Int.J.HIT.TRANSC:ECCN.Vol.6: Issue 2A (2020) Page 1-9 ISSN: 0973-6875

Available Online at www.hithaldia.in/locate/ECCN
All Rights Reserved

ORIGINAL CONTRIBUTION

A Sentiment Analysis on Tweeter Opinion of Drug Usage Increase by TextBlob Algorithm Among Various Countries During Pandemic

Sandipan Biswas[1], Shivnath Ghosh[2] and Sandip Roy[3]

[1,2,3]Department of Computer Science and Engineering, Brainware University, West Bengal, INDIA

ABSTRACT

Twitter, as a social media platform, is a great place where people from all parts of the world can make their opinions heard. Around 500 million of tweets daily which amounts to about 8TB of data are produced in Twitter, which can be very useful for analyzing important information via opinion mining about any news or launch of a product or a certain kind of trend to be observed well in twitter data. The main aim sentiment analysis (or opinion mining) is to extract emotion, opinion, subjectivity and attitude from a natural text on use of medicine/drug for the treatment of COVID-19. In twitter sentiment analysis, we have categorized tweets into positive and negative sentiment to conduct a study and find out that the use of clustering can quickly and efficiently differentiate tweets on the basis of sentiment scores. This paper surveys different approaches of clustering with respect to sentiment analysis and presents a way to find relationships between the tweets of political heads of different countries on the basis of polarity and subjectivity on the increase use of drug/medicine for the treatment of COVID-19.

Key words: Cluster, Opinions, Sentiments, Twitter.

1. INTRODUCTION

The way people thinking and expressing themselves have changed in last few years by internet which has changed their view points, expressions and platform to convey these expressions [1]. The platform like twitter has now become a large hub where people express their opinions they heard. It also acts as a great platform for many businesses to connect with clients and an also ideal place for celebrities to share their thoughts. Millions of users share their thoughts every moment by posting their tweets, which can be of no more than 140 letters to make the users to write short and to the point [2].

These sentiments found in the tweets can be categorized as positive, negative or neutral. Positive sentiments are those, which express good words or appraisals for a certain statement, news, expression, event, movie or product review. As well as, the negative sentiments are those, that contains bad words or criticize any event, product, movie etc. Similarly, neutral sentiments are those which are neither positive or negative [1][4]. These sentiments along with tweet attributes are usually categorized into many classes, but in this paper, we also express the focus on clustering with sentiment analysis.

Clustering is a job in which we can assign a certain group or class to certain object such that the object within the same group or class are more similar than that in the other distinguished group or class [3][5].

Corresponding Address: sandipan diet@rediffmail.com,

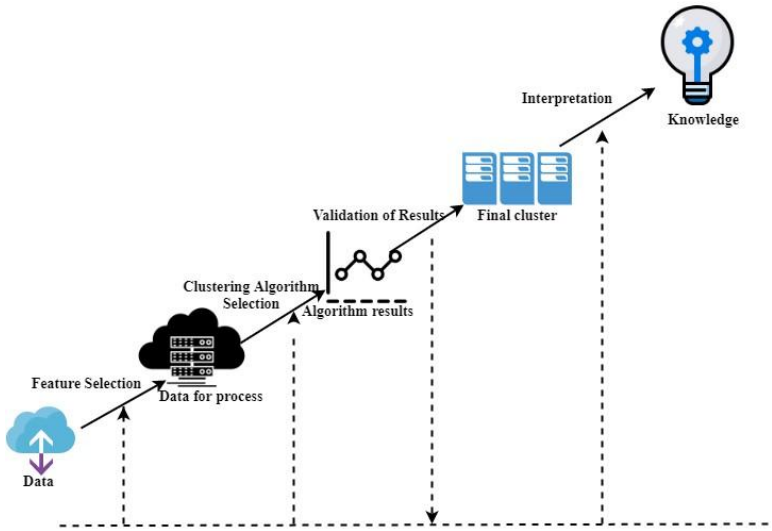

Figure 1: Stages of clustering

In or paper, sentiment scores are calculated for each tweet and then unsupervised techniques are applied on them to categorized similar cluster group. So clustering is an essential unique group within the same set of objects or patterns [7]. The stages of clustering are shown in Figure. 1.

2 .RELATED WORK

Bergsma et al. [6] on the basis of twitter data, have focused some invisible features such as ethnicity and gender by performing clustering on visible parameters such as name, location and friend-list. Romero, et al. [7] have expressed the need to be classify data and study of prediction parameters for any data analysis project. Vanessa Friedmann et al. [8] with the help of PCA, trimmed the data and altered it into a lower dimensional feature space, which can be used to be passed into the k-means algorithm to segment the samples into clusters.

Go et al. [9] in 2009 have expressed the emoticons as part of distant learning to obtain emotion or sentiment in which tweets ended with,were labelled as positive tweets and the tweets which ended with / were labelled as negative tweets. Barbosa et al. [10] have proposed a polarity prediction, which have used the data from 3 websites as unwanted labels so as to train a model and further use about thousand manually labelled tweets for adapting and another thousand manually labelled tweets for checking its reliability. Luiz F. S. Coletta et al. [11] have considered the classifiers combination and clusters in the classification of tweets sentiment by using C3ESL, in which the auxiliary data is provided by data divisions made from “bag-of-words” description with lexicon scores that can improve the classification correctness generated by classifiers.

Phan et al. [12] have proposed a framework for grouping tasks which involved short and sparse text segments. The structure of this framework is the collection of the appropriate section “universal dataset” from hidden topics that can be deter-mined and applied to the grouping or classification task, Gimpel et al. [13] have in-traduced a POS tagger for twitter data for Twitter POS tagging and conducted experiments to evaluate them, and provided their annotated corpus and trained POS tag-ger to the research community.

3. CLASSIFICATION ALGORITHM

3.1 Naïve Bayes

Naïve Bayes classifiers which are a family of simple “probabilistic classifiers”, ap-plies Bayes’ theorem with strong (naïve) independence assumptions among the features

of machine learning and it is simplest Bayesian network models [14]. Naïve Bayes has been studied and was presented into the text retrieval community for application in the early since1960s [15] and remains a baseline model for text categorization, to one of the categories or the other with word frequencies as a common feature. It has a huge application in automatic medical diagnosis [16].

3.2 K-Mean

This is one of the simplest unsupervised techniques for clustering, where the sample observations are divided into K clusters which are far from each other and then each data point gets assigned to the cluster which is nearest to it. Similarly, each data point is assigned one of the chosen clusters [17] and after that the cluster centres are calculated again and the same process goes on until the centres stop changing positions.

3.3 Fuzzy C-Mean

In this algorithm, the distance among the cluster centre and data point is used to label membership function to each data point with respect to each cluster centre.

If there is higher the proximity of the data set to the centre of the cluster, then more the membership function would be progressing towards that cluster centre [18]. Among each data point the total sum of membership values should have unity. The membership and cluster centres are updated after each iteration operation.

3.4 Decision Tree

It is a flowchart-like structure in which each internal node represents a test on a feature, then each leaf node represents a class label (decision taken after computing all features) and branches that represent combinations of features lead to those class labels and the paths from root to leaf characterize classification rules.

Decision tree is one of the predictive modelling approaches used in statistics, data mining and machine learning which are constructed via an algorithmic approach to identify ways to split a data set based on different conditions. It is one of the popular used and practical methods for supervised learning.

Decision trees where the target variable can take continuous values which are typically real numbers called as regression trees. Classification and Regression Tree (CART) is general term regression tree. There are numerous implementations of decision trees, one of the most well-known is the C5.0 algorithm [19].

4. PROPOSED MODEL

This section describes each component of the proposed model. We explaining the following model in a flowchart in Figure.2.We have also discussed our proposed algorithm in Algorithm 1.

4.1 Data Description

Twitter, which acts as a social networking platform, allows its users to send and read micro-blogs of up to 280-characters known as “tweets” and enables registered users to read and post their tweets through the web sites, short message service (SMS), and mobile applications.

As a global real-time communications platform, Twitter has more than 500 million monthly visitors and 270 million monthly active users around the world. Twitter’s active group of registered members includes World leaders of various countries, major athletes, star performers, news organizations, and entertainment outlets and it is currently available in more than 35 languages.

4.2 Proposed Model Component Description

In this section each component of the proposed model is described briefly. We begin by explaining the first component:

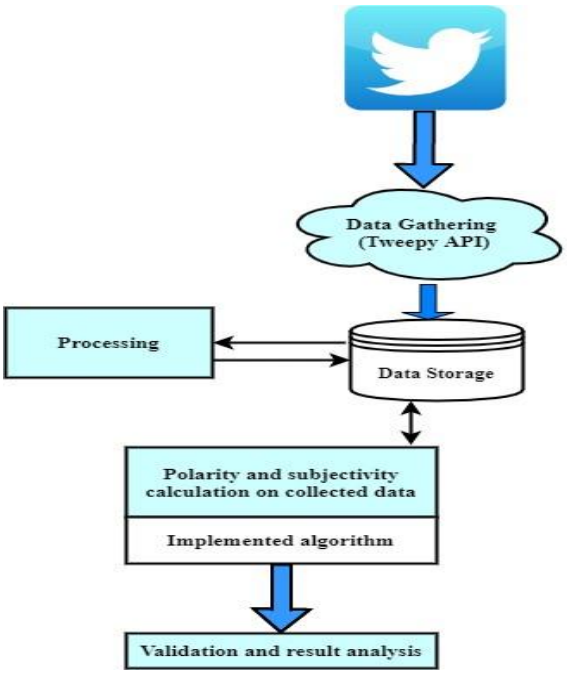

Figure 2: Proposed Algorithm Flowchart

ALGORITHM 1: Sentiment Classification of Tweeter Comments using Text Blob

Input:	Text File (Tweeter comments which include Nouns, Adjectives, Adverbs)
Output:	Values > 0 (Positive), Values < 0 (Negative), Values = 0 (Neutral)

Begin:

```
1.  Sentiment Analysis () ← File
2.  For each row in rows
3.    if Sentiment Polarity Score > 0 then
4.      Sentiment ← Positive
5.    else
6.      if Sentiment Polarity Score < 0 then
7.        Sentiment ← Negative
8.      else
9.        if Sentiment Polarity Score = 0 then
10.         Sentiment ← Neutral
11.     else
12.     end
13.   end
14. end
```

End:

15. **end**

4.2.1 Sentiment Analysis Component The below section describes in details of the proposed model step wise in Table 1.

4.2.2 Programming Language and simulator

Python: Python, a interpreted, high-level, general-purpose programming language and created by Guido van Rossum in 1991, for code readability that have notable use of significant whitespace. This object-oriented approach aims and helps programmers to write clear, logical code for small and large-scale program in various scientific and commercial projects [20].

Anaconda Navigator: Anaconda is a free and open-source [21] distribution (running in Window and Linux) for Python and R

programming languages that is used in scientific computing like data science, machine learning applications, large-scale data processing and various predictive analytics, etc. The aim is to simplify package management and deployment which are managed in the package management systemconda.[22][23].

Table 1: Stepwise Sentiment analysis Components

Steps No.	***Steps***	***Description***
1.	*Tokenization*	*Tokenization refers by dividing text or a sentence into a sequence of tokens, which roughly correspond to "words "which is one of the basic tasks of NLP and using TextBlob, the two steps are followed: i) A TextBlob object is created and passed a string with it and ii) TextBlob function is called in order to do a specific task.*
2.	*Noun Phrase Extraction*	*Here we extract the words in the previous section, instead of that we can just extract out the noun phrases from the TextBlob. Among them noun phrase extraction is particularly important when you want to analyze the "who" in a sentence.*
3.	*Part-of-speech Tagging*	*Part-of-speech tagging, which is also known as grammatical tagging, is a method to mark words present in a text on the basis of its definition and context.*
4.	*Words Inflection and Lemmatization*	*Inflection is a process of word formation where characters are added to the base form of a word to express grammatical meanings and it is very simple, i.e., the words we tokenized from a TextBlob can be easily changed into singular or plural.*
5.	*N-grams*	*It is a combination of multiple words together where N grams (N <1) are generally more informative as compared to words, and can be used as features for language modelling which can be easily accessed in TextBlob using the N-grams function.*
6.	*Sentiment Analysis*	*Sentiment Analysis [28] is basically a process of determining the attitude or the emotional expression of the writer, i.e., whether it is positive or negative or neutral and the sentiment function of Text-Blob [27] returns two properties, polarity, and subjectivity.*

4.3 Hydroxychloroquine and COVID-19

WHO is currently judging the use of hydroxychloroquine and remdesivir on COVID-19 patients within the early treatment. The hydroxychloroquine and remdesivir arm have been paused as a precaution for the safety data is being re-viewed.

Every country, particularly those with regulatory authorities, has advised its citizens regarding the use of any drug. Hydroxychloroquine and chloroquine are already li-censed products for treating other diseases, at this stage, but these drugs have not been found to be effective in the treatment of COVID-19. WHO has also cautioned physicians against recommending unproven treatments to COVID-19 patients, and cautioned people against self-medicating. The consensus among world experts is that the potential exists

but that far more re-searches are urgently needed to fix whether existing antiviral drugs can be operative to treat COVID-19 and if these treatments prove to be effective, they could reduce the burden of COVID-19.

The COVID-19 pandemic has spread out of control and has caused considerable morbidity and mortality in several countries including India. In this unique situation, clinicians have tried all kinds of treatments in an effort to stop the progression of this virus. One treatment that the empirical use of anti-malarial CQ/ HCQ has received huge attention. There is no strong and enough scientific and clinical data to support their use, but several countries have already included CQ/HCQ in COVID-19 treatment protocols [24],[25],[26].

5. RESULT ANALYSIS

Here all the results are calculated on the database of twitter data scraped file on specific time period. Twitter, as a social site, can express a person's own expression on a current scenario or past event or upcoming event and on the basis of the comments/reviews given by people on those twitter comments sentiment results are generated. If the review is positive, it is supposed to be good to otherwise negative for bad. Here the analysis was performed not only a particular comment, but overall comments also.

The following two graphs figure 3 and 4 below represent the comparison of the various political heads of different countries with the result and also represent the de-mand trend of HCQS as a drug/medicine for the treatment of COVID19. We can conclude that or both the two results i.e., polarity analysis and subjectivity analysis have a positivity attitude for the use of HCQS for the treatment of COVID-19. We will work further more on the opinions/sentiments using other optimization techniques for drug/medicine demand in COVID-19 pandemic.

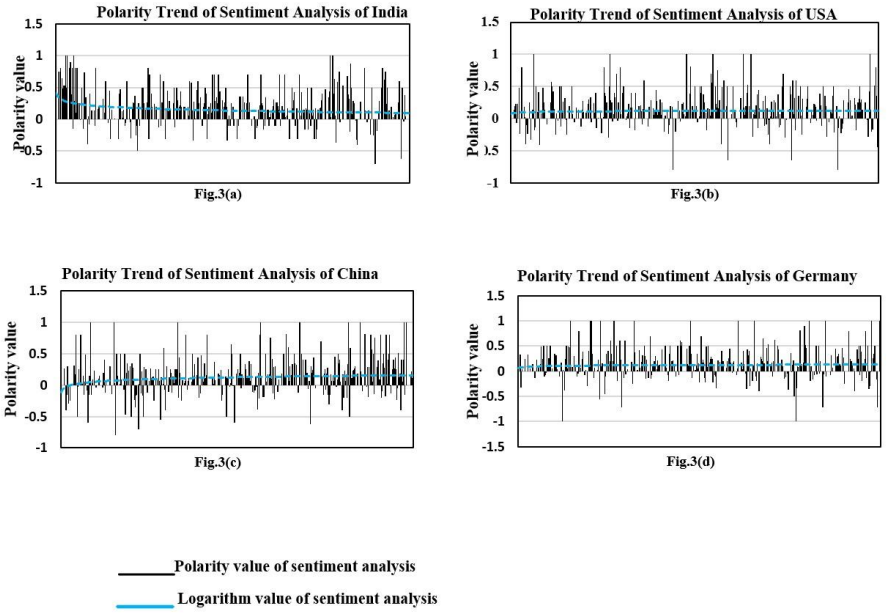

Figure 3: Polarity comparison of Tweeter sentiment analysis of political heads of different countries

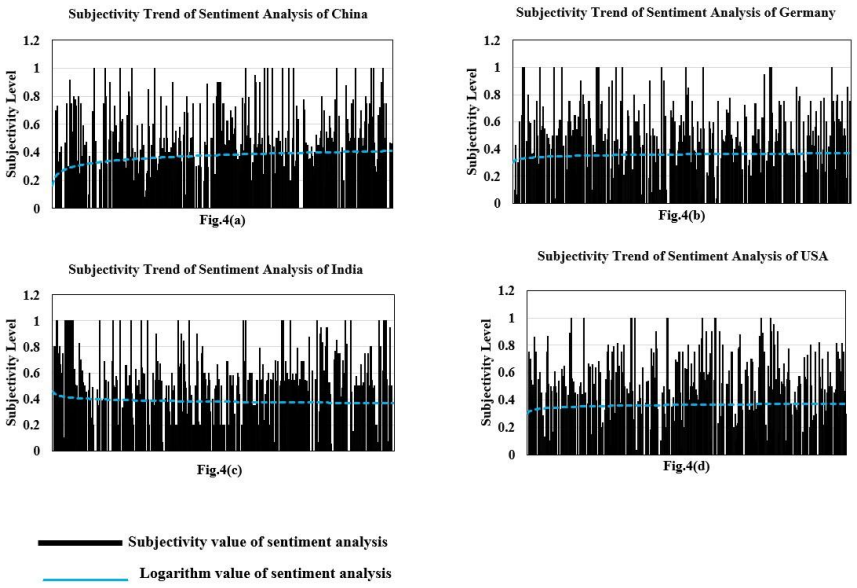

Figure 4: Subjectivity comparison of Tweeter sentiment analysis of political heads of different countries

6. CONCLUSION

In our paper on the simultaneous effect of analyzing different twitter comments of different countries' political heads are analyzed for understanding the use and demand analysis of medicine trend with respect to COVID-19. The model has worked on the views/ opinions of the twitters views of the experts affect to the traders who want to invest into the medicine market. This comparative study is done to prove the results for more accuracy. In this paper, TextBlob methods lead to the effective performance on data collected. The results show a strong co-relation be-tween twitter comments and up/down in sentiment polarity and subjectivity. Further, research work can be done in sequence to get more improved results in this area.

REFERENCES

[1] Alexander,P. , Patrick,P. : Twitter as a Corpus for Sentiment Analysis andOpinion Mining. Proceedings of LREC. 10,(2010).

[2] Rani,M., Arora,J.: A review of data analysis of twitter.International Journalof Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 5, ISSN: 2277 128X,(May.2016).

[3] P.N.Tan.:Introduction to data mining. Pearson, (2006).

[4] vetlana Kiritchenko,S., XiaodanZhu ,X., Mohammad,S. M.:Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research Vol 50,723–762,(2014).

[5] Dunn,J. C.:A fuzzy relative of the ISODATA process and its use in detectingcompact well-separated clusters. Journal of Cybernetics, 3: 32-57,(1973).

[6] Bergsma,S., Dredze,M.: Broadly improving user classification vi communication-based name and location clustering on twitter.HLTNAACL,(2013).

[7] Romero,C.,Ventura,S., Bra,P. de.,Castro,C.:Discovering prediction rules in aha! courses. Proceedings of the International Conference User Modelling, 2534, (2003).

[8] Friedmann, V.:Clustering a customer base using twitter data. CS229, (2015).

[9] Go,A.,Bhayani,R., Huang, L.: Twitter sentiment classification using distant supervision.Technical report,Stanford, (2009).

[10] Barbosa,L.,Feng,J:Robust sentiment detection on twitter from biased andnoisy data.Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36-44, (2010).

[11] Luiz, F. S., Coletta, N˙adia F. F. da Silva, Eduardo R. Hruschka, Estevam Rand Hruschka Jr.:Combining classification and clustering for tweet sentiment analysis.Brazilian Conference on Intelligent Systems,(2014).

[12] Phan,X.,Nguyen,L.,Horiguchi,S.:Learning to classify short and sparse text andweb with hidden topics from largescale data collections.In Proceeding of the 17th International conference on World Wide Web, ACM, New York, USA, 91-100, (2008).

[13]Gimpel,K.,Schneider,N.,O'Connor,B.,Das,D.,Mills,D.,Eisenstein,J.,Heilman,M.,Yogatama,D.,Flanigan,J., Smith,N.A.:Part-of-Speech tagging for twitter: annotation, features and experiments. CMU, Pittsburg, USA, (2011).

[14] Rish, Irina,R.:An empirical study of the naive Bayes classifier. IJCAI Workshop on Empirical Methods in AI, (2001).

[15] Andrew, M.:Graphical Models, Lecture2: Bayesian Network Representation. (22 October 2019).

[16] kang H., YooS.J.:Senti-Lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications Volume 39, Issue 5, Pages 6000-6010,(2011).

[17] J. M. Pena,J.M., Lozano, J.A., P. Larranaga:An empirical comparison of fourinitialization methods for the K-Means algorithm.Pattern Recognition Letters, vol. 20, no. 10, pp. 1027– 1040, (1999).

[18]Babuska,R.:Fuzzyclusteringlecture.http://homes.di.unimi.it/valenti/SlideCorsi/Bioinformatica05/Fuzzy-ClusteringlectureBabuska,(Oct2014).

[19] Amiri,I.S.,Akanbi,O.A., Fazeldehkordi,E.:A Machine-learning Approach toPhishing Detection and Defense". Syngress, (2014).

[20] Dave,K.:A Python Book: Beginning Python, Advanced Python, and PythonExercises. Section1.1,(23 June 2012).

[21] Open-Source Community. Anaconda. Retrieved 2019-02-27.

[22] Conda – Conda documentation. Retrieved February 25, 2016.

[23] Anaconda Navigator. docs.anaconda.com

[24] LenzerJ.:Covid-19: US gives emergency approval to hydroxychloroquine despite lack of evidence. BMJ 2020,369:m1335, (2020).

[25] Indian Council for Medical Research. Recommendation for empiric use of hy-droxychloroquine for prophylaxis of SARS-CoV-2 infection. https://www.icmr.gov. in/index.html,(9 April 2020).

[26] Gao,J., Tian,Z.,Yang,X.: "Breakthrough: chloroquine phosphate has shown apparent efficacy in treatment of COVID-19 associated pneumonia in clinical studies". Biosci Trends 2020;14:72–3.

[27] Biswas, S., Sarkar, I., Das, P., Bose, R., S, Roy.: Examining the Effects of Pandemics on Stock Market Trends through Sentiment Analysis. Journal of XidianUniversity.Issue 6, 1163-1176,(2020).

[28] Biswas,S., Ghosh, A. K., Chakraborty, S., Roy, S., and Bose,R.:Scope of Sentiment Analysis on News Articles Regarding Stock Market and GDP in Struggling Economic Condition. Interna-tional Journal of Emerging Trends in Engineering Research.8(7),3594-3609,(2020).