@gmail.com

**Ms.A.C.Athithiya**, Final Year Student, B.Tech IT, Kongu Engineering College, Erode. Email: ac.athithiya@gmail.com

**Mr.P.Kani Kumar**, Final Year Student, B.Tech Information Technology, Kongu Engineering College, Erode.

The researchers used data consisting of tweets with emoji's. Finally, model was trained using Maximum Entropy, Naive Bayes classifiers, Support Vector Machine (SVM). The outcome shows that SVM model was more effective than other models. [14]

Data analysis of twitter includes machine learning and lexicon-based approach. Comparatively, various research being done with sentiment. The study defines the concept of opinion in sentiment analysis in twitter. The result shows that machine learning method such as Naive Bayes and SVM have the highest accuracy, it improves the robustness and performance of twitter. [16]

Sentimental analysis is used for classifying the positive and negative opinion using Machine Learning Techniques with the help of SVM algorithm which shows maximum accuracy. Sentimental analysis is most effectively used for analyzing the people opinion. It is used for result comparison purpose. [18]

Tweet sentimental analysis has been undiscovered in the literature. Marketing the product which is useful for the customer to search the products and brand. Here, Automatic tool is used for classifier ensembles and lexicons. This analysis is used for classifying problems like opinions, attitudes, and emotions. [19]

## III. PROPOSED METHODOLOGY

The proposed methodology consists of four modules: (1) Data pre-processing: to pre-process the data (2) Feature Extraction (3) Machine learning classifiers.

### DATA PRE-PROCESSING:

Data pre-processing can be owned to reduce the size of the feature for machine learning algorithms. This concept is essential because the tweets contain several feature as discussed earlier.

- Punctuations of all kinds are eliminated
- URL is removed from tweets
- Emoji's or Emoticons are therefore restored by similar meaning since they appear as an important feature to detect sentiments.
- Stop-words that are useless words are removed from the tweet.
- Question-mark that are not necessary are removed
- Upper case alphabets are automatically converted to lower case.

**Example:**

| | |
|---|---|
| 12222 | glad rt bet bird wish flown south winter |
| 3936 | point upc code check baggag tell luggag vacat day tri swimsuit |
| 367 | vx jfk la dirti plane not standard |
| 12257 | tell mean work need estim time arriv pleas need lapt op work thank |
| 2957 | sure busi go els airlin travel name kathryn sotelo |
| 1166 | four schedul flight reserv liter not take one unreal |

### ALGORITHM1:PRE-PROCESSING TWEETS

Begin
Input Query string
Until the data is retrieved from Twitter streaming API, Do:
Filter English Language Tweets
Case conversion
For each tweet, do:
Procedure Pre-processing:
Remove Twitter symbols (#topic, @user name, retweet (RT))
Remove URL ("http://url")
Remove all symbols, number, Emoticons, and punctuations
Avoiding misspelling and slang words
Remove repeated letters
Remove Stop Words
Tokenization
Stemming
Return tweet
End Procedure

### ALGORITHM 2: FEATURE EXTRACTION

Procedure for Feature Extraction :
Withdraw features using (TF-IDF) with the suitable format
Extract features using (Count_Vectorizer) in machine learning techniques with the suitable format
Extract features using (Word_to_Vec) in machine learning techniques with the suitable format
End Procedure

### ALGORITHM 3: BALANCING AND SCORING:

Procedure Balancing and scoring features:
To calculate the polarity status of the tweets, balancing and finally labelling tweets
End Procedure

### ALGORITHM4: SENTIMENT CLASSIFICATION:

Procedure Sentiment Classification features:
Process to classify the tweet by using the machine learning algorithms (Multinomial Naïve Bayes and Logistic Regression)
End Procedure
End Until
End.

## IV. FEATURE EXTRACTION

Used to extract traits to develop a catagorizing model. The traits which are extracted will be in a style offerable to support cordinally to the machine learning techniques from datasets which contain primary data of different form like text, images or may be a train of symbols. Use of certain feature extraction techniques such as count vectorizer, TF-IDF (frequency inverse

document frequency) and word_to_vec to extract features from given tweets.

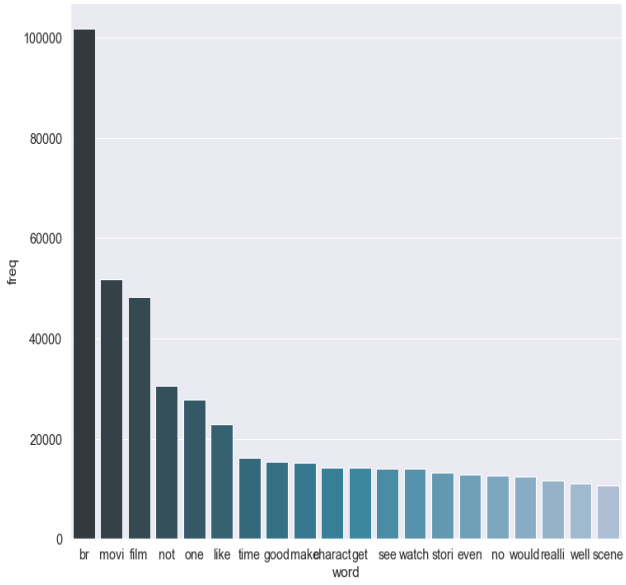
## V. TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF)

One of the statistical measure and a loading technique is TF-IDF. Common verbs, adjectives and nouns are been extracted from the processed dataset which are used to calculate the sentiment polarity such as positive, negative and neutral in a sentence in order to find out people's opinion on the desired concept by replica such as unigrams, bigrams, or n-grams.

To calculate the importance of a word to a document in a respected dataset we use TF-IDF approach where weight is assigned to each word in the document. It extracts the features based on the count of the words, by providing lesser weight to some frequent words and more weight to the rare words.

## VI. COUNT VECTORIZER

The count_vectorizer provides the simple way to build the vocabulary of words and also to tokenize the collection of text documents and encode the documents using that vocabulary. This can be done by creating an instance for the count vectorizer class then call the fit() function in order to learn the vocabulary from one or more documents then finally class the transform() on one or more documents as needed to encode each as a vector. Then this encoded vector will return the length of the vocabulary and an integer count to find the number of times the respected word appears in the document.




### WORD TO VECTOR

Word_to_vec is not a single algorithm but a combination of two techniques which are CBOW (continuous bag to words) and skip gram model. This helps to derive the relationship between a word and its contextual word. CBOW model takes the context of each word as input and it will try to predict or identify the word based on the context. In skip gram model we input the target word into the network and it will output the C probability distributions.

## VII. MACHINE LEARNING CLASSIFIERS

The Sentiment Analysis from the tweet can be detected using two basic approaches, one is lexicon based approach and another one is machine learning approaches. Use machine learning techniques of various forms for text catagorising and sentimental analysis of twitter data. These techniques are used to practice the algorithms by the task of involving the algorithm with the train data to the test data. Number of algorithms include Multinomial Naive Bayes and Logistic regression used to find the accuracy of the particular dataset.

### MULTINOMIAL NAIVE BAYES:

Naive Bayes is used to predict the tag of a text. These are commonly used in NLP algorithms. Used to calculate the prospect of each label for a given text and then it will output label with the highest one. Bayes theorem classify the probability p(c|x) where c represents the grade of feasible outcome and X is the instance used to present some common forms.

P (c|x) = P(x|c) * P(c) / P(x)
For example we consider the movie review dataset,

P (positive|overall liked the movie) = Prospect that the label of the stanza is positive.

P (negative|overall liked the movie) = Prospect that the label of the stanza is negative.

### LOGISTIC REGRESSION:

Regression model used for classification purpose is logistic regression. To relate a single categorical variable to one or more independent variable Logistic regression is used. A hyper-plane which is used to maximize the separating gap between classes is found by logistic regression. C = .01 and max_iter = 100 are the parameter values of the logistic regression classifier.

$$Y=e \wedge (b0 + b1*x) / (1+e \wedge (b0 + b1*x))$$

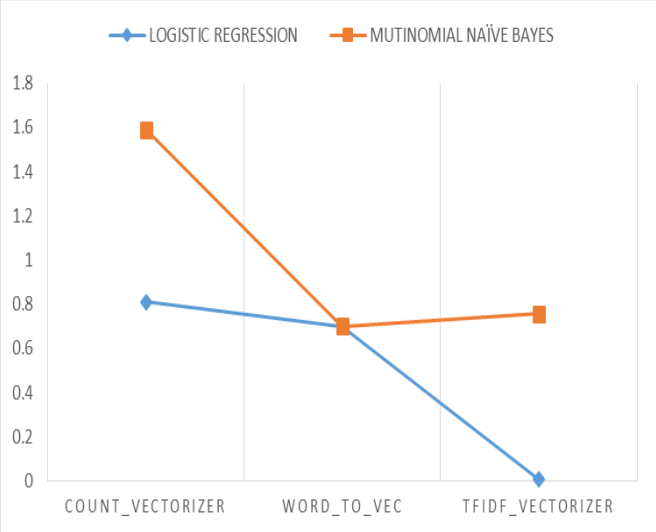
Here y is the output predicted, b0 is the bias and b1 is the single input coefficient value(x).

| METHODS | ALGORITHMS | |
|---|---|---|
| | LOGISTIC REGRESSION | MULTINOMIAL NAÏVE BAYES |
| COUNT_VECTORIZER | 88.4% | 87% |
| TFIDF_VECTORIZER | 88.1% | 83% |

## VIII. RESULT

**ACCURACY TABLE**

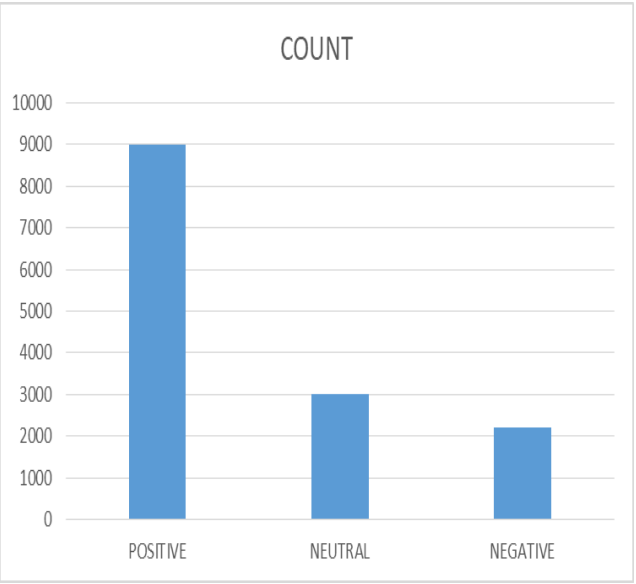
| METHOD | ALGORITHM |
|---|---|
| | LOGISTIC REGRESSION |
| WORD_2_VEC | 80% |




**Graph based on the accuracy table:**

### CALCULATION ON POSITIVE, NEGATIVE AND NEUTRAL TWEETS

The calculation on positive, negative and neutral tweets from the given dataset. Datasets considered here are airline sentiment dataset and IMDB review dataset.




## IX. CONCLUSION

In machine learning both classifiers achieve the best results when using the features of the Count_Vectorizer. Overall, Logistic Regression outperforms the Multinomial Naive Bayes classifier. The best performance on the test set comes from the Logistic Regression with features from Count_Vectorizer. This can be further implemented using the deep learning techniques in order to get the more accurate results.

## REFERENCE


1. Ankita, Nabizath Saleena "An ensemble classification system for twitter sentimental analysis",International Conference On Computational System for Twitter Sentiment Analysis (ICCIDS) Procedia Computer Science 132(2018)937-946
2. M.A.Cabanlit and K.J.Espinosa, "Optimizing N-gram based on text feature selection in sentiment analysis for commercial products in Twitter through polarity Lexicons",inprcc.5thInt.Conf.Inf.Intell.Syst.Appl.(IISA),Jul.2014,pp.94-97.
3. Zhao Jianqiang , Gui Xlaolin, and Zhang Xuejun "Deep convolution neutral network for twitter sentimental analysis" Jan 1,2018 ACCESS.2017.2776
4. Brendan o' Connor , Ramnath Balasubramaniyan , Bryan R. Routledge , Noah A. Smith " Linking Text Sentimental to public opinion time series" International Conference On Web and Social Media(ICWSM),2010,volume 11,nos.122-129,pp.1-2
5. M.Bouazizi and T.Ohtsuki, "A Pattern-based approach for multi-class sentiment analysis in Twitter ",IEEE Access, vol.5,pp.20617-20639,2017*
6. Bala Durga Dharmavarapu,Jayang Bayana "Sarcasm dection in twitter using sentimental analysis" International Journal of Recent Technology and Engineering(IJRTE) volume 8,issue -1 may 2019
7. R.Sara, R.Alan, N.Preslav, and S.Kirithchenko, S.Mohammad, A.Ritter, and V.Stoyanov,"SemEval-2016 task 4: Sentiment analysis in Twitter", in proc.8th Int.Workshop Semantic Eval., 2014, pp.1-18.
8. Akshi Kumar and Teeja Mary Sebastian "sentiment analysis on twitter" International Journal of computer
9. science Issues volume 9 Issue 4, No 3, july 2012
10. S.Rosenthal, P.Nakov, S. Kiritchenko, S.Mohammad, A.Ritter, and V.Stoyanov, "Semeval-2015 task 10: Sentiment analysis in Twitter",in Proc.9th Int .Workshop Semantic Eval.(SemEval), Jun.2015,pp.1-18
11. P.Nakov,A.Ritter,S.Rosenthal, F.Sebastiani, and V.Stoyanov, "SemEval-2016 task 4: Sentiment analysis in Twitter," in proc.10th Int.Work.Semant.Eval.,Jun.2016,pp1-18.
12. A.P. Jain and P. Dandannavar "Application of machine learning techniques to sentiment analysis" in proc.2ndInt.Conf.Appl.Theor.Comput.Commum.Technol.(iCATccT),Jul.2016,pp. 1-6.200
13. A.K.Jain, and R.P.W.Duin, and J.C.Mao,"Statistical pattern recognition: A review,"IEEE Trans.Pattern Anal. Mach.Intell., vol.22.no.1,pp.4-37, Jan.2000
14. I.H.Witten, E.Frank,M.A.Hall, and C.J.pal, Data Mining: PracticalMachine Learnimg Tools and Techniques.San Mateo,CA.USA:Morgn Kaufmann,2016
15. G. O, R. Bhayani, L. Huang " Twitter sentiment classification using distant supervision" volume 150 ,no.12,pp 1-6.2009
16. V.Cherkassky and F. M. Mulier, Learning From Data: Concepts, Theory,and Methods. Hoboken, NJ, USA: Wiley, 2007
17. Avinash Surnar, Sunil Sonawane "Review for twitter sentiment analysis using various method" International Journal Of Advanced Research In Computer Engineering And Technology(IJARCET) Volume 6,Issue 5, May 2017
18. P. A. Gutierrz, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervas-Martinez, "Ordinal regression methods: Survey and experimental study," IEEE Trans. Knowl. Data Eng., vol.28, no. 1,pp. 127-146, jan. 2016.
19. N. M. Dhanya and U.C Harish "Sentimental analysis on twitter data on demonetization using machine learning techniques" Springer International Publishing AG 2018 DOI:10.1007/978-3-319-71767-8_19
20. Nadia F.F de Silva Eduardo R. Hruschker, Estevam R. Hruschka. J. R "Tweet Sentiment analysis with classifier ensembles" Decision support Systems 66(2014)170-179
21. L. Li and H. Lin, "Ordinal regression by extended binary classification,"in Proc. Adv. Neural Inf. Process. Syst., 2007,pp. 865-872.

22. J. D. Rennie and N.Srebro, "Loss functions for preference levels: Regression with discrete ordered labels," IJCAI Work. Adv. Preference Handling,Jul. 2005,pp. 180-186.
23. Z.Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation, " in Proc. IEEE Conf. Comput.Vis Pattern Recognit.,Jun. 2016, pp. 4920-4928.

## AUTHORS PROFILE

**K.Sentamilselvan,** completed BE in Computer Science and Engineering from Anna University, Chennai, TN, India in 2010. Further completed M.Tech degree with specialization of Information Security from Pondicherry Engineering College, Pondicherry, India in 2013. And currently working an Assistant professor in the Department of Information Technology at Kongu Engineering College, Perundurai, Erode, TN, India. Life member of Computer Society of India (CSI). Research interest is Hacking, Web application Security, Information Security, Cloud and Grid computing.

**D.Aneri,** currently studying B.Tech Information Technology Final year at Kongu engineering college, Erode- referred many journal papers and finally got an idea to work with sentiment analysis and machine learning, a challenging task to do twitter sentiment analysis using machine learning classifiers.

**A.C.Athithiya,** currently studying B.Tech Information Technology Final year at Kongu engineering college, Erode. Referred many papers related to sentiment analysis and found which machine learning algorithm suits the best for the following dataset.

**P. Kani Kumar,** currently studying B.Tech Information Technology Final year at Kongu engineering college, Erode. Referred many papers related to sentiment analysis and which dataset works the best for the given scenario.