International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Trends Manipulation and Spam Detection in Twitter

Dr. P. Maragathavalli [1], B. Lekha [2], M. Girija [3], R. Karthikeyan [4]
[1, 2, 3, 4] *Information Technology, Pondicherry Engineering College, India*

Abstract: ***Social network has emerged as a very popular way for the users to communicate and interact online. Users spend plenty of time on Twitter by reading news, discussing events and posting messages. Data is collected using the live public tweets in real time, twitter trends are manipulated by using Naive Bayes classifier. This trending topic also attracts a significant amount of spammers who continuously expose malicious behaviour leading to great misunderstanding and inconvenience on users social activities. Spammers pose tweets either as advertisements, scams and help perpetrate phishing attacks or the spread of malware. In building an online spam-detection service, that would take the URL of a tweet and categorize it as spam or not. To classify tweets into spam and non-spam, the tweets are being applied into Random Forest Classifier. It tries to detect, is a tweet is spam or not based on properties of the tweet like the presence of some keywords, types of hash tags present. The proposed work shows that it is capable to provide better performance with twitter trending topics and true positive rate of spam and non-spam.***

Keywords: ***Classification, Hashtag, Social Network, Spam messages, Twitter trends.***

I. INTRODUCTION

Online social networks (OSNs), which include Twitter, Facebook, and other social network , have emerge popular within few years. Twitter is a social networking and micro-blogging service. Users post updates called tweets containing up to 280 characters of text and HTTP links. Twitter's main page features a regularly update a list of trending topics in which tweets are instantly exploded in volume. Trending topics tells us about what people are attracted to, what people think is important and what people make the effort to pass on or share. At each people in the network, the probability of transmission and the number of connections to other people determines how fast and how far the information spreads.

Twitter supports a hashtag annotation format [10] so that users can indicate what their posted messages are about. This general "topic" of a tweet is, by convention, indicated with the hash sign, #. Twitter users often use hashtags to identify the topic of the messages. Hash tags that appear in tweets at a given time in Twitter's list of trending topics. Topics that are very popular in real time are called "trends" on Twitter. A trend on Twitter [10] consists of one or more terms and a time period, such that the volume of messages posted for the terms in the time period exceeds some expected level of activity .Trends were initially created from the most used hashtags, but now Twitter also shows keywords as part of the trends. Twitter trends are dynamic and change depending on the tweets being created at real time. The way Twitter finds trends from published tweets has not been disclosed by them but they have stated that trends are generated by their proprietary algorithm, which tries to find the topics that are being mentioned currently in tweets more than they were being mentioned in the past that is, their rate of occurrence in the tweet stream is increasing. They have explained that a topic can start trending when the number of times it is being mentioned increases suddenly and dramatically. This also means that highly popular all-time topics may not be included in the trends. In other words, Twitter trends prioritize novelty over popularity to determine trends.

According to Twitter, trends were designated to help the user find breaking news and hot emerging topics on social media in real time. Trends are showcased in the Twitter homepage after signing in, so the trending topics/hashtags are given the maximum visibility. Thus trends may change the perception of the individual user, in the sense that the user may be more inclined to tweet about topics that are extremely popular at that moment. As the Twitter user community as a whole has the ability to change trending topics, so do trends with regard to convincing individual users to tweet about a trending topic. The trending topics are valuable for informing users of current trends. The processing steps involved in classifier in Fig 1.

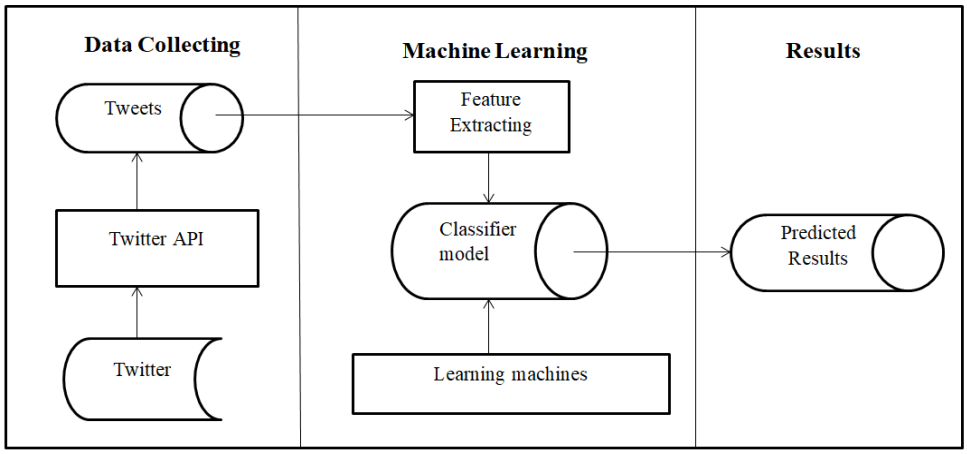

Fig. 1 Processing steps involved in classifier

A spam is an unwanted data that a web user receives in the form of messages. This spamming is actually done by sending unsolicited bulk messages to indiscriminate set of recipients for advertising purpose. These spams can also be used for some attacks which is used to destroy user's information or reveal his identity or data. Spam messages are the messages that the receiver does not wish to receive. Increasing volume of such spam message is causing serious problems for internet users, Internet Service Providers, and the whole Internet backbone network.

In response to come across Twitter junk mail, there were some works brought. Most of these works are making use of machine getting to know set of rules to separate spam and non-spam. Some initial works, together with , made use of account and content capabilities, which include account age, quantity of followings, URL ratio, and the length of tweet to distinguish spam and non-spam. Extracted the distance and connectivity among a tweet sender and a receiver to decide whether or not the tweet is spam or not. However, amassing these capabilities are very time-consuming and useful resource-consuming, because the Twitter social graph is extraordinarily large. When the behaviours of spammers are analysed within the scope of tweet-based features, these facts are observed: Spammers tend to use links to direct legitimate users to their malicious purposes, Spammers tend to use lots of mentions to attract the attention of more legitimate users, Spammers tend to use lots of hashtags to reach more users, Since spammers' tweets are unsolicited, the number of likes and retweets their tweets have received are much lower compared to legitimate users.

II. EXISTING WORK

Twitter trends, a timely updated set of top terms in Twitter, have the ability to affect the public agenda of the community and have attracted much attention. Unfortunately, in the wrong hands, Twitter trends can also be abused to mislead people. Zhang et al.[1] attempt to investigate whether Twitter trends are secure from the manipulation of malicious users and the twitter trends manipulation by using SVM classifier. With the datasets collected via Twitter API, Twitter trending topics are manipulated. Then, by employing the classifier to explore how accurately factors at the topic level could predict the trending. A linear influence model to capture the network impact on the diffusion of a topic in Twitter in order to find the evidence of manipulation. The application of the model is limited to linear scenarios.

Kathy Lee et al. [9] Although Twitter provides a list of most popular topics people tweet about known as Trending Topics in real time, it is often hard to understand what these trending topics are about. Therefore, it is important and necessary to classify these topics into general categories with high accuracy for better information retrieval. In text-based classification method, by constructing word vectors with trending topic definition and tweets, and the commonly used weights are used to classify the topics using a Naive Bayes Multinomial classifier. In network-based classification method, by identifying top 5 similar topics for a given topic based on the number of common influential users. Experiments on a database of randomly selected 768 trending topics show that classification accuracy of up to 65% and 70% can be achieved using text-based and network-based classification modelling respectively. Supervised learning techniques are used to classify the twitter trending topics. They downloaded trending topics and definitions every 30 minutes from what the Trend and all tweets that contain trending topics from Twitter while the topic is trending. All the

tweets containing a trending topic constitutes a document. Our results show that network-based classifier performed significantly better than text-based classifier on our dataset.

In [2] the author has discussed about the statistical features based on detection techniques. Recent works focus on applying machine learning techniques for Twitter spam detection, which make use of the statistical features of tweets. In the labelled tweets data set, however, that the statistical properties of spam tweets vary over time, and thus, the performance of existing machine learning-based classifiers decreases. This issue is referred to as "Twitter Spam Drift". The proposed scheme can discover "changed" spam tweets from unlabelled tweets and incorporate them into classifier's training process. Recent works focus on applying machine learning techniques for Twitter spam detection, which make use of the statistical features of tweets. The proposed scheme can discover "changed" spam tweets from unlabelled tweets and incorporate them into classifier's training process. LDT is used to deal with a classification scenario where there is a sufficiently robust algorithm, but in lack of more data. It can not only eliminate unuseful information in the training data but also make it faster to train the model as the number of training samples decrease. The benefit of "old" labelled spam is to eliminate the impact of "spam drift" to classify more accurate spam tweets.

In [3] the author has discussed about the Performance, Stability and Scalability of twitter spam detection .The current solutions fail to detect Twitter spams precisely and effectively .By comparing the performance of a wide range of main stream machine learning algorithms, aiming to identify the ones offering satisfactory detection performance and stability based on a large amount of ground truth data. The performance study evaluates the detection accuracy, the true/false positive rate and the F-measure, the stability examines how stable the algorithms perform using randomly selected training samples of different sizes. The scalability aims to better understand the impact of the parallel computing environment on the reduction of the training/testing time of machine learning algorithms.

In [4] Twitter Spammer Detection introduce features which exploit the behavioural-entropy, profile characteristics, spam analysis for spammer's detection in tweets. By taking a supervised approach to the problem, but leverage existing hash tags in the Twitter data for building training data. Spammer tweets pose either as advertisements, scams and help perpetrate phishing attacks or the spread of malware through the embedded URLs. In this project, twitters tweets are fetched for a particular hashtag. Each hashtag may have 1000 of comments and new comments are added every minute, in order to handle so many tweets we are using twiter4j API and perform pre-processing by removing quotes, hash symbols and spam analysis through URL.

In [8] the author has discussed about the spam detection by using Random Forest Classification. This paper described classification of emails by Random Forests (RF) Algorithm. RF is ensemble learning technique. The Random forest is a meta-learner which consists of many individual trees. Each tree votes on an overall classification for the given set of data and the random forest algorithm chooses the individual classification with the most votes. If identified category is 0 then e-mail is marked as non-spam e-mail otherwise if identified category is 1 then e-mail is marked as spam e-mail. The existing work collected the dataset for particular time period and analyse the result using the classifier. We are interested in building an online spam-detection service that would take the URL of a tweet and categorize it as spam or not. The data is collected in real time using streaming API and manipulate the trends and the spam will detected in real time by collecting tweets.

III. PROPOSED WORK

The main objective of the paper is to manipulate the trend and spam detection. Tweets are extracted in a streaming way and Twitter offers the Streaming API for researchers to get entry to public tweets in real time and perform preprocessing by removing quotes, hash symbols and spam analysis through URL. Based on the frequency of the topic used along with hashtag, endogenous factors are calculated. Labeled tweet data set are trained using Naive Bayes classifier in order to Trends Manipulation. The output from the trends manipulate are collected. Labeled tweet data set are trained using Random Forest classification. Classify unlabelled tweets based on trained dataset. Add Spam label to labelled data set. The spam tweets are added to labelled data set based on some keywords that classified. Re-train the classifiers, In order to detect spam tweets from the labelled tweets. By using Random Forest Classification unlabelled tweets are tested to detect the spam and non-spam.

Real-Time Detection of Twitter Trends Manipulation and Twitter Spam was identified using Naive Bayes and Random Forest classification. The proposed work is shown in Fig 2.In building an online spam-detection service, that would take the URL of a tweet and categorize it as spam or not. They try to detect is a tweet is spam or not based on properties of the tweet like the presence of some keywords, types of hash tags present etc.

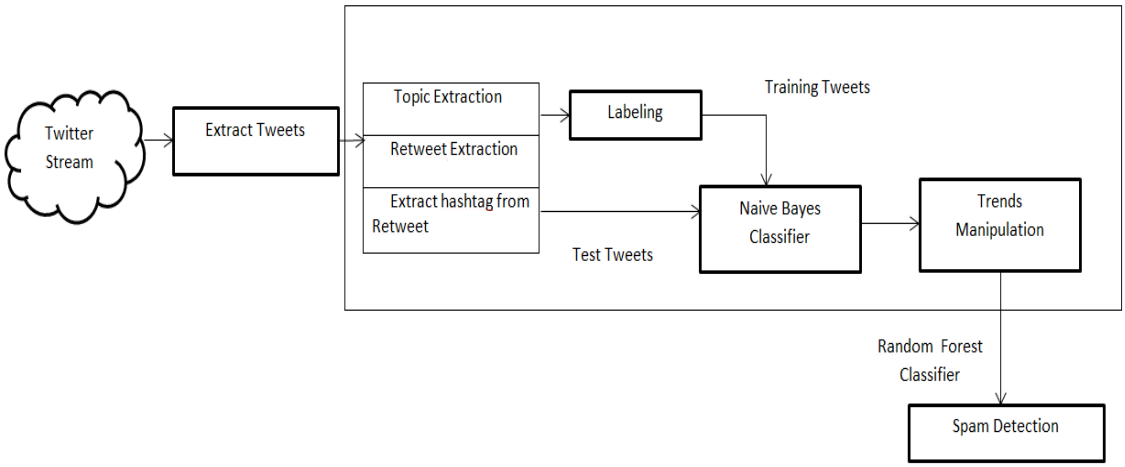

Fig. 2 System Architecture

A. *Twitter Stream*

Tweets[4] are retrieved in a streaming way and Twitter offers the Streaming API for developers and researchers to get entry to public tweets collected. Each hash tag may have 1000 of comments and new comments are added every minute, in order to handle many tweets T_t shown in Table.1,we are using twiter4j API.

B. *Extract Tweets*

By using this method tweets are extracted from Twitter using Twitter4j API in java based on user input. The input considered for the proposed module is query keywords. The output considered for the proposed module is Extracted Tweets. From the collected tweets, topics are extracted based on hash tag (#). The input considered for the proposed module is Tweets. The output considered for the proposed module is Topics. The topics are from the tweets based on non-linear model. It does not use fixed time window to extract the topic.

C. *Retweet Extraction*

The output from extract tweets are collected. From the collected tweets Re tweets are extracted based on RT symbol. The topics are extracted from the re tweet module. The hash tag usage of re tweets are considered for this module. The endogenous factors T_f is calculated by considering both public tweets and re tweets. The occurrence of topics with hash tags are considered, in order to find the trend detection by using the naive bayes classifier.

D. *Trends Manipulation*

Once the data was gathered, our next task was to develop a collection of tweets labelled into training and testing tweets categories. For such a collection to be useful, it had to contain adequate numbers of tweets, include tweets from a range of times and topics, and be as unbiased as possible given these constraints.

Naive Bayes Classifier is used for our approach. It is one of the widely popular Supervised Machine Learning algorithms used for text classification, and clearly provides less computational time from other Supervised Machine Learning Techniques.

Naive Bayes Classifier [12] primarily works on the conditional probability theory. It offers the assumption of a particular feature from a class of features. But not necessarily, it will come out as the accurate one.

The word naive indicates the novice assumption of each conditions. For any given case, where something else has already occurred, in that scenario, by using the conditional probability, which is the base of Naive Bayes algorithm, we can calculate the probability of occurrence of a future event using the prior knowledge gained from the previous sample case. Naive Bayes Classifier is the type of classifier, which predicts all possible membership probabilities for each class, i.e. the probability that is already recorded for a particular class. Naive Bayes is one of the most popular methods, despite of its relatively easier approach to implementation, it is fairly complicated, and often outperforms most other complicated algorithms in performance. The trends are manipulated by using naive bayes classifier.

Notation	Data Source	Selection for analysis
T_t	Twitter own trends as retrieved from the Twitter API	Selected from complete set of trends published by Twitter
T_f	Trends computed from raw Twitter data using term frequency measures	Selected from top-scoring terms

Table. 1 Analysis of data for tweet collection

E. Spam Detection

Twitter spams usually refer to tweets containing advertisements, messages redirecting users to external malicious links including phishing or malware downloads. Spams on Twitter not only affect the online social experience, but also threatens the safety of cyberspace. To classify the majority of the dataset, we employed random forest classifier. These algorithms are first trained on the labelled data to develop classification models that are then applied to unlabelled data to predict which tweets are in the spam class and which are in the non-spam class. There are particular words used in spam emails and non spam emails. These words have particular probability of occurring in both emails[12].

The spammers might target topics that rank higher. For the purpose of analysis, Spam Incidence is the percentage of tweets our classifier labelled as spam divided by the total number of tweets. The spammers might target topics that stays in the trending topics list longer, since that the longer a topic stays, the more likely it will catch spammers' attention.

The presence of URLs was a key attribute, since most spammers display messages with the hope of attracting users to follow a link. The fact that the number of words and the number of characters provided essentially no predictive power was also not surprising given the diversity of both spam and non-spam messages in Twitter. As expected, spam messages had URLs with much higher frequency, and perhaps more numeric characters as a result of that in combination with monetary values. Spammers used hashtags more often than regular users, perhaps as a way to ensure their messages would be grouped with the trending topics. Interestingly, spam messages targeted topics with a lower mean ranking than non-spam messages.

.The performance analysis is calculated was shown in Fig.4.

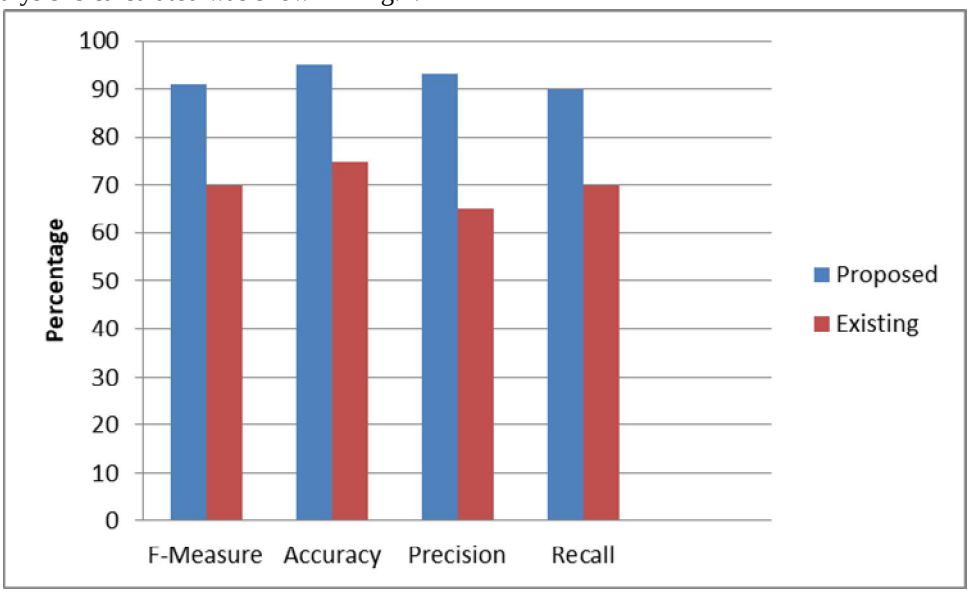

Fig.4 Performance Analysis

IV. CONCLUSION

This paper, Real-Time Detection of Twitter Trends Manipulation and Drifted Twitter Spam was developed using Naive Bayes and Random Forest classification has been proposed in order to identifying the twitter trend and twitter spam. Based on the frequency of the topic used along with hashtag, endogenous factors are calculated. We employed Twitter's streaming API to gather over a million tweets on hourly trending topics. Processing the raw data from the API, we extracted tweet features relevant to spam detection as identified by previous research. We trained a naive Bayes classifier for tweet classification on a hand-labelled random sample of nearly 1000 tweets and verified its effectiveness. Filtering the trending topic data with this classifier, we obtained results on the prevalence of spam overall, between topics, and the effect of spam on topic ranks. Labelled tweet data set are trained using Naive

Bayes classifier in order to Trends Manipulation. The output from the trends manipulate are collected. With the goal of achieving twitter trend manipulation we further evaluated the algorithms in terms of the security to detect spam from the trending topic tweets. Labelled tweet data set are trained using Random Forest classification. Twitter has not been able to solve the problem of spam perfectly, which is the reason trends on Twitter are susceptible to manipulation and misinterpretation which can be very harmful to the society as a whole.F1 scores of the classifiers on both the training and test sets. A deeper examination of our results revealed that RF was predicting 90% of the test set. Spam frequency in the trending topics seemed to correspond with previous results suggesting a spam rate in Twitter messages. Re-ranking of topics after applying our spam filter changed existing rankings very little. Furthermore, users of Twitter can take precautions regarding spam when certain topics are involved. Random Forest performed better than other classifier according to precision

REFERENCES

[1] Yubao Zhang , Xin Ruan,Haining Wang,Su he "Twitter Trends Manipulation:A first look inside the security of twitter trending"in IEEE Transactions on information forensics and security,2017,pp.1-12.
[2] Chao Chen, Yu Wang, Jun Zhang, Yang Xiang, Wanlei Zhou and Geyong Min, "Statistical Features-Based Real-Time Detection of Drifted Twitter Spam", IEEE Transactions On Information Forensics and Security,2017,pp.1-12.
[3] Guanjun Lin, Nan Sun, Surya Nepal, Jun Zhang, Yang Xiang,"Statistical Twitter Spam Detection Demystified: Performance, Stability and Scalability",IEEE Access,2017,pp.1-13.
[4] Shradha Hirve, Swarupa Kamble ,"Twitter Spam Detection", International Journal of Engineering Science and Computing,2016,pp.1-5.
[5] C. Chen, J. Zhang, X. Chen, Y. Xiang, and W. Zhou, "6 million in Spam tweets: A large ground truth for timely twitter spam detection," in IEEE Communication System Security Symposium, 2015,pp.1-6.
[6] Cho Chen,Jun Zhang,Yi xie,Yang xiang,Wenlai zhoi,Majed alrubaian,"A Performance Evaluation of machine learning based streaming Spam Tweets Detection",IEEE Transactions on computational social systems,2015,pp. 1-
[7] Grant Stafford,Louis Li Yu,"An Evaluation of the Effect of Spam on Twitter Trending Topics", International Conference on Social Computing,2015,pp.1-8.
[8] Bhagyashri, Gaikwad, Halkarnikar," Spam E-mail Detection by Random Forests Algorithm", International Journal of Advanced Computer Engineering and Communication Technology,2013,pp.1-8
[9] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, "Twitter Trending Topic Classification", IEEE Computer Society,2011,pp. 251-258.
[10] Mor Naaman , Hila Becker and Luis Gravano,"Hip and Trendy: Characterizing Emerging Trends on Twitter", Journal of the American Society for Information Science and Technology,2011,pp. 902-918.
[11] Sourav Das and Anup Kumar Kolya," Sense GST: Text Mining & Sentiment Analysis of GST Tweets by Naive Bayes Algorithm ",IEEE Computer Society,2017.
[12] Rekha , Sandeep Negi ," A Review on Different Spam Detection Approaches ", International Journal of Engineering Trends and Technology,2014.
[13] Luca Maria Aiello, Georgios Petkos," Sensing trending topics in Twitter",IEEE,2013,pp 1-14.