International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 21 (2018) pp. 14956-14963

Machine Learning based Analysis of Twitter Data to Determine a Person's Mental Health Intuitive Wellbeing

Rakshitha C L
Department of Computer Science & Engineering, Dr.Ambedkar Institute of Technology, Bengaluru, India.

Gowrishankar S
Department of Computer Science & Engineering, Dr.Ambedkar Institute of Technology, Bengaluru, India.

Abstract

Social media has become a space for individuals to convey their views, thoughts, and emotions publicly on diverse topics. Twitter is a popular microblog which is one of the widest communication platforms. We discuss about the about mental health being of a person using tweets that were extracted based on emotional keywords. Sentiment analysis of tweets is carried out using Machine learning based algorithms like Support Vector Machine(SVM), and Naive Bayes algorithm to classify the mental health of a person based on their intuitive wellbeing.

Keywords: Social media, Sentiment Analysis, Machine learning, SVM, Naive Bayes.

INTRODUCTION

Individuals share their feelings, views, opinion through various Microblogging sites nowadays. Technology has made headway in the way individuals interact with each other formally or informally. Internet, through its various channels, has enabled individuals to share their joys, sorrows, disappointment, feelings with the world. One such platform is Twitter which has millions of users who share tweets on different topics [17]. Tweets are short message with 280 characters, which can be viewed by anyone instantly.

There are different approaches to perform sentiment analysis like machine learning approaches, lexical based approaches and hybrid approaches.

Machine learning based method is emphatically reliant concerning size and quality of training corpus which requires manual human mediation for labeling tweets into different classes. It is essentially partitioned into two fundamental methodologies - Supervised and Unsupervised. Supervised machine learning has a training set along with a testing set which should be classified. Classification can be done using machine learning algorithms like SVM, and Naive Bayes. Unsupervised machine learning does not contain a training set, yet, it instead ingests information and endeavors to analyze the information in parts and gathers the analyzed parts into bunches. Lexical approach is classified into dictionary and corpus based approaches. Dictionary based approaches utilize a dictionary to analyze the polarity of a sentence. Hybrid based approaches use both supervised and unsupervised methods for classification.

Machine learning approaches give more accuracy using SVM algorithm [4]. In this paper, we use machine learning approaches to determine the mental health of a person based on their intuitive wellbeing. We consulted a Psychiatrist for guidance on the selection of specific words known as emotional keywords, which is used by individuals who are rationally stable however because of a few conditions they might be discouraged and their psychological wellness may not be sound in their correspondence, by which we can see if the individual has a psychological issue.

The contents of this paper are as follows. Section 2 deals with the literature survey. Section 3 gives the details of Tools and Algorithms utilized in the paper. Section 4 exhibits the proposed tweets sentiment classification approaches. Section 5 presents the experiment results to examine the execution of our classifier. Section 6 concludes the paper.

LITERATURE SURVEY

Paper by Pang et al.[1] is one of the earliest conventional text which used the standard feature framework using machine learning for classification of data into two classes for unigram as a feature in classification which goes well with Naive Bayes and SVM on a movie review. [11] reported enhancement by attaching a preprocessing filter to expel target sentences which enabled the classifier to concentrate just on subjective sentences, raising the precision from 82.9% to 86.4% of every motion movie surveys dataset. Sentiment analysis of short messages is considered as a relatively more difficult issue than that of customary content, for example, film survey records. This is halfway because of the confinement of short length, the regular utilization of casual words and the fast advancement of language in short messages [3]. Machine learning based systems achieve domain based sentiment classification for training data and to use most efficient classification algorithm. The size of the training set plays an important role to build a high quality classifier [4] a classifier such as SVM is more appropriate for the large training set.

Using Latent Dirichlet Allocation (LDA) model, Phan et al. [5], based on extern Wikipedia corpus, presented a framework through which they expanded short texts by attaching hidden topical names.

Banerjee et al.[6], by improving their delineation with extra aspects from Wikipedia such as the titles of select Wikipedia articles, proposed a method of enhancing the precision of the search in the grouped short text.

By the bootstrapping ensemble framework we can extemporize class imbalance problems and linguistic variations in the text can be overcome [7]. Coletta et al.[8] exhibited the execution of SVM, consolidated bunch gathering arrangement on Twitter information.

Kouloumpis et al.[9] focused on the regularities in the informal practices followed by Twitter users to propose a system for sentiment analysis. He classified the hashtags used in tweets as positive and negative in polarity. The content-based ranking method was proposed by Phan Ngoc and Myungsik Yoo [10].

KNIME which provides differences and analysis possibilities for online reviews and tweets. This system was proposed by Ana Minanovic [2] for the collection of data and sentiment analysis.

Alexander Porshnev et al.[12] proposed a combined Neural Networks and Support Vector Machine based classifier to examine the tweets on a specific stock market data. Christos Troussas [13] used Naive Bayes based classification to identify the sentiment on a particular status in Facebook comment text for analysis. Zhao and Gui [14] have put forth different strategies with preprocessing writings. According to them, elimination of numbers and stop words have next to nothing impact on text classification instead of expanding acronyms. Likewise, it has set up through various examinations, where Naive Bayes and Random Forest Classifiers are more delicate than Logical Regression and Support Vector Machine when modern preprocessing systems are utilized.

As an expansion of angle based sentiment mining approach proposed by Bing Liu, a unique deterministic approach was proposed by the creators in [15]. This new approach has a better performance than the model proposed by Liu. It enhances Recall is a functional estimate of success of the prediction, which is sensitivity. Because of the new and complex Natural Language Processing based tenets that have been created for both subjective and sentiment classification. A topic on adaptive sentiment classifier was formally proposed by. Liu, S. Et al.[16] as an adaptive multiclass SVM model which transfers from an initial common sentiment classifier.

TOOLS AND ALGORITHM

Python Programming Language : We use Python on a Windows system. Python is an open source software and a high-level programming language. It is easy to learn and use due to its simple syntax. Python comes with an extensive set of standard libraries. Nowadays Python has become the ultimate language for analyzing data.

ParallelDots : ParallelDots is a Cloud based service and provides various APIs to perform text analytics. It offers services that provide advanced natural language processing over raw text and analyzes text by providing polarity. ParallelDots APIs can ingest each text per request for classification.

MonkeyLearn : MonkeyLearn is an online text classification service which provides the user to upload their data set to train an algorithm. It also provides some settings like different classifier type, stop words, and n-gram range. Monkey learn is intended for someone new to machine learning, where they can start text classification immediately without any prior knowledge.

Python libraries like Scikit-learn, NLTK toolkit, Tweepy, Matplotlib, TextBlob are used. Scikit-learn is used for classification of tweets. Scikit-learn is one of the leading platforms for implementing a Python program which also provides many classification algorithms. NLTK provides a collection of text processing libraries. TextBlob is a python library for processing textual data. Matplotlib is a plotting library for 2D graphics.

Naive Bayes : Naive Bayes algorithm takes the advantages of probability and Bayes theorem for prediction of a text. Multinomial Naive Bayes works excellent with natural language processing problems. In Multinomial Naive Bayes, all features are conditionally independent of their class.

$$P(A|B)=P(A)P(B|A)/P(B).$$

Where B represents the text to classify, A is an individual class P(A) and P(B) is a probability of text and class identity, P(B/A) represents the probability of a text B appearing gives class A .

Support Vector Machine(SVM) : In broad terms, an SVM models endeavours to find the best splits in the training data after the split, the new data is placed on one side of the split and given a score so that there is a likelihood that the data placed is on the correct side of the split. Fig 1 shows the SVM illustration, where solid line is the maximal margin hyperplane the SVM learns from the training data and dashed lines represents the closest training data point in the respective classes, support vector are the points closet to decision boundary.

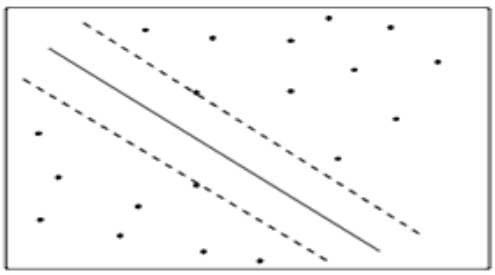

Figure 1. Support Vector Machine

SYSTEM ARCHITECTURE

We use Machine learning approaches, supervised approaches which have both training and testing data and classification done using a machine learning algorithm. In this research, we

use two approaches for classification. The first approach using the MonkeyLearn API is shown in Fig 2.

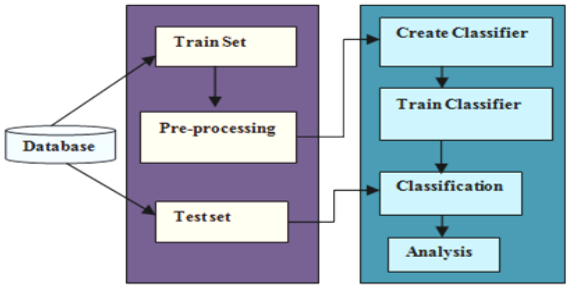

Figure 2. First Approach For Text Classification

In the ***first approach***, the dataset of tweets for a particular emotional keyword is given as input and the output from this process are fed to MonkeyLearn API for classifying text by using Multinomial Naive Bayes algorithm.

Data Collection : The data corpus is collected using some specific words which is related to the mental health of a person. For this research, we consulted a Psychiatrist for guidance on the selection of specific words known as emotional keywords, by which we can find out whether the person has a mental disorder. Using these emotional keywords, data is collected from Twitter. In this research, we are collecting and storing the data in the database. First, we need to create a database for a few selected data fields saved in Twitter, as shown in Fig 3. These are the data fields used in this research.

User_ID-The unique user id of the author.

User_Name-The name registered by the user. *User_Screen_Name*-The screen name of the user. *User_Description*-The user profile of the author. *User_Account_Created_At*-The date and time the user account was created.

User_Language-The language of the account as reported by the user.

Friends_Count-The number of counts the user's account is following. *Followers_Count*-The number followers the account has. *Tweeted_At*-The date and time the tweet was created. *Tweet_ID*-The unique ID of the tweet as assigned by Twitter. *Tweet_Text*-The text generated by the author.

Tweet_Interest-The emotional keyword for which tweets are searched. *Tweet_Language*-The language of the tweet. *Tweet_Likes_Count*-The number of count of likes for a tweet. *Retweet_Count*-The number of count the tweet has been retweeted.

Time_Zone-The time zone of the user.

Tweet_Source-The source of the tweet.

User_Location-The self reported location of the user. *User_Verified*- Boolean value indicating if the user is verified accounts.

As shown in Fig 4, we need to create a Twitter account and create a Twitter application, and we need to generate keys and access token. Tweets are extracted using access token. Using the emotional keywords, the tweets are extracted from Twitter. After data collection, the data is stored into database accordingly. For this research, we have collected 65,994 tweets using Twitter API.

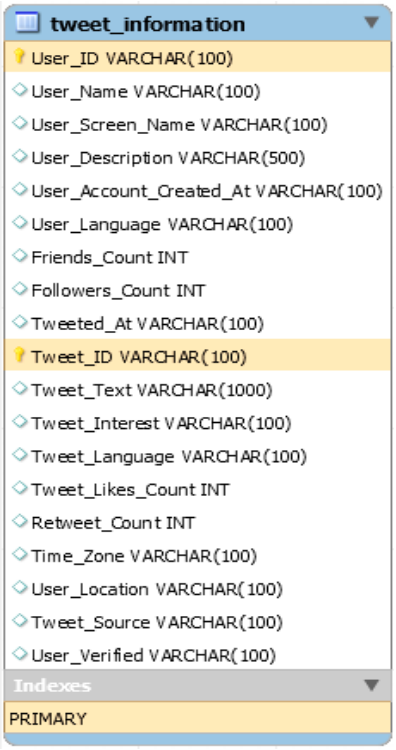

Figure 3. Data Attributes Created In Database

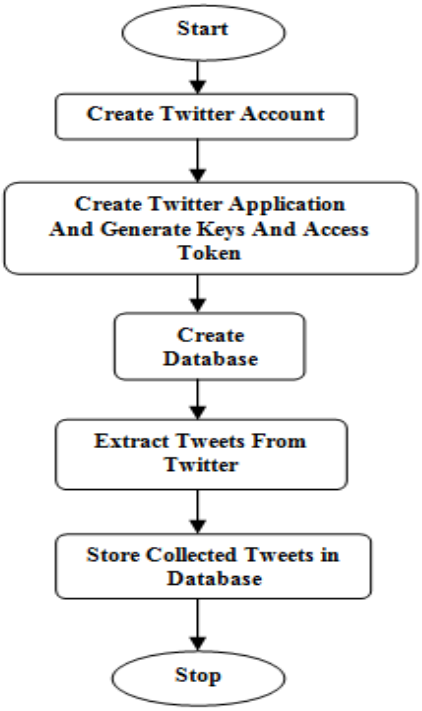

Figure 4. Data Collection Phase

Data Pre-processing : The raw data collected from Twitter contains a large amount of surplus information which is unstructured data, so we need to pre-process it to get a simple and structured data. We use the NLTK toolkit for pre-processing the data by using tokenization, stop words removal and stemming as shown in Fig 5. Tokenization is splitting the words in the sentence by using a delimiter, white space and removal of punctuations is done in this stage. Using stop words, useless words which are used in a sentence, such as "the", "a", "an", "in" that do not affect the sentence is removed. Stemming is the removal of derivational affixes, after performing these steps, we will have a structured data.

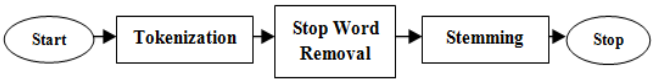

Figure 5. Pre-Processing Phase

Create Classifier : MonkeyLearn provides a REST API that can be directly accessed and also accessed through the client library to classify the tweets. First, we need to install MonkeyLearn and authenticate using user access token and grant access to MonkeyLearn API. Then, we need to create a new classifier and categories as shown in Fig 6 which represent the scenario.

Train Module and Classification : In train module, we manually label the tweets for training, which are classified using ParallelDots for each tweet based on its category, and use 20% of the collected data set for training the classifier. The classification is made using a machine learning algorithm like Multinomial Naive Bayes. In the result part, sentiment analysis is carried out for each emotional keyword collected where 20% of the dataset is used for training and classification using Multinomial Naive Bayes and prediction to determine whether the dataset is a positive class or negative class. Positive class denotes that the dataset has more positive sentences and if the dataset has more negative sentences, then it is classified under the negative class.

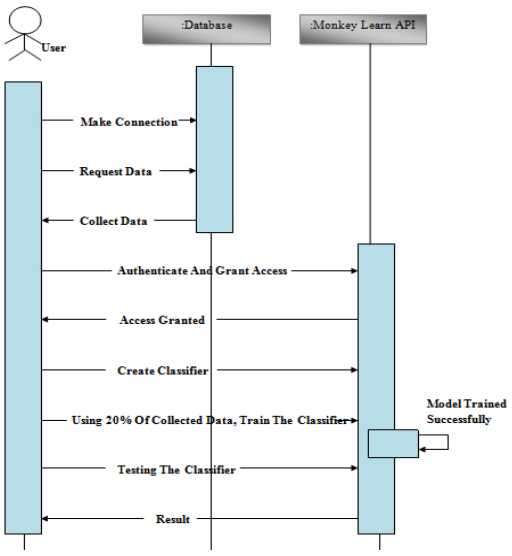

Figure 6. Text Classification Using Monkey Learn API

Fig 6 gives the clear idea of the proposed system easily which represents the schema diagram. It shows the overall outline for text classification using MonkeyLearn.

The ***second approach*** is building own classifiers in Python using a machine learning algorithm like SVM available in Scikit-learn library. The main TfidfVectorizer function is used to build classifiers.

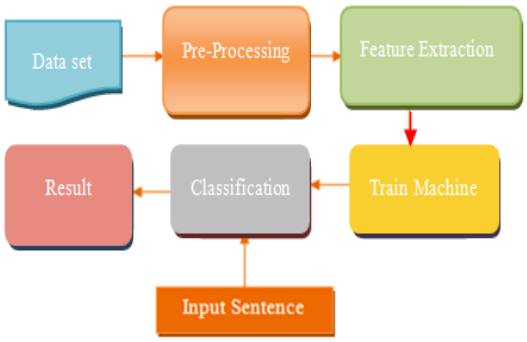

Figure 7. Second Approach For Text Classification

Fig 7 shows the supervised machine learning approach.. Steps for this approach are as follows: 1. Data Collection 2. Pre-Processing 3. Feature extraction 4. Training the Machine 5. Classification 6. Prediction.

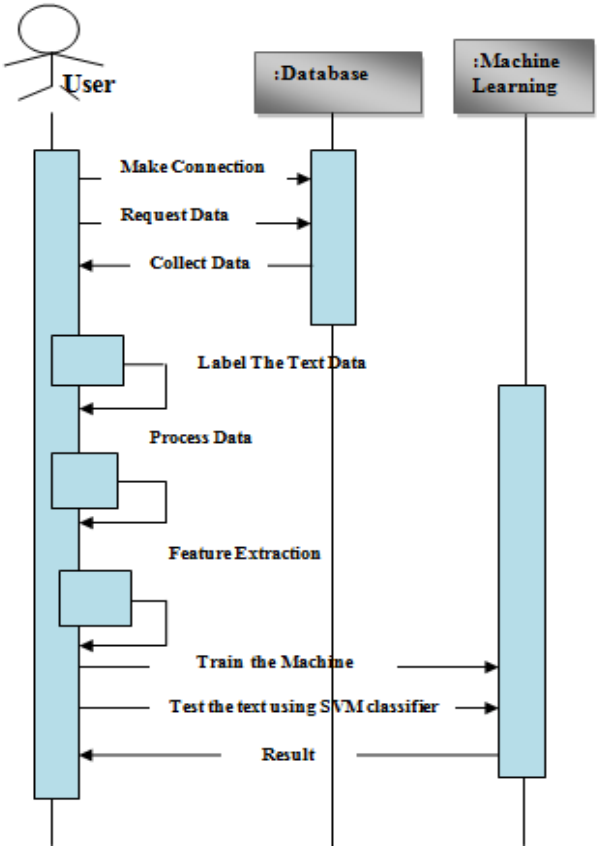

Figure 8. Analysis Using Supervised Machine Learning

In the second approach, the input is given in the form of text sentence, where output displays the accuracy score of a machine and polarity by using SVM algorithm.

As outlined above, data collection and data pre-processing done for the first approach remain the same for the second approach also.

Feature Extraction : Term Frequency–Inverse Document Frequency(TF-IDF) which is used to convert the text into decimal vector [19] is the most commonly used weighting method for document description. There are many feature selections like chi and select feature according to the k highest score

Training The Machine : 20% of the collected dataset is used for training purpose. In this approach also, we use ParallelDots API which is used to classify each tweet. Classified tweets are used for manual labeling of tweets for the training set. We use two classes, positive and negative, labeled to all tweets. More the number of training sets, more the accuracy in a supervised approach.

Classification : Alec Go et al.[18] founded that SVM works better than Naive Bayes, so we adopted SVM for building the classifier. In classification, text classifier has been built on our own using SVM classifier built in Python. The main function to build classifiers includes tfidtransformer (TF-IDF), both in training and classification stage. In the result part, analysis of the text is done by providing an input sentence, and sentiment analysis is done for a particular tweet as having positive or negative polarity.

RESULTS

We are dealing with the human mental health, where data analysis plays an important role. By using Emotional keywords collected based on commonly expressed words by depressed and mentally ill persons, which are recommended by a Psychiatrist, the tweets are extracted from Twitter. Fig 9 shows all the emotional keywords used in this research and represented in the WordCloud which visualizes the most used emotional keywords from the entire dataset.

The data is collected from Twitter using Twitter API and stored in a database. For this research 65,994 tweets were collected. In this paper, where data analysis is done by both supervised and unsupervised machine learning approaches. Using MonkeyLearn API, text prediction is done based on Naive Bayes algorithm and it is also done using TextBlob for text classification which is an unsupervised approach.

Table I. Sentiment Analysis using Monkeylearn and Textblob

Tweet Interest	Number Of Tweets	TextBlob	MonkeyLearn
		Negative	Negative
Tearful	892	0.43	0.43
Resigned	830	0.35	0.41

TABLE I. shows the sentiment analysis done using TextBlob and MonkeyLearn API. On examination of the highest number of tweets collected for each emotional keyword, the analysis was done as shown in TABLE I. Tearful was the highest Tweet_Interest collected in a period of 9 days. Using the TextBlob classifier, it was determined that 0.43% of tweets statements have negative connotations and also by using MonkeyLearn API it was determined that tearful, emotional keyword have 0.43% of Negative tweets. We can analyze all the remaining emotional keywords based on the above analysis as shown in Table I.

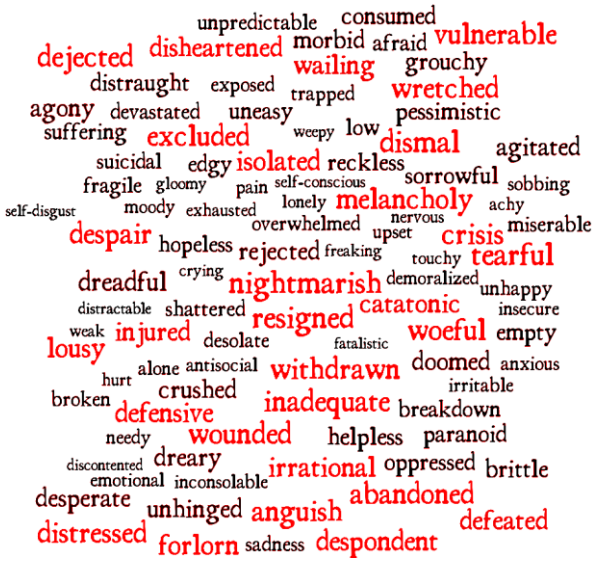

Figure 9. WordCloud which visualizes the emotional keyword used in this research

All graphs are plotted using Matplotlib plotting library. Fig 10 shows Day v/s No of Tweets. By looking at the graph, it can be determined that more tweets with negative connotations are tweeted on Wednesdays. By this graph, we can analyze that more people get emotionally and mentally weak on Wednesdays and Thursdays which may be due to work pressure and the ratio is more when compared to weekends.

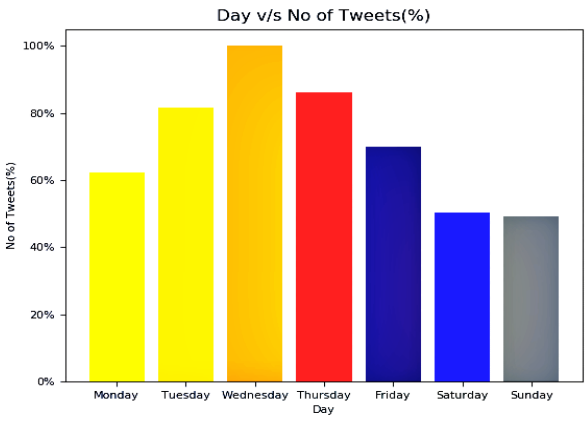

Figure 10. Day v/s No Of Tweets

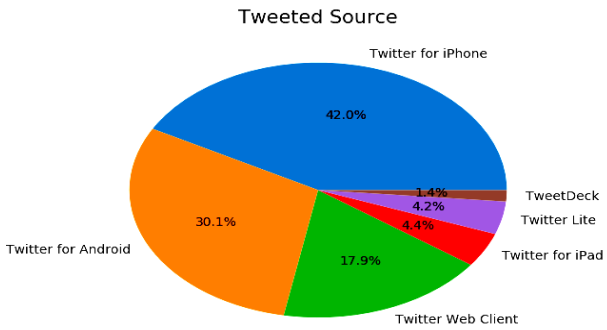

Figure 11. Tweets Collected By Tweeted Source

Fig 11, which shows tweets collected from different source. From this graph, we can analyze that more no of tweets are from the iPhone and iPhone users may be more prone to mental illness.

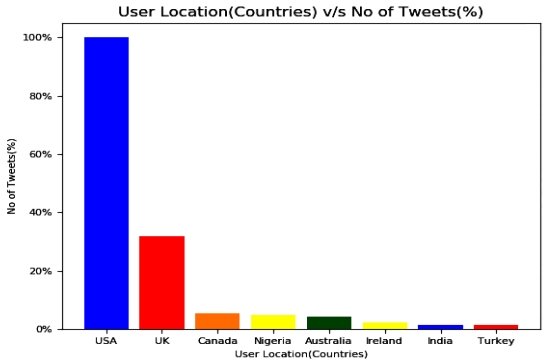

Figure 12. User Location v/s No of Tweets

Fig 12 shows User location v/s No of Tweets. The USA takes the first place for more number of users tweeting. It can be analyzed that people get depressed in that area due to loneliness. Interestingly, even India ranks in the top 10 countries for the highest number of tweets in the world, could be due to a high number of population.

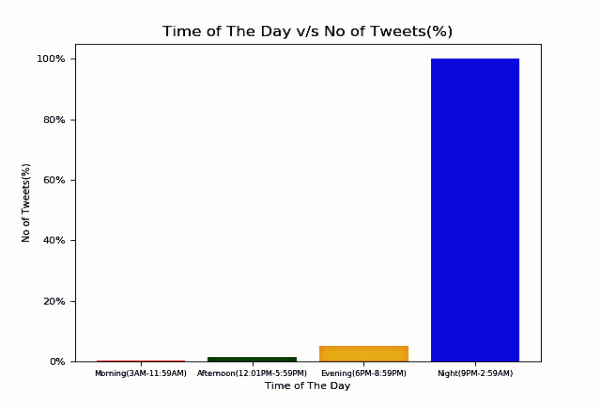

Figure 13. Time of The Day v/s No of Tweets

Fig 13 shows Time of The Day v\s No of Tweets. A maximum number of tweets are tweeted at night time. We can analyze that in developed countries, the majority of the people live an independent lives, and they experience more mentally weakness during night time.

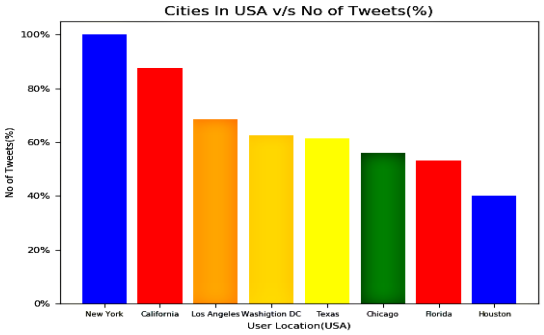

Figure 14. Cities In USA v/s No of Tweets

Fig 14 shows Cities in the USA v/s No of Tweets. This graph is represented because as shown in Fig 12, USA is the first place, so we can move forward and look into cities in the USA. As per the national institute of science, people easily get into mental illnesses in America, 26% of the Americans over the age of 18 meet the criteria of having a mental illness. New York stands in the first place. Through this, we can analyze that most developed cities are prone to more illness.

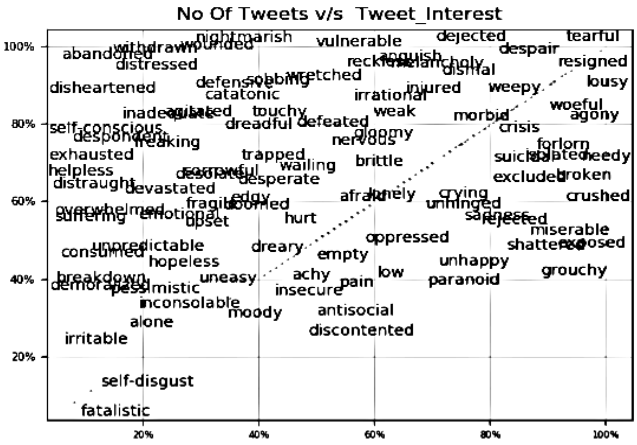

Figure 15. No of Tweets v/s Tweet_Interest

Fig 15 represents the highest common words used by a depressed person versus the least common words used.

To analyze the mental health of a person from a text classification using the supervised approach proposed in this paper. In this research, input for the result is given by a text sentence. The output is given in the form of polarity, positive or negative. Mental health of a person is predicted based on the output. The output is as shown, in Fig 16, 17 and 18 for analysis of text using machine learning. For example, in Fig 16 input is text sentence - "i am feeling sad" and the output is "Negative" for the input text.

Figure 16. Text Analysis Using Machine Learning

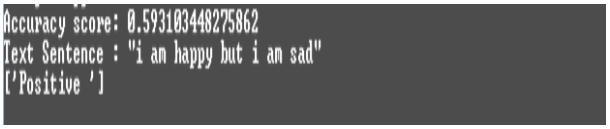

Figure 17. Text Analysis Using Machine Learning

Figure 18. Text Analysis Using Machine Learning

CONCLUSION

Social media has become a part of daily communication in social life. Twitter is one of the microblogging sites which is used worldwide for interacting with one another. Nowadays mental health is also a main issue in the society. We identify the mental health of a person using machine learning approaches, the data collection is done by Twitter API using emotional keywords and classifier are built for analysis using two classes to classify the polarity of tweets.

ACKNOWLEDGEMENT

The second author would like to acknowledge that this research work was supported in part by the VGST grant of Govt. of Karnataka, India, under the RGS/F scheme.

REFERENCES

[1] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[2] A. Minanovic, H. Gabelica, and Z. Krstic, “Big data and sentiment analysis using knime: Online reviews vs. social media,” in Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on. IEEE, 2014, pp. 1464–1468.

[3] Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. In The Semantic Web–ISWC 2012 (pp. 508-524). Springer Berlin Heidelberg.

[4] H. Cui, V. Mittal, and M. Datar. Comparative Experiments on Sentiment Classification for Online Product Reviews. In Proceedings of AAAI-06, pp.1265-1270, 2006.

[5] Phan, X. H., Nguyen, L. M., & Horiguchi, S. (2008, April). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proceedings of the 17th international conference on World Wide Web (pp. 91-100). ACM.

[6] Banerjee, S., Ramanathan, K., & Gupta, A. (2007, July). Clustering short texts using wikipedia. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 787-788). ACM.

[7] A. Hassan, A. Abbasi, and D. Zeng, “Twitter sentiment analysis: A bootstrap ensemble framework,” in Social Computing (SocialCom), 2013 International Conference on. IEEE, 2013, pp. 357–364.

[8] L. F. Coletta, N. F. F. d. Sommaggio Silva, E. R. Hruschka, and E. R. Hruschka, “Combining classification and clustering for tweet sentiment analysis,” in Intelligent Systems, 2014 Brazilian Conference on. IEEE,2014, pp. 210–215

[9] E. Kouloumpis, T. Wilson, and J. Moore, “Twitter sentiment analysis:The good the bad and the omg!” ICWSM, vol. 11, pp. 538–541, 2011.

[10] P. T. Ngoc and M. Yoo, “The lexicon-based sentiment analysis for fan page ranking in facebook,” in Information Networking (ICOIN), 2014 International Conference on. IEEE, 2014, pp. 444–448.

[11] Pang, B., and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[A]. 42ndMeeting of the Association for Computational Linguistics[C] (ACL-04). 2004, 271-278.

[12] A. Porshnev, I. Redkin, and A. Shevchenko, “Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis,” in Data Mining Workshops (ICDMW),2013 IEEE 13th International Conference on. IEEE, 2013, pp. 440-444.

[13] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno, and J. Caro, "Sentiment analysis of facebook statuses using naive bayes classifier for language learning,” in Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference on. IEEE, 2013,pp. 1–6.

[14] Zhao Jianqiang, Gui Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis" in IEEE Access, 2017

[15] Edison, Juan, “A novel deterministic approach for aspect-based opinion mining in tourism products reviews," Published in:Expert Systems with Applications, Volume 41, Issue 17, pp.7764-7775, Elsevier, 2014.

[16] Liu, S., Li, F., Li, F., Cheng, X., & Shen, H.. Adaptive co-training SVM for sentiment classification on tweets. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (pp. 2079-2088). ACM,2013.

[17] F. Bravo-Marquez, M. Mendoza, and B. Poblete, "Combining strengths, emotions and polarities for boosting twitter sentiment analysis," in Pro-ceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. ACM, 2013, p. 2.

[18] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. Technical report, Stanford University, 2009.

[19] Introduction to TF-IDF [EB/OL]. [2015-08-06].http://baike.baidu .com/view/ 1228847.htm.