International Journal of Engineering & Technology, 7 (2.33) (2018) 619-621

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Analysis of sentiment in twitter using logistic regression

Rayasam Lakshmi [1] *, Satya,R. B. Divya [1],R. Valarmathi [2],

[1] Student, Department of Computer Science and Engineering, Sri Sai Ram Engineering College, Chennai
[2] Associate Professor, Department of Computer Science and Engineering, Sri Sairam Engineering College, Chennai
**Corresponding author E-mail: rayasamlakshmisatya@gmail.com*

Abstract

Social Platforms such as Twitter, Facebook are not always the good places and when explored there exists a dark side to it. The main objective of this research is to identify the sentiment of a tweet in twitter and also further analyse a twitter accounts activity. Logistic regression and text blob are used to identify the sentiment of the tweets, as for the taken datasets they provided the highest accuracy when compared with other algorithms such as GaussianNB, BernoulliNB, SVM. The datasets are extracted from twitter and split into training and testing data using which the model is trained to classify the sentiments of a tweet and then the analysis of a twitter account is done.

Keywords: *Sentiment Analysis; Twitter; Logistic Regression.*

1. Introduction

With the increasing popularity of the social media platforms it has become easier for people to communicate their opinions and views. Platforms such as twitter enables users to communicate their thoughts and views. Humans portray varying sentiments on a daily basis on everything and some do so by publicly putting out either their happiness, sadness, anger, frustration and various other sentiments in the form of a tweet in twitter. Accessing the datasets of tweets from twitter help us in analysing the range of emotions a user goes through a day of whichever they have posted as a tweet and then analyse it as to the number of people supporting and further propagating the different kinds of emotions and the time span of these tweets with varying sentiments. Such deep analysis of a particular individuals social media platform has many purposes even though a few drawbacks such as fake emotions on tweets and fake content does exist.

On achieving higher accuracies in such sentiment analysis of tweets this research work can be further extended to detect the expression of violence in twitter. In recent times Terrorist groups such as ISIS have panned out across such social Medias bringing in the need for more stringent rules for publishing content on such social media platforms. In the recent times many terrorist activities have been posted on social media platforms like twitter prior to the actual attack.

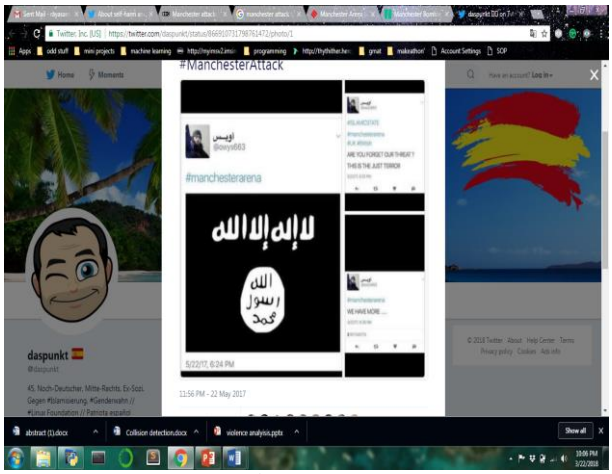

Fig. 1: An Example of a Tweet with Violent Sentiment.

2. Related work

Nakov et al [1] presented an evaluation of a Semantic analysis task in which two tasks were performed – the sentiment analysis of the social media content and the natural language processing of the text. It compares various approaches for analysing the social media content such as twitter.

Ristea et al [2] focuses on analysing the correlation between crime and social media. Using spatial correlation and spatial statistics the crime around stadiums using tweets from twitter are analysed.

Miral et al [3] analysed the raw twitter datasets and compared the results with various approaches such as naïve Bayesian, Support vector machine and Random forests.

Desmond et al [4] explored and analysed as to how street culture is translated online through the conventions of Twitter Internet threats usually ends in violence or homicide.

Iguider et al [5] proposed his work based on Shukran Sentiment Analysis system, In this, sentiment analysis is performed and po-

larity of the text is identified using Naive Bayes classifier and SentiWordNet and SenticNet. The detection and translation of language is performed using TextBlob which is a library from python for processing textual data.

3. Proposed method

In the proposed system the tweets are analyzed to identify expressions of varying sentiments. In this the datasets are first extracted from twitter and noise removal is done on the data (i.e) useless tweets such as advertisements are removed and then the datasets are analyzed.

The tweets are analyzed using the logistic regression algorithm and Text Blob for text translation and analysis. For the obtained datasets, Logistic regression algorithm gave a result of highest accuracy when compared to other algorithms such as Naïve Bayesian, Support vector machine, GaussianNB, RandomForest classifier [3]. After performing sentiment analysis and identifying if the nature of the tweet is violent or not the twitter account is analyzed for its activity. Activity such as Timeline of the recent tweets and the number of likes and retweets for each tweet and the source of creation of each tweet are represented graphically

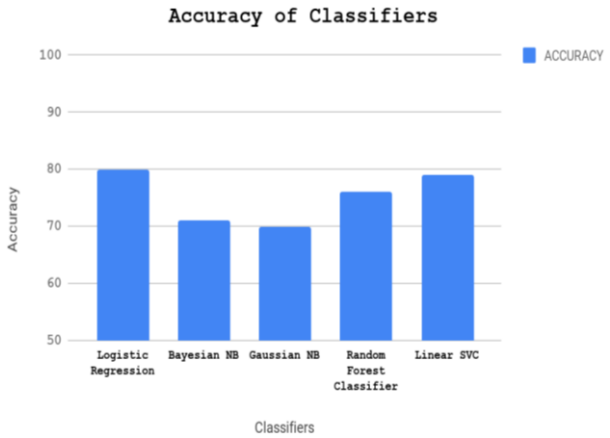

Fig. 2: Accuracy of Each Method of Analysis for the Taken Datasets.

3.1. Block diagram

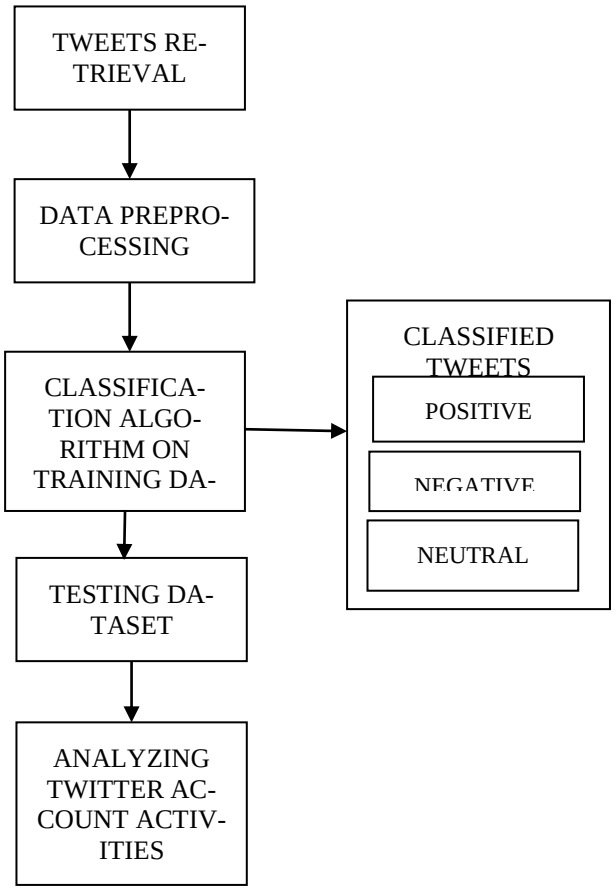

Fig. 3: Steps for Training a Classifier for Sentiment Analysis.

3.1.1. Obtaining datasets from twitter

1) The initial step is to obtain datasets from Twitter. This can be done by signing up for a developer account which further provides us with a secret access key through a web app. The Tweepy package of python uses these 4 codes obtained from the developer account to obtain the raw datasets from twitter which are saved in .csv format. A maximum of 300 tweets can be downloaded per run of the code.
2) Sign up for a developer account in twitter
3) Note down the consumer key, consumer secret, access key, access secret
4) The tweets are mined using python code with Tweepy package
5) A sample of the mined tweets:@elonmusk|961359174230773765|2018-02-07 22:01:02|b'Last pic of Starman in Roadster on its journey to Mars orbit and then the Asteroid Belt https://t.co/IWSjRyTr8V'

3.1.2. Performing sentiment analysis on the dataset

1) Identification of algorithm to use- Logistic regression with the help of blob words.
2) The Logistic regression algorithm is chosen as it yields an accuracy of 80% which is better than Naïve-Bayesian System which has an accuracy of just 70%.
3) The sentiment analysis of the tweets results in 3 values, 1-Positive, 0-Neutral, -1-Negative

3.1.3. Analyzing the twitter account

1) The twitter account which is identified of expressing violence via tweets is analyzed.
2) The particular twitter accounts source of creation of the tweets is identified which would be helpful in tracing back the location from the tweet was tweeted.

3) The timeline of tweets can also be identified and plotted as a graph and the people who are siding with the violent expression can also be identified with the number of likes of the tweet and also the number of retweets.

3.2. Experimental work

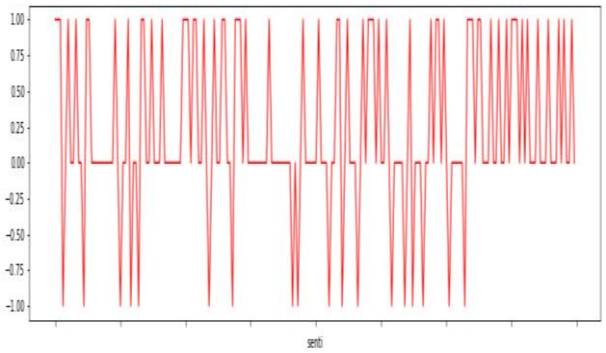

Fig. 4:. Polarity of the Sentiments of Each Tweet in the Account.

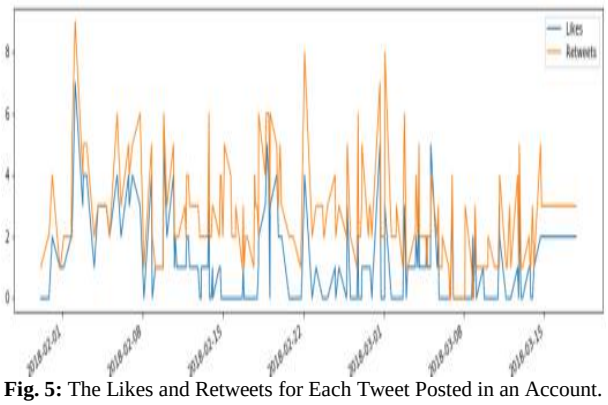

Fig. 5: The Likes and Retweets for Each Tweet Posted in an Account.

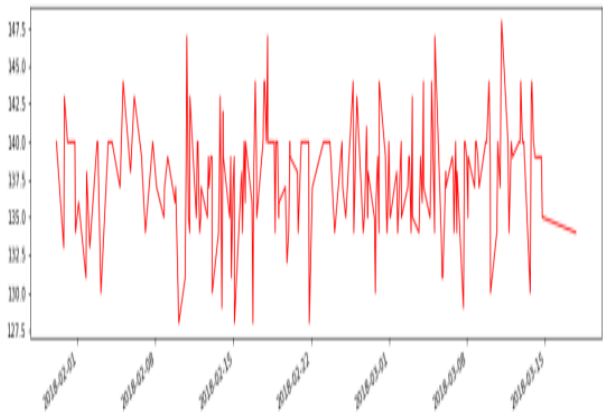

Fig. 6: The Timeline of Tweets in an Account.

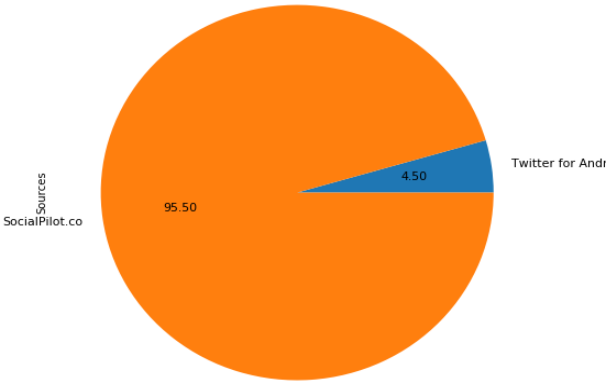

Fig. 7:. The Sources of Creation of Tweets.

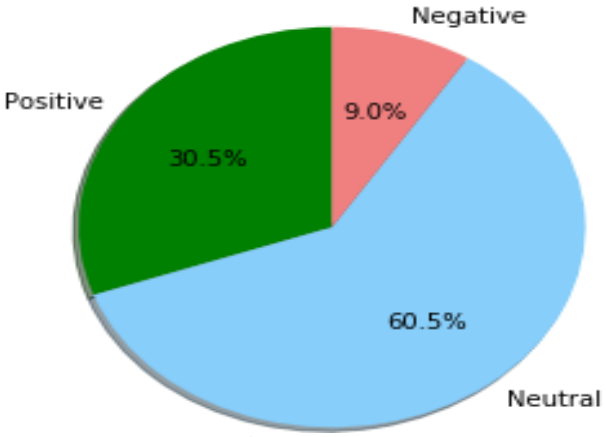

Fig. 8: The Percentages of the Sentiment in a Particular Account.

4. Conclusion

Perfecting this research work so as to achieve higher accuracies will help us to identify the sentiment of tweets which will further help in identifying if there are any violent intention n a specific tweet. There is a clear potential to be explored adopting the detection mechanisms in addition to the existing methods of detecting terrorism in social media.

There are still few challenges to overcome involving this system. Among them are the improvement of accuracy and adaptation to all languages. Despite the above identified issues, the system promises a bright future in terms of its implementation.

References

[1] Nakov, P., Rosenthal, S., Kiritchenko, S. et al. Lang Resources & Evaluation (2016) 50: 35. https://doi.org/10.1007/s10579-015-9328-1.

[2] A. Ristea, C. Langford and M. Leitner, "Relationships between crime and Twitter activity around stadiums," 25th International Conference on Geoinformatics, Buffalo, NY, 2017, pp. 1-5. doi:10.1109/GEOINFORMATICS.2017.8090933.

[3] M. Meral and B. Diri, "Sentiment analysis on Twitter," 22nd Signal Processing and Communications Applications Conference (SIU), Trabzon, 2014, pp. 690-693. doi: 10.1109/SIU.2014.6830323.

[4] Desmond, Jeffrey Lane, Patrick Leonard, Jamie Macbeth, Jocelyn R Smith LeeGang violence on the digital street: Case study of a South Side Chicago gang member's Twitter communication, Volume: 19 issue: 7, page(s): 1000-1018, https://doi.org/10.1177/1461444815625949.

[5] Iguider W., ReforgiatoRecupero D. (2017) Language Independent Sentiment Analysis of the Shukran Social Network Using Apache Spark. In: Dragoni M., Solanki M., Blomqvist E. (eds) Semantic Web Challenges. SemWebEval 2017. Communications in Computer and Information Science, vol 769. Springer, Cham.