An Integrated approach for Malicious Tweets detection using NLP

Sagar Gharge
Dept. of Computer Science and Engineering,
Walchand College of Engineering,
Sangli, India,
email: sagar.gharge@walchandsangli.ac.in

Mr. Manik Chavan
Dept. of Computer Science and Engineering,
Walchand College of Engineering,
Sangli, India,
email: manik.chavan@walchandsangli.ac.in

Abstract— Many previous works have focused on detection of malicious user accounts. Detecting spams or spammers on twitter has become a recent area of research in social network. However, we present a method based on two new aspects: the identification of spam-tweets without knowing previous background of the user; and the other based on analysis of language for detecting spam on twitter in such topics that are in trending at that time. Trending topics are the topics of discussion that are popular at that time. This growing micro blogging phenomenon therefore benefits spammers. Our work tries to detect spam tweets in based on language tools. We first collected the tweets related to many trending topics, labelling them on the basis of their content which is either malicious or safe. After a labelling process we extracted a many features based on the language models using language as a tool. We also evaluate the performance and classify tweets as spam or not spam. Thus our system can be applied for detecting spam on Twitter, focusing mainly on analysing of tweets instead of the user accounts.

Keywords— Spam detection, Social network, Statistical natural language processing, Machine Learning.

I. INTRODUCTION

Social networking sites now-a-days have become the most popular medium for people to spend their time and make more friends. Online Social networks (OSNs), such as Twitter, Facebook, and many enterprise social networks, have become very much popular in the last few years. People are spending a huge amount of time on OSNs making friends, with people whom they know or with people they are interested in. Twitter, which came into existence in 2006, became one of the most popular site. Now-a-days, over 300 million Twitter users create around 600 million new micro blogs which are called as "tweets" per day. Twitter now has become the most popular online micro blogging service, that encourages its users to send and read image based and text based posts of up to 140 characters. Now-a-days, billions of users use Twitter to be in contact with their friends and followings. On social networking sites, people share and spread a lot of knowledge and information which proves beneficial to others and help them in many ways. But you never know who is keeping track of your information and can use that information against you with a malicious intension. There are many people who share their personal information on social networking sites daily and such people fall prey to some or the other malicious activities and their personal information is misused

Now-s-days, social networking sites are a medium of analytics on a large amount of user data for many companies based on which many prediction models are being built, and recommended systems are been prepared for end users. But this is only the single side of a coin. On the other side there are many such people who are waiting for just a single chance of a single wrong step by the user. Social networking sites have become a platform for the promotion of businesses, classes and may more activities, but some people are using it in a very wrong way. They provide the users with links in a promotional form and redirect them to many misleading and misguiding sites where, the content is either malicious or in such a way that it will change their thoughts negatively. The main problem which comes as "spam" is here, "spam" is a form of content which is irrelevant or unsolicited message which is sent with a purpose of advertising, phishing, spreading malware etc. There arises a big threat of malware. Many people fall prey to spams and are redirected to sites from where a malware is been downloaded to their personal machines, which is continuously sending their valuable information to an unknown server.

Twitter is a place where more such spammers are active and they are circulating spams continuously. Hence, there is a need to detect such spams, which contain URLs and are redirecting people to different locations. The work in this paper deals with detecting such spam tweets and many new features based on language models that help to improve spam detection. The underlying idea of the paper is as follows: we analyse the use of languages in the tweets and the page of where the URL is linked to it. In the scenario of spam the language models differ from each other. Hence, the role of the person posting the spam i.e. "spammer", is to divert people to another content which is totally irrelevant. We analyse the divergence ratio between the language models in order to classify the models as spam or not spam.

978-1-5090-5297-4/17/$31.00 ©2017 IEEE

The remaining of the paper proceeds as follows: Section 2 presents the previous works in the spam detection research area; Section 3 shows the architecture of the system; Section 4 is devoted to the proposed methodology; Section 5 describes the observation and analysis; Finally, Section 6 draws the conclusions.

II. Related Work

The work in [1] focuses on the streaming data on twitter which is classified as spam or not. Twitter APIs which are available are used to collect data from twitter and then analyses it based on the machine learning methods. The work in [2] focuses on the use of enterprise social networks. The use of enterprise social media by the employee is analyse according to the fact that, what is beneficial to employee and what is not. The paper focuses on the factors influencing the decision of employee to use the social networks for their work purpose. Work in [3] shows the detailing of analysing the information related to the sentiments of people which is captured during the twitter data gathering. The twitter data has become the area of wide research through which the people can analyse the sentiments of the user. It proposed many algorithms that classify the sentiments from the tweets into a data stream and to decide whether the tweet specifically is subjective or objective, that is positive or negative sentiments are classified.

For a given set of documents it becomes very difficult to find similar kinds of sentences and key concepts within the sentences. The work in [5] focuses on such analysis of documents, which crawls through the document and then analyses and identifies the type of sentences within it and then the way in which sentences show their uniqueness. The problem of spam on web is growing day by day. The paper [6] focuses on scenarios of such spam detection based on various classifiers, which analyse the URLs in the content and crawls the pages associated to URLs to identify the divergence ratio from the original content to the content on the page where the URL redirects you to.

The researches on smoothing techniques are very less. The technique focuses on analysing the various factors of the data being used. Paper [7] analyses the factors such as training data size, test data size etc. It calculates the variations in the existing methods. Cyber bots are increasing day by day as the world is moving towards automation in social media too. Some bots which are reliable and legitimate, they focus on doing the daily task for which they are programmed to but some bots designed by humans who are ready to violate the privacy of some users, are used to peep into other users data and secretly retrieve the information. The analysis and difference between these two types of bots is shown in [8]. By referring [9] we get to know the concept of relativity between contents and events. Event relatedness is analysed and the differences between events are presented.

III. Architecture

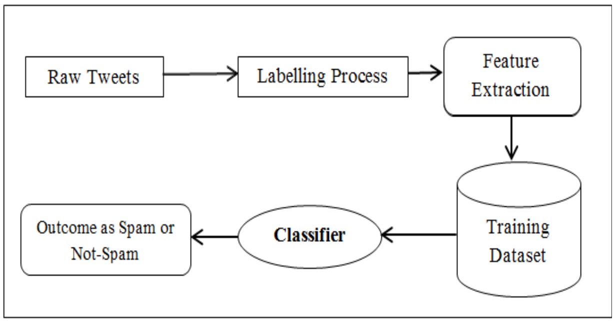

Fig 1. Architecture of the proposed system.

This section consists of the architecture of our proposed system to detect spam tweets. Fig. 1 shows the proposed model.

The system comprises of the following five processes:

A. Tweets Collection

First the system retrieves a set of tweets based on the given topic which is the trending one on twitter. The tweets are captured and stored in a specific file format and then they are all set to be analysed. Trending topics can also be retrieved from twitter separately and based on that other tweets are collected to analyse.

B. Spam labelling

The second process is the labelling process where the system checks through all the available datasets to detect malicious URLs. The label sets obtained will be used to detect spams by training the system accordingly.

C. Feature extraction

In the third step, the labelled tweets are matched against certain language based features, which will help us to calculate the divergence ratio and decide whether the tweets are spam or not. Each tweet will be represented using natural language processing and content analysis technique.

D. Classification

The prepared dataset consisting of shortlisted set of twees where each tweet is represented by a set of features are given to the classifier to train the model and to gain necessary knowledge to detect spams.

E. Spam Detection

The classification algorithm takes a tweet form the user as input and reply back to the user whether it is spam or not. If any miss classification is been noticed by the supervisor, then the supervisor reports to the system about the corrections, which results in updating the dataset from time to time and the classifier goes through a continuous training process.

IV. Methodology

A. *Tweets collection and labelling*

For the evaluation of our approach to detect spams in the tweets, we need a pre-classified collection of tweets as spam or not spam. Such collection is publically available. We then designed a system which can browse through Twitter using Twitter APIs and methods to collect different trending topics and tweets associated to it. We collected the trending topics from the English speaking countries (US, England, Canada, Australia). Trending topics were collected within a five minutes time interval, there were around 15 K trending topics. A total of approx. 70 K tweets based on English language were collected, from which almost 25 K tweets were containing URLs whereas 8 K tweets were labelled as malicious.

The above mentioned tweets were labelled as spams but it cannot be said that the remaining tweets are not spam, perhaps, for proving our model they will be considered as reliable tweets. Hence the remaining tweets need to be analysed again as there may exist another technique through which some of the tweets between them can be detected as spam.

B. *Feature Extraction*

In this section we will look forward for various new language based features to improve spam detection. These features play an important role, as here are not many previous works in statistical analysis of language for spam tweets identification up till now. The result of feature extraction is our final dataset used by classifier. This dataset will contain two sets of tweets, which will be labeled as spam or not spam and it will also contain the tweets which are related to. In the case that a user determines that a tweet has been misclassified, all tweets affected by the same URL may be re-labeled and updated in the dataset.

The language model is based on analysis of txt in the tweet and its associate language. It works over the slices of the text identifying of the probability distribution into it. Generally the original and legitimate language is unknown and is identified based on a sample of text representing that language. The probability distribution can then be identified and known by estimation of different models for each of different texts. The method for extracting such language based features is given as follows.

The first step is to get each trending topic and then one by one the tweets associated to the topic. The tweets found are divided into two categories, one those are not containing any kind of links and the other that contain atleast one URL. The tweets that does not contain any kind of links are separated and are labeled as harmless. Rest of which contain any URL are proven as suspicious. Such tweets are sent for further process. The further evaluation of suspicious tweets is done in two ways, first the tweets are partitioned in a way such that the text is separated and after that the URL associated to the text is separated. The ULR are analyzed whether they are malicious or not, this becomes one of our feature to analyze the tweets. These tweets will be consider for core classification. The features are not just limited to the one mentioned. The text separated from the URL of the suspicious tweets and other features extracted from the tweet. These features are languages based features. Here the natural language processing part comes into picture. The text is processed by removing the stop words and finding the meaning of each separated word. He words are stored in dictionary. There are many other features through the text such as retweet count, number of #hashtags per tweet, number of followers of user and relative posts.

One of our features works on divergence ratio between two separated part. The divergence between the meaning of text and the page found by visiting and exploring the associated URL. Both of their types are checked and the divergence between them is calculated which is counted as our next feature. Spammers now a days are more clever, they know what keywords can be suspicious and what cannot be determined easily. Hence they are also using the most common keywords to carry out their tasks. However on the other side there isn't any connection between the adjacent posts. The purpose of this feature is to analyze the tweets relative to that, to know the divergence point. With the help of the divergence points we are able to build new language model. The cost of computation for extracting these features is very low but they are not effective as expected.

Considering all such scenarios we were able to shortlist all the above mentioned features and we are looking for ten more features which will help us to classify the tweets as spam or not, based on the content and the context of the tweet.

C. *Classification*

If the datasets are pre-defined properly on which the classification process is to be carried out, then the classification can be performed. The classification process can either be performed using the programming method or it can also be done through existing set of tools. We have prepared a setup of Weka (Whitten and Frank 2005) as it contains almost each and every algorithm for machine learning and to carry out data mining tasks. The prepared dataset is evaluated against many of the machine learning algorithms in the Weka tool. The main classifier we followed was the Support vector machine. In previous works this algorithm was used with the default options, it was seen that the algorithm gave the accurate results as compared to other existing classifiers. As the results are not still up-to the mark and the evaluation has to be carried out more precisely modifying the datasets and preparing them according to the requirements

V. Observation And Results

We prepared an experimental setup of machine learning tool "Weka", which performs analysis based on many of the machine learning algorithms. As we have chosen SVM as our main classifier, the results are based on those experiments. For determining the accuracy of the system we worked on a random set of sample 1000 tweets, from which 60% were legitimate and the rest were malicious. As the classes for these users were known already, out of those 1000 tweets 95-97% were classified without mistake.

Confusion Matrix

	Spam	**Legitimate**
Spam	38%	2%
Legitimate	2%	58%

Accuracy

Accuracy	**True Positive**	**False Positive**
93%	0.975	0.05

VI. Conclusion

The main objective of the paper was to analyze the tweets on twitter by applying the mentioned classification algorithms. The SVM classifier gives the standard results. The paper presented a new methodology for detecting spam tweets from twitter which was different from the previous works which only focused on detecting spam accounts. The use of language models were with the purpose to detect the divergence from one place to another, from which some of them further proved to be malicious.

Acknowledgment

I would like to thank my Guide Mr. M.K. Chavan for their continuous help and encouragement. Also a word of appreciation for Department of CSE, Walchand College for providing useful resources.

References

[1] Chao Chen, Jun Zhang, *Member, IEEE*, “A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection”, IEEE Transactions on Computational social systems, vol. 2, no. 3, september 2015.

[2] C. P.-Y. Chin, N. Evans, and K.-K. R. Choo, “Exploring factors influencing the use of enterprise social networks in multinational professional service firms,” *J. Organizat. Comput. Electron. Commerce*, vol. 25, no. 3, pp. 289–315, 2015.

[3] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, “Twitter spammer detection using data stream clustering,” *Inf. Sci.*, vol. 260, pp. 64–73, Mar. 2014.

[4] Juan Martinez-Romo, Lourdes Araujo, “Detecting malicious tweets in trending topics using a statistical analysis of language”, Expert Systems with Applications 40 (2013) 2992–3000.

[5] Allan, J., Wade, C., & Bolivar, A. (2003). Retrieval and novelty detection at the sentence level. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. SIGIR ’03.(pp. 314–321). New York, NY, USA: ACM. http://dx.doi.org/10.1145/860435.860493.

[6] Araujo, L., & Martinez-Romo, J. (2010). Web spam detection: new classification features based on qualified link analysis and language models. IEEE Transactions on Information Forensics and Security, 5(3), 581–590.

[7] Chen, S. F., & Goodman, J., 1996. An empirical study of smoothing techniques for language modeling. In Proceedings of the 34th annual meeting on Association for Computational Linguistics. ACL ’96. Association for Computational Linguistics, Stroudsburg, PA, USA (pp. 310–318). http://dx.doi.org/10.3115/981863.981904.

[8] Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: are you a human, bot, or cyborg? IEEE Transactions on Dependable and Secure Computing, 9(6), 811–824.

[9] Lee, C.-H. (2012). Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams. Expert Systems with Applications, 39(18), 13338–13356. <http://www.sciencedirect.com/science/article/pii/S0957417412007841>.

[10] Nishanth, K. J., Ravi, V., Ankaiah, N., & Bose, I. (2012). Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts. Expert Systems with Applications, 39(12), 10583–10589.

[11] K. Kandasamy, P. Koroth, “An Integrated Approach to Spam Classification on Twitter Using URL Analysis, Natural Language Processing and Machine Learning Techniques”, IEEE Students’ Conference on Electrical, Electronics and Computer Science 2014.

[12] Mochamad Vicky Ghani Aziz, Ary Setijadi Prihatmanto, “Design and Implementation of Natural Language Processing with Syntax and Semantic Analysis for Extract Traffic Conditions from Social Media Data”, IEEE 5th International Conference on System Engineering and Technology, Aug. 10 - 11, 2015.