Detection of Depression and Mental illness of Twitter users using Machine Learning

M. Ambika, K.V. Devakrishnan, A. Divya, R. Gowtham Raj, K. Kaviyaa

Abstract*: Today Micro-blogging has become a popular Internet-user communication tool. Millions of users exchange views on different aspects of their lives. Thus micro blogging websites are a rich source of opinion mining data or Sentiment Analysis (SA) information. Due to the recent emergence of micro blogging, there are a few research works devoted to this subject. We concentrate in our paper on Twitter, one of the prominent micro blogging sites to analyze sentiment of the public. We'll demonstrate, how to gather real-time twitter data for sentiment analysis or opinion mining purposes, and employed algorithms like Term Frequency - Inverse Document Frequency (TF-IDF), Bag of Words (BOW) and Multinomial Naive Bayes (MNB). We are able to determine positive and negative sentiments for the real-time twitter data using the above chosen algorithms. Experimental evaluations below shows that the algorithms used are efficient and it can be used as a application in detection of the depression of the people. We worked with English in this article, but for any other language it can be used.

Keywords: Machine Learning, Python, Sentiment analysis, Twitter.

I. INTRODUCTION

Micro blogging[1] has become one of the most commonly recognized channels of communication used by people across the globe. Text messages appear daily on popular websites offering micro blogging services such as Reddit, Pinterest, Facebook, Twitter and so on. Such text messages are released on social networks in order to share views on different issues and thus address the latest problems Due to the unlimited formation of messages and also the simple accessibility of microblogging sites, Internet users have been inclined from conventional communication to microbloggingThe rapid increase in the number of user posts about consumer goods and services or views on political and religious issues is making microblogging sites such as Twitter more common for analyzing sentiments. Any registered user on Twitter can post a tweet with a maximum length of 140 characters[5]. Sentiment Analysis (SA)[1] suggests the customer whether or not the item data is successful before they get it. Marketers and companies use this research data to understand goods or services in such a way that they can be delivered in keeping with the consumer requirements [2]. For the following reasons, conducting the Sentiment Analysis (SA) on tweets is aimed at best:

1. Tweets which are in nature abstract.
2. Research can be performed in real time.
3. The variety of tweets for the study can be retrieved[3].

II. LITERATURE SURVEY

[1] Twitter Sentiment Classification Using Distant Supervision:

There is no previous model correlated with classifying message sentiments on microblogging sites such as Twitter, Facebook, etc. This paper presented the results of the Machine learning(ML) algorithms used by distant supervision to characterize the Twitter messages sentiments. The components are used as both unigrams and bigrams. Accuracy improved compared with unigram components are Naive Bayes (81.3% from to 82.7%) and Maximum Entropy (from 80.5 to 82.7). There was however a decline for SupportVecMachine (from 82.2% to 81.6%) [4].

[2] Stock prediction using twitter sentiment analysis:

We apply sentiment analysis and machine learning ways in this project to find out the connection between people sentiment and business sentiment. To forecast stock market movements, we use twitter data to predict people's mood and use expected sentiment and the DJIA values of previous days. We introduced new cross validation(CV) method to test the results and that got 75.56 % accuracy using the DJIA values and Twitter data of Self Organizing Fuzzy Neural Networks (SOFNN)[6].

[3] Sentiment analysis of movie reviews in Twitter using machine learning techniques:

In this nominal we examined Movie reviews using different Machine Learning (ML) techniques such as Naive Bayes, K-Nearest Neighbor and Random Forest. The polarity of the tweets is identified using various techniques. The performed algorithms were: Naive Bayes(NB), K-Nearest Neighbor (KNN), Random Forest. The best results are given by Naïve Bayes classifier (NBC). The Naïve Bayes classifier (NBC) achieved accuracy of 81.45%, the Random Forest classifier achieved accuracy of 78.65%, the K-Nearest Neighbor classifier achieved accuracy of 55.30%[7].

III. SYSTEM ARCHITECTURE

The system architecture consists of the components as shown in Figure 1 such as tweets extraction from twitter, data preprocessing, extraction of features, training set are specified for the analysis provided.

Revised Manuscript Received on April 25, 2020.
* Correspondence Author

M. Ambika*, Pursuing, (B.E) degree at Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology, Coimbatore.

K.V. Devakrishnan, Pursuing, (B.E) degree at Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology, Coimbatore.

A. Divya, Pursuing, (B.E) degree at Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology, Coimbatore.

R. Gowtham Raj, Pursuing, (B.E) degree at Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology, Coimbatore.

K. Kaviyaa, Pursuing, (B.E) degree at Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology, Coimbatore.

Retrieval Number: D8314049420/2020©BEIESP
DOI: 10.35940/ijeat.D8314.049420
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication
© Copyright: All rights reserved.

The training set is obtained from a predefined set of positive or negative tweets that can be achieved using the algorithm and the result of positive and negative tweets are obtained. The used Classifier classifies the tweets according to the training set and regulates the performance. We will examine the microblog[1] named as Twitter in this article, classifying the "tweets" into positive and negative sentiment. We use Twitter hashtags (e.g., # best feeling, # DonaldTrump, # love) to explore the method for creating these data to classify positive and negative tweets to be used to train three-way sentiment classifiers. Thus, these tweets and hashtags are important to evaluate individual people's level of thought.

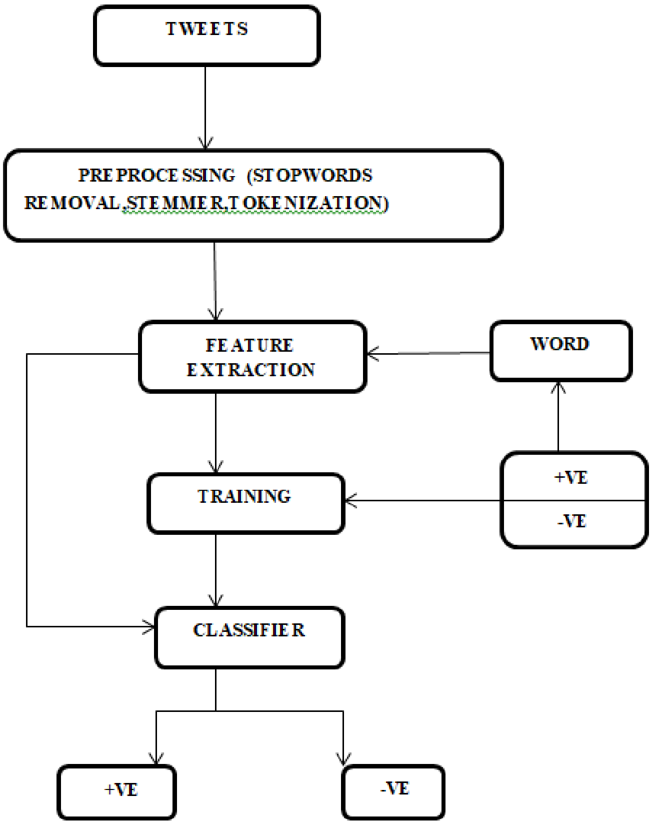

Fig 1 System Architecture

IV. METHODOLOGY

The tweets are imported on Twitter using the supported TwitterAPI[3]. From these API specific unique keys can be scrapped for each user such as consumer key, consumer password, access token password. We can analyze the different famous person thoughts on a Twitter event or occasion after collecting those details. Then pre-processing of the derived data is carried out. Pre-processing involves removing unnecessary words such as Removal of Stopwords[2], Stemmization and Tokenization that are necessary for analysis. We may label the tweets with the scraped tweets whether it's positive or negative.

A.Description of Dataset:

In this system, we have used the dataset called sentiment_tweets. csv for training the model. It includes the following fields: User ID, message (Tweets) and label .The CSV (Comma Separated Value) file contains 10314 data of which 8000 data are used for training and 2314 are used for testing. The below Table 1 shows the sample dataset.

Table 1:Sample Dataset

User Id	Message	Label
106	just had a real good moment. i missssssssss him	0
217	is reading manga	0
220	@comeagainjen http://twitpic.com/2y2lx- http://www.youtube.com/watch?v=zoGfqvh28	0
288	@lapcat Need to send 'em to my accountant tomorrow. Oddly, I wasn't even referring to my taxes. Those are supporting evidence, though.	0
540	ADD ME ON MYSPACE!!!myspace.com/LookThunder	0
624	so sleepy. good times tonight though	0
701	@SilkCharm re: #nbn as someone already said, does fiber to the home mean we will all at least be regular now	0
1193	nite twitterville workout in the am -ciao	0
1324	@daNanner Night, darlin'! Sweet dreams to you	0
1332	Good morning everybody!	0

B.Software Description:

With the help of Spyder and Jupyter notebook, the data visualization like wordcloud, confusion matrix is created in this system[3]. Pandas, numpy, matplotlib, pyplot, list, Dictionary are the predefined functions. Pandas, which is used to convert csv file to dataset. Numpy, which is one of Python's essential calculation library. Python is made up of several inbuilt categories such as list, dictionary, set, and tuple. A list is the equivalent of a Python array, but is resizable and can contain elements of different types. A dictionary stores pairs (key, value), like a Java Map or a JavaScript object[3]. Tweepy is used for Twitter API access[3] and is open source. The Wordcloud is a visual representation of text data which displays the word frequency. The confusion matrix is also used to characterize the model's output on a collection of test data for which it is known the true positive, true negative, false positive, false negatives. It shows an algorithm's visualised results.

C. Pre-processing of Data:

All the texts are divided down into tokens. This process is known as tokenization. For example “this is an amazing phone” is divided into this‟, „is‟, „an‟, „amazing‟ and „phone‟ as individual tokens. On a space, a token is identified. Stop words like articles, prepositions, conjunctions, and pronouns are also removed. Stop words provide little or no information [2]. Stemming is is done, which the process of reducing the change in form of words to their root terms such as mapping a group of words to the same base stem word[2].

D. Extraction of the features:

This is the main function related to classification. This includes removing irrelevant words or terms which convey no sentiments at all. Unigram (n=1) is used for extraction of a feature[1]. The unigram stands for single and distinct words.

E. Algorithms used (Classification):

Term Frequency - Inverse Document Frequency (TF-IDF):

The TF-IDF[8] assigns each word a score. The term frequency is determined by counting the number of times a given word or phrase as used in the document and the inverse document frequency(IDF)[8] is determined by dividing the total number of documents by the number of documents containing the given term given by the formula:

$$w_{i,j=tf_{i,j}\,X\,\log\left(\frac{N}{df_i}\right)}$$

$tf_{i,j}$= the number of occurrences of i in j

df_i= the number of documents containing i

N= the total number of documents

Bag of Words (BOW):

The Bag Of Words (BOW)[8] is an algorithm that calculates the total number of occurrences of a word in the whole document. The BOW lists the terms that occur as a pair, with their word counts per text. In the table where the words and documents that are vectors are stored, each row is a word, each column represents the document, and each cell represents the word count.

	Doc1	Doc2	Doc3
Hurt	20	7	15
Suffer	2	21	0
Suicide	32	0	12

Multinomial Naive Bayes:

In the model of multinomial naïve bayes, samples or feature vectors indicate the frequencies with which multinomial produces such events ($p_{1,\ldots\ldots,}p_n$) where p_i is the likelihood of the event *i* occurs. The $x = (x_{1,\ldots\ldots,}x_n$) sample or feature vector is a histogram, with x_i counting the number of times event *i* was seen in a given instance. It is the event model usually used for the classification of documents, with events reflecting the occurrence of a word in a single document (like the word bag). The histogram(x) probability is given as follows,

$$p(x|C_{k)=}\frac{(\sum i\, x_i)!}{\Pi i x_i\,!}\ \Pi i p_{ki}{}^{x_i}$$

F. Data Visualization:

Data visualization is a graphical representation of data. This involves in bringing forth the images that helps to understand the raw data clearly. Here we used two visualization techniques called Word cloud and Confusion matrix. The Word cloud (or tag cloud) is a visual display of text data. Words and their meaning are displayed in various font sizes or colors. Figure 1 represents the positive word cloud and figure 2 represents the negative or Depressed word cloud. The confusion matrix is a table that explains how a classification model performs on the test data that truth values are known.The terminology related to the confusion matrix are True positive (TP) gives the number of positive tweets correctly marked, as positive and false positive (FP) is the number of negative tweets wrongly marked as positive. True negative (TN) is the number of negative terms correctly classified as negative and false negatives (FN) is the number of positive words wrongly classified as negative tweets. The Figure 3 represents the confusion matrix,

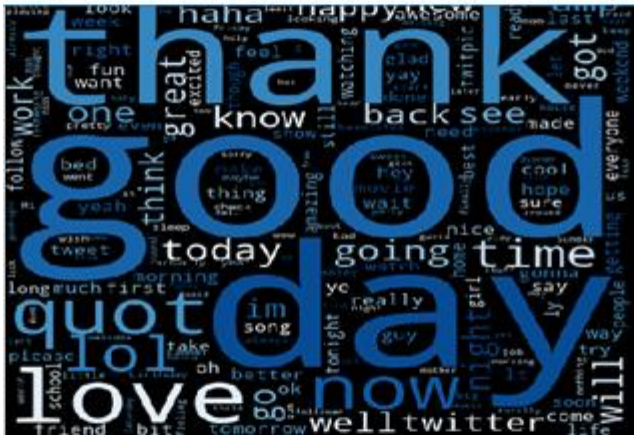

figure 1: Positive word cloud

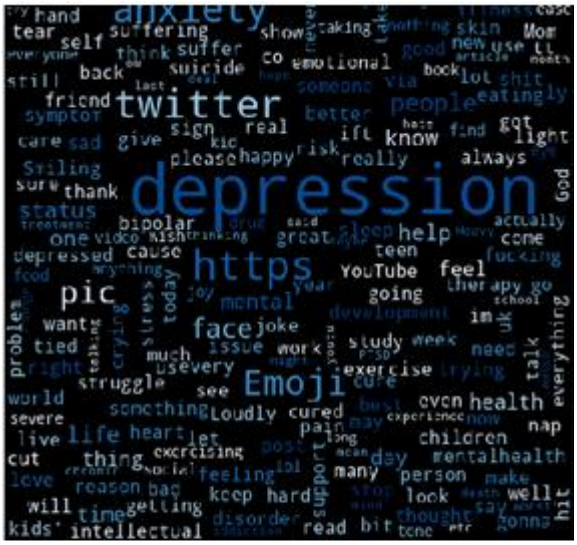

figure 2: Depressed word cloud

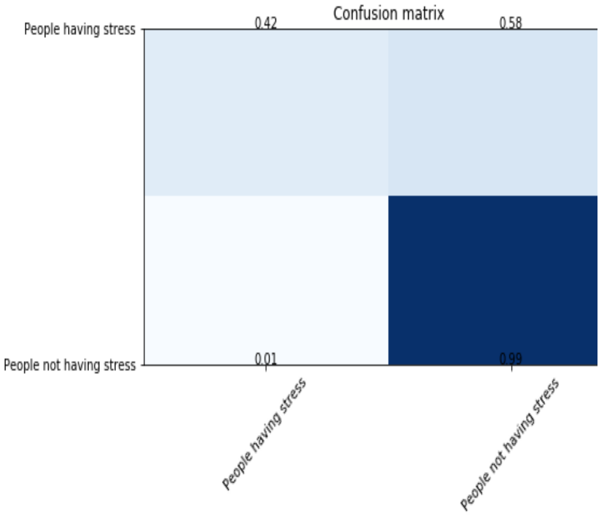

Figure 3:Confusion matrix

V. CLASSIFICATION REPORT

The performance of the algorithms are analysed with the four parameters which are Accuracy, Recall, Precision, and F-measure [8]. Here ,We used three algorithms for Sentiment Analysis of Twitter data. We have observed that Multinomial Naïve Bayes (MNB) has given more accuracy of 91% compared to other two algorithms where, TF-IDF has given 85% of accuracy and Bag of Words (BOW) has given 81% of accuracy. The figure 4 categorical graph represents the parameters of the algorithms used.

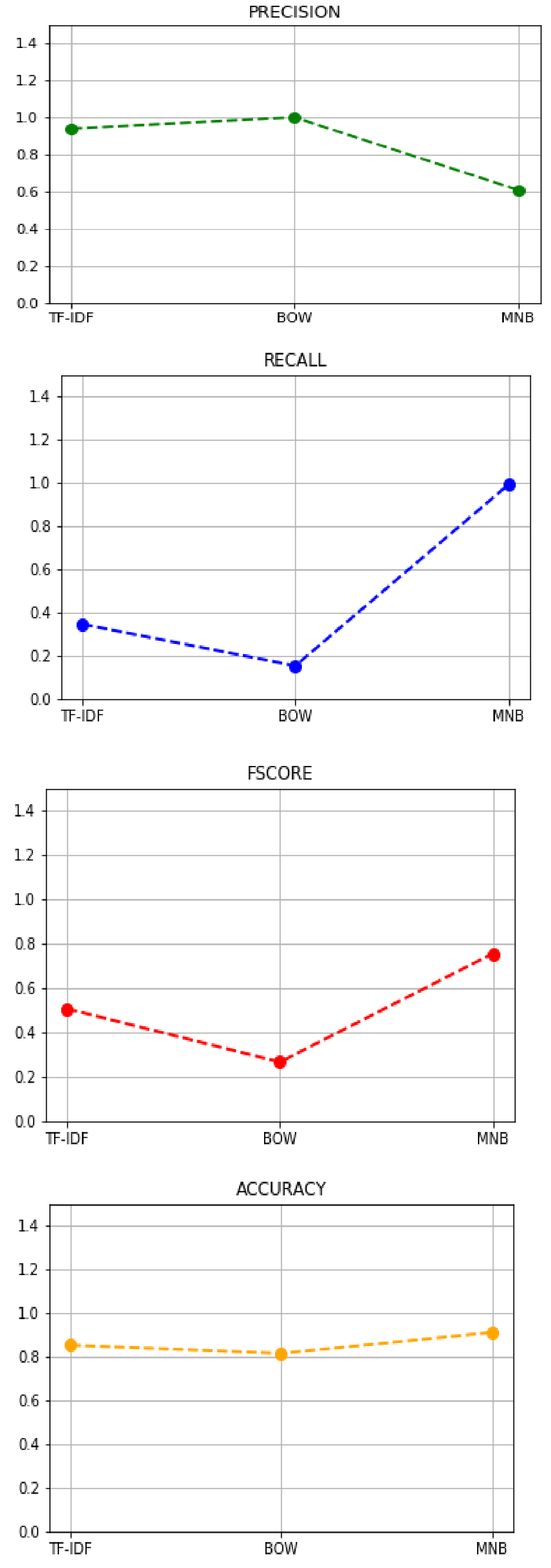

figure 4: Categorical Graph[8]

VI. CLASSIFICATION RESULT

The table shown below defines the Classification result of the algorithms using the calculation of the parameters.

Table 2: Classification result [8]

	TF-IDF	BOW	MNB
Precision	0.9473	1.0	0.6103
Recall	0.3461	0.1538	0.9907
F-score	0.5070	0.2666	0.7553
Accuracy	0.8529	0.8151	0.9110

VII. CONCLUSION

Analysis of the twitter sentiment falls under the division of the mining of text and opinion. This focuses on analyzing the tweet sentiments and providing the data to the Machine Learning (ML) model in order to train it and then test its accuracy so that we can use this model for potential use as illustrated in the results. It consists of stages such as data collection, pre-processing of documents, sentiment detection, classification, training and model testing. In this paper ,We worked with various algorithms and sentiments classification are identified .Thus the accuracy of Multinomial Naïve Bayes is found to be better than TF-IDF and Bag of Word algorithm as mentioned before. This subject has grown over the last decade with models ranging from nearly 85% to 90% performance. Yet this fails in the field of data diversity. With the slang and the short types of words used it has many implementation issues. When the number of classes is increased, many analyzers don't perform well. Therefore the study of sentiments has a very bright outlook for future growth.

REFERENCES

1 Pak and P. Paroubek. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”. In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010.

1 V.Lakshmi, K.Harika , H.Bavishya and Ch.Sri Harsha, “Sentiment Analysis of Twitter Data”, 2017.

2 Hetu Bhavsar and Richa Manglani, “Sentiment Analysis of Twitter Data using Python”,2019.

3 Alec Go ,Richa Bhayani and Lei Huang, “Twitter Sentiment Classification Using Distant Supervision”, Stanford University, CA 94305.

4 Malhar Anjaria and Ram Mohana Reddy Guddeti,“A novel sentiment analysis of social networks using supervised learning”,2014.

5 Anshul Mittal and Arpit Goel, ” Stock prediction using twitter sentiment analysis”, Stanford University.

6 Palak Baid,Apoorva Gupta, and Neelam Chaplot “Twitter sentiment analysis of movie reviews using machine learning techniques”,2017.

7 Metin Bilgin and Haldun Koktas,“Sentiment Analysis with Term Weighting and Word Vectors”,2019.

AUTHORS PROFILE

M. Ambika completed Bachelor of Engineering(B.E) in Computer Science and Engineering (CSE) and Master of Engineering (M.E) in Software Engineering(SE). She is currently working as Assistant Professor in the Department of Computer Science and Engineering at Sri Shakthi Institute of Engineering and Technology. She is currently pursuing her research in Machine Learning.

ambikase@gmail.com

K. V. Devakrishnan is currently pursuing his UG Bachelor of Engineering (B.E) degree at Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology, Coimbatore.

kvdkrish@gmail.com

Divya is currently pursuing her UG Bachelor of Engineering(B.E) degree at Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology,Coimbatore.

divyaanandan0@gmail.com

R. Gowtham Raj is currently pursuing his UG Bachelor of Engineering (B.E) degree at Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology, Coimbatore.

rgowthamraj00@gmail.com

K. Kaviyaa is currently pursuing her UG Bachelor of Engineering(B.E) degree at Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology, Coimbatore.

kaviyaa0204@gmail.com