Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

Social Media Mining for Suspicious and Offensive Keywords : Twitter

Saurabh Ruikar
Computer Engineering, MITCOE
Pune, India

Prof. Sushila Palwe
Computer Engineering, MITCOE
Pune, India

Tanmay Sangwan
Computer Engineering, MITCOE
Pune, India

Avanee Sarnaik
Computer Engineering, MITCOE
Pune, India

***Abstract*— Increasing terrorist organizations are using the internet to brainwash individuals and promote terrorist activities through provocative web pages that negatively influence the youth of the nation. The growing popularity of microblogging sites like Twitter has sparked a corresponding rise in social networking scams. An avalanche of data is generated daily. This growing microblogging phenomenon therefore allows spammers to disseminate malicious tweets quickly and massively. Trending topics capture the emerging Internet trends and topics of discussion that are on everybody's lips. In this paper, we describe a complete process to automatically collect suspect tweets according to a vocabulary of topics frequently associated with threats. The method has been validated on real datasets. Detection of tweets containing threats is a recent area of research in which most previous works had focused on the identification of malicious tweets and the application of a statistical analysis of language to detect offensive language in trending topics. In this paper we, present the first work that tries to detect opprobrious tweets in real time using language as the primary tool. We first collected and labeled a large dataset with trending topics and tweets. Then, we have conducted an extensive evaluation process that has allowed us to show how our system is able to distinguish between malevolent benevolent messages. Thus, our system can be applied to Twitter threat detection in trending topics in real time due mainly to the analysis of tweets.**

Keywords- Social networking, social media, threats, text classification.

I. INTRODUCTION

Social networking sites have become prevalent since the last decade. Lately, social media has become one of the fastest ways for individuals to communicate and exchange information. Twitter is one of the most popular online social networking and microblogging services that enables its users to send and read text-based posts of up to 140 characters, known as ‘‘tweets.” Nowadays, millions of users use Twitter to keep in touch with friends through text messages, images, audio clips, video clips, etc , to meet people and discuss about the trending issues. It is common to observe that certain individuals have especially strong influences on others. The information shared on social networking sites may contain some data which might be offensive to some people. Also, the shared media may contain some illegal information which can spread wrong message in the society. One of the most popular tools in twitter is the list of trending topics that capture the hottest emerging trends and topics of discussion.

Using this feature of twitter, people can quickly gather news about a topic or learn at a glance which are the topics on which most people speak. Unfortunately, this growing microblogging phenomenon allows spammers to disseminate malevolent tweets. Twitter provides several methods for users to report spam and these reports are investigated by Twitter and the accounts being reported are suspended in case of spam. However, reporting spam abuses using these methods is not very useful for trending topics because the suspension process is slow while the trending topics are ephemeral in most cases and they last for a few hours or a day at most.

According to recent studies, it has been observed that the increase in the use of social media is attracting more people to participate and express their point of views about a variety of subjects on a daily basis. However, there are a huge number of comments which are offensive and sometimes politically incorrect and therefore, must be obstructed from coming up online. The tweets/comments are usually made from fake, that is, spam accounts and such users are commonly known as “trolls.” This is pushing the service providers to be more careful with the contents they publish to avoid judicial claims. This work proposes the use of automatic textual classification techniques to identify and only allow to go online harmless textual posts and other content. Different sites use different methods to moderate the textual content. Twitter manually moderates the content. In manual moderation of content, manpower is required, and the moderators have to go through a lot of mental stress while moderating the data. For example, they might be forced to view content that is inhumane which will lead to severe post-traumatic stress disorder. Thus, manual moderation of abusive content is malicious for the person moderating the content as it causes harmful effects on them. Therefore, there is a need for an efficient technique to monitor hate speeches and offensive words on social networking sites which may be a threat to others.

II. PROPOSED SYSTEM

On Twitter, every user communicates through messages commonly known as tweets. The intake of these tweets will

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

be done through **username**, **hashtags** and **keywords**. The username type and the tweets are displayed to the person handling the account. The tweets are extracted independently as per user's choice. These extracted tweets are then stored separately in a json file respectively. Twitter allows us to access tweets using their API with the Tweepy library. However, it only allows access to tweets that go back to two weeks earlier than current date. For example, tweets from about a month ago will not be accessible.

As mentioned above, the tweets are received through three modes. These modes, i.e., username, hashtag and keywords are trained to check if the tweets received are threats or not. The model classifier is basically made of training set and test set. The model is trained in such a way that the training data (hate speech dataset) determines whether the data is malicious or not. Each individual mode is to be presented. A single username maybe associated with hundreds of tweets. Our model will segregate and find those tweets and accordingly divide them in two categories namely: offensive or non-offensive. Offensive tweets containing threats or disparate messages are labelled as **1** whereas normal tweets are labelled **0**. The below table represents a sample dataset in which tweets are segregated accordingly.

	A	B	C
1		Holy shit, Freddie Highmore was in Charlie and the Chocolate Factory"Me: *rolls on the floor, laughing*	1
2		" I got hicks lol	0
3		This new twitter is confusing the shit out of me." Go back to south america	1
4		As a woman you shouldn't complain about cleaning up your house. as a man you should always take the trash out...	0
5		I want to kill you and set your house on fire..	1
6		Teanna Trump probably cleaner than most of them but....."	1
7		Let me eat a Oreo do these dishes." One oreo? Lol	0
8		I been kidnapped and assaulted.	1
9		Harlem shake is just an excuse to go full retard for 30 seconds."	1
10		god teeth , gold chainwhite , cocaine	1
11		Twitter is not a source of news broadcast you stupid retards" it is when nbc, abc, cnn, Fox, etc are all keeping quiet.	1
12		firefighter is a job for white trash	1

Fig 1: Example of Dataset

Once we have data in the form of tweets, we need to filter that data which in other words is known as data preprocessing. As shown in the figure below, data preprocessing takes place with the use of multiple steps. The first step involves breaking down the tweet into tokens. This is achieved by using the tokenize module available in Python. Now that the tokens are generated, stemming and stop words removal takes place. Stemming helps us find the root node of a word by removing the prefixes or suffixes whereas stop words removal is the technique used to remove all the unnecessary words in a tweet. Any word that will not be needed by classifier to find the context of the tweet is removed during stop words removal process. Next is the Lemmatization process which works in grouping of inflected words such that the root word is present in the language. For example run, ran are forms of the word run. Thus we can say run is the lemma of these two words. After all the above processes for preprocessing the data, the tokens are converted back to strings using the detokenize module in Python. This collection of twitter strings is then converted into vectors which are used to help us in training our model classifier.

During training, classifier model is constructed from the vectorized sentences prepared by data mining component and label (Offensive/Normal) which are already present in the dataset. Further, this trained classifier model is used for predicting a given sentence whether it's offensive or not thereby helping us check if the tweet is a threat or not. Classifier predicts the outcome accurately and precisely. For this purpose, we have chosen to use the Random Forest classifier algorithm.

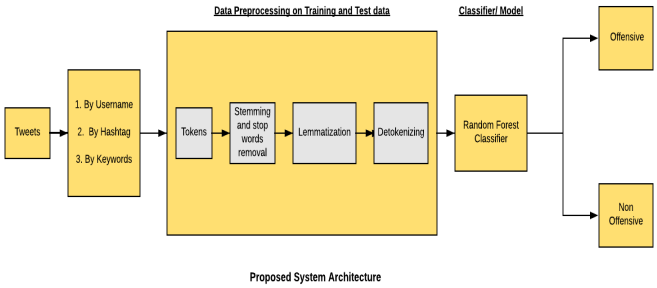

Fig 1.2: Proposed System Architecture

III. MATHEMATICAL MODEL

Here, we use the Random Forest classifier model for performance comparison. Random forests, also known as random decision forests, are a widely used method in which many decision trees are constructed at a time and then merged together to get a more precise and accurate prediction. Random forest aims to reduce the correlation issue by choosing only a subsample of the feature space at each split.

It can be mathematically represented as:

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T}$$

Several problems are overcome with Random Forest including:

1. Less Variance
2. Reduces Overfitting

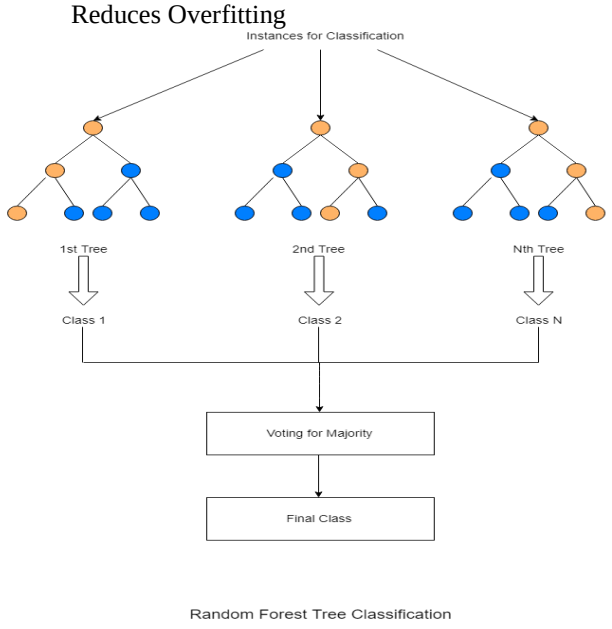

Fig 2: Classification of Random Forest

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

RESULT

The dataset which was subjected to the tests was modified accordingly and is fit for further classification. The dataset was initially divided into two categories: **malicious tweets** and **normal tweets**. Those two categories are used for the experiment. The system appropriately distinguishes between the two. The proposed model is implemented with data pre-processing in order to obtain results. Following graph shows the comparison of various predictive metrics for 2 models which are used for the Training.

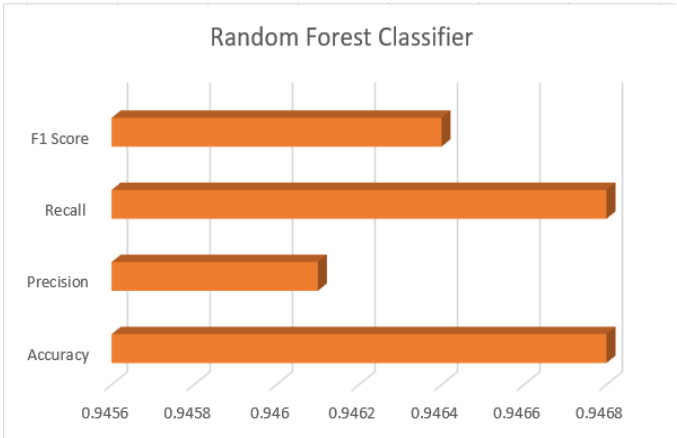

Fig 4: Predictive metrics comparison graph

Results	Random Forest Classifier
Accuracy	0.9468876141916295
Precision	0.9461987731815997
Recall	0.9468876141916295
F1	0.9464790348593033

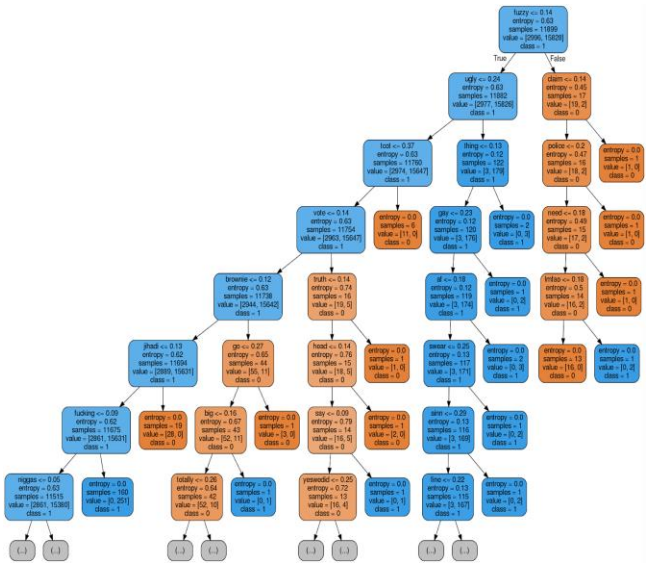

FIG 3: SAMPLE DECISION TREE ESTIMATOR VISUALIZATION

CONCLUSION

In this system, a new methodology using web mining and data pre-processing is used to detect portent tweets in Twitter trending topics. Our study is based on the analysis of the language used in each tweet, to identify those tweets whose purpose is to spread opprobrious messages. It mainly focuses on categorizing text data in two categories namely malevolent and benevolent. Two tools that are available are the 140 characters in a tweet and the linked pages. In addition, because of growing microblogging phenomenon and trending topics, spammers can disseminate malicious tweets quickly and massively. Our project therefore aims at establishing a better and peaceful internet culture. With our proposed system in action, Twitter can be a much safer environment for everyone to use and will lower the instances of hate speech and threats in the form of tweets or retweets significantly.

FUTURE SCOPE

We clearly acknowledge the limitations of our analyzed dataset and consider that it can further be enhanced and used on a national level which will enable the government to keep track of the suspicious activities. The system can be further expanded to understand regional language. A geotag feature used for IP tracing can be added to get much more effective and accurate results. The overall precision would also reduce the total time taken. Despite this, we also believe that the scientific community should work towards finding a common evaluation framework which compares the threat detection systems.

ACKNOWLEDGMENT

It gives us great pleasure to present a seminar on “Social Media Mining for Hate Speech and Offensive Keywords Prominence.” In preparing this seminar number of hands helped us directly and indirectly. Therefore, it becomes our duty to express our gratitude towards them. We are very much obliged to subject guide Prof. Sushila Aghav, in Computer Engineering Department, for helping and giving proper guidance. Her timely suggestions made it possible to complete this seminar for us. All efforts might have gone in vain without her valuable guidance. We also express a great sense of gratitude to the Director Prof. (Dr.) R. V. Pujeri, Principal Prof. (Dr.) A. S. Hiwale, Head of Computer Engineering Prof. Bharati Dixit and the entire staff members in Computer Engineering Department for their cooperation.

REFERENCES

[1] Davidson, Thomas and Warmsley, Dana and Macy, Michael and Weber, Ingmar. ”Automated Hate Speech Detection and the Problem of Offensive Language”. In proceedings of the 11th International AAAI Conference on Web and Social Media 2017, (Pg. 512-515).

[2] Scikit-learn: A module for machine learning. https://scikit-learn.org [Access Date: 19 Dec 018].

[3] Kumar Abhishek, Hiteshwar Kumar Azad, Semantic-Synaptic Web Mining: A Novel Model for Improving the Web Mining in 2014 Fourth International Conference on Communication Systems and

[4] Aakash Negandhi, Soham Gawas, Prem Bhatt, Priya Porwal, Detect Online Spread of Terrorism Using Data Mining in IOSR Journal of Engineering, volume 13 (2018).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

[5] P. Sampath, C. Ramesh, T. Kalaiyarasi, S. Sumaiya Banu, G. Arul SelvanAn, Efficient Weighted Rule Mining for Web Logs Using Systolic Tree in IEEE- International Conference On Advances In Engineering, Science And Management (ICAESM -2012) March 30, 31, 2012.

[6] S.Gowri, G.S.Anandha Mala, G.Divya,Suspicious data mining from chat and email data in International Journal of Advances in Science Engineering and Technology, ISSN: 2321-9009 Volume- 2, Issue-2, April-2014.

[7] Amber Sinha, Social Media Monitoring,The Centre for Internet and Society, India. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] T. Ho, "The random subspace method for constructing decision forests, " IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832-844, 1998.