International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

Spam Detection Techniques for Twitter

Rutuja Katpatal[1], Aparna Junnarkar[2]

[1,2]Department of Computer Engineering
P. E. S. Modern College of Engineering, Pune-05

Abstract - *Twitter is one of the most popular social networking sites among people. Many people use this platform to communicate, share their views, comments regularly. As its popularity is increasing, spammers are also targeting twitter. Twitter spams is becoming a basic issue these days. Twitter have devoted themselves to make a spam-free platform. Various companies are trying to detect spam by applying different schemes. Trend Micro filters spam URL by blacklisting them. But it fails due to time lag. Other machine learning schemes have been introduced which makes use of different machine learning algorithm to classify tweet as spam and non-spam.In this paper, different techniques of twitter spam detection is studied on their detection rate-measure and accuracy.*

Key Words: Twitter Spam Drift, Lfun, Classifier, Detection Rate, False Measure, accuracy

1. INTRODUCTION

Twitter is a social networking site where people interact with each other through messages and post which are called tweets. Only the registered users can post the tweets. Nowadays, use of internet has increased and with its increase use, cyber-attacks have also increased. These attacks not only hampers the security but also destroys the whole internet. People are afraid of using the internet. These attackers send spam messages to users. The social networking site make information available to users and connect them. But these spammers use this freely available information and try to attack user account through which they can get access to their other accounts. It is necessary to save users and system from such spammers. These spammers target the social networking sites. As the twitter is growing, it is more prone to spam attack. Tweets contain URL and links which after clicking directs users to some website which contain viruses, malware, scams etc.[1] .Apart from spamming, phishing, attacks by virus, these social networking sites should keep user data confidential and secure.

Many security companies are trying to find the spam tweets and make twitter safe to use. Trend Micro is another company who is struggling to make twitter spam free. It uses a blacklisting service called Web Reputation Technology system. It filters spam URLs for users who have its products installed [27]. But due to its time difference it is not able to protect user from spam because before it could blacklist particular URL, the user has already visited that URL. Every tweet comprises of different statistical properties like number of followers, number of words per tweet, number of hash tag included in tweet, number of URL in tweet etc.

Different Machine Learning algorithm uses these characteristics of tweet to detect whether it is spam or no spam. They first extract these statistical properties of tweets. These properties help us to differentiate between spam and no spam. Then with spam samples a training data is formed. This training data trains the classifier which in turn detects the spam tweets. However the properties of tweets vary over time. The training data set to train classifier is not updated with changed samples. This issue of changing characteristics over time is called "Twitter Spam Drift" problem. It happens because spammers change the text in tweets keeping semantics same as they are avoiding being detected by security companies. Thus Lfun approach is proposed which tackles twitter spam drift problem. It updates the training data set with changed samples so that new incoming tweets can be correctly classified. It has been observed that only few tweets without URL are classified as spam. Spammers take help of URL which they attach with tweets so that user can be directed to site where malware, viruses can be downloaded. So only spam tweets with URL are considered.

2. RELATED WORK

Twitter is attracting spammer due to its increasing popularity. As more and more people are using twitter daily, it is necessary to protect it from these spammers. Many security companies are trying to find the spam tweets and make twitter safe to use. Trend Micro is another company who is struggling to make twitter spam free. It uses a blacklisting service called Web Reputation Technology system. It filters spam URLs for users who have its products installed [27. But due to its time difference it is not able to protect user from spam because before it could blacklist particular URL, the user has already visited that URL. In order to avoid blacklisting, some researchers used rule to filter spam. Reference[2] filtered spam on three rules: suspicious URL searching, keyword detection and username pattern matching. To eliminate impact of spam, References[3] removed all tweets which has more than three hash tag.

Later machine learning algorithms were applied which extracted statistical features of tweets and formed training data set. A use of account and content based

features[4] like length of tweet, no. of followers, no. of characters in tweets, account age etc were made to detect spam and spammers. It used Support Vector machine. Some researchers trained RF-classifier[5] and then used this classifier to detect spam on social networking sites like Twitter, Facebook and MySpace. \par{} Features discussed in [4] and [5] can be manipulated easily by mixing spam with normal tweets, purchasing more followers etc. Some researchers proposed robust features which was based on social graph so that feature modification can be avoided. A sender and receiver concept was used[6] where the distance and connectivity between tweet sender and receiver was extracted to find out whether it is spam or no spam. Due to this performance of various classifiers were greatly improved. A more robust features such as Local Clustering Coefficient, Betweeness Centrality and Bidirectional Links Ratio were proposed[7] to detect spam tweets.

It has been observed that most of the spam tweets contain URL. Hence it is necessary to study tweets with URL. Various URL based features like domain tokens, path tokens and query parameters of the URL,along with some features from the landing page, DNS information, and domain information have been used to detect spam tweets[8]. In [9] ,the researcher classified tweets as spam using characteristics of Correlated URL Redirect Chains, and further collected relevant features, like URL redirect chain length, Relative number of different initial URLs etc.

Though the above mentioned method can be used to detect spam, it cannot tackle spam drift problem. Various models were built [10] for each user like Language model and Posting Time model. It was found that when these models behaved abnormally ,there is a compromise of the account and then this account is used to spread spam. But it did not identify spamming accounts.

3. SPAM DETECTION METHODS

After analyzing different research paper on spam detection in Twitter, I have considered below three papers for survey. They are-

1) Lfun technique.

2) Asymmetric self learning technique.

3) Binary Detection Model.

A main objective of paper is to analyze these three methods, ensure high detection rate, ensure high accuracy and consistent F-measure.

Lfun(Learning from unlabeled tweets) is used to detect spam tweets. It has two framework:LDT and LHL. It takes labeled tweets to train the classifier. The classifier in turn is used to detect whether the incoming tweets is spam or not.F-measure and detection rate is used to evaluate the performance of method. F-measure is an evaluation metric which combines precision and recall to measure the per-class performance of classification or detection algorithms. Detection Rate is defined as the ratio of those tweets correctly classified as belonging to class spam to the total number of tweets in class spam. Lfun can reach over 90% Detection Rate and can effectively improve the F-measure and the improvement is up to 25% in the best case.

In asymmetric self learning technique, there are three components: Training Stage, Online Detection, and ASL. In the Training Process, a number of tweets are collected first. Then these offline tweets are labeled, and light-weight statistical features are extracted to represent each tweet to form the training set. This training data is used to train a supervised classifier. The trained classifier is then used to detect online tweets in the Online Detection process. Features of incoming tweets are extracted to represent their statistical characteristics. After that, these tweets can be labeled as spam or non-spam by the ML classifier. In ASL scheme, after a pre-defined time window th, the classified spam tweets are added into the training set, and the ML classifier will be re-trained. The re-trained classifier is then used to classify new incoming tweets. The re-training process will be repeated after the defined time th. F-measure and Detection Rate is used to measure the performance. After evaluating this method it was found that the F-measure is improved at least 10% for Bayes Network when used to detect later days' spam, while the Detection Rate is improved more than 20% for all the three tested ML algorithms.

Binary Detection Model does not do feature extraction of tweets. In this method, first Word2Vec is applied to pre-process the tweets instead of feature extraction, where the technique adopted is an advanced language processing method in deep learning and it can convert word or document to representative vector. A binary detection model is built on the basis of several machine learning algorithms to distinguish spam and non-spam. Next parameter setting is assigned for spam filtering. The performance of different classifiers are studied, and then compared existing method to other existing text-based methods. It was found that its detection rate is 92% ,F measure is 94% and it has accuracy of 92%.

4. PERFORMANCE PARAMETERS

For a given system, the detection rate, False measure and accuracy is calculated.

Detection Rate :

Detection Rate is the ratio of tweets which are classified as spam to the total spam tweets.It can be given as
Detection Rate =TP/TP+FN
where
TP is True Positive and FN is False Negative

The higher the detection level, the more messages classified as spam.

False Measure:

F-measure is used as one of the evaluation measure to detect spam. Precision and recall is used to calculate its value. It can be given as
False measure ={2*Precision*Recall}/{Precision+Recall}

Accuracy:

Accuracy is the total number of tweets correctly classified as spam and non spam.
Accuracy={TP+TN}/{TP+TN+FP+FN}

5. RESULTS

The result for above three models are given below

- The detection rate for Lfun,ASL and Binary Detection Model are given below

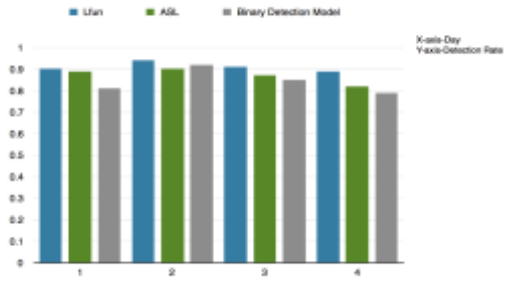

Fig.1.Evaluation of Detection Rate

Table.1.Detection Rate

Parameter	Lfun	ASL	Binary Detection Model
Detection Rate	94%	90%	92%

Fig. shows that Lfun has more detection rate than other two. It has detection rate of more than 94%.

- The F-measure for Lfun,ASL and Binary Detection Model are given below

Fig.2.Evaluation of F-measure

Table.2.F-measure

Parameter	Lfun	ASL	Binary Detection Model
F-measure	83%	85%	94%

Fig. shows that Lfun has consistent F-measure than other two.

- The accuracy for Lfun,ASL and Binary Detection Model are given below

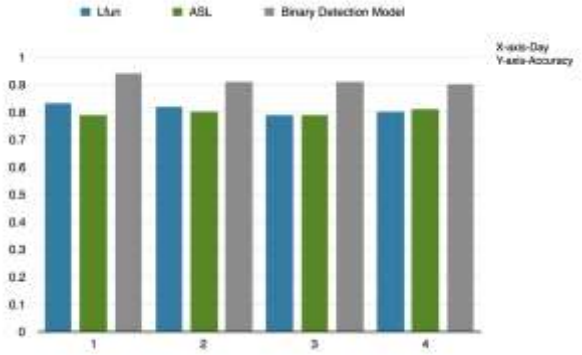

Fig.3.Evaluation of Accuracy

Table.3.Accuracy

Parameter	Lfun	ASL	Binary Detection Model
Accuracy	83%	81%	92%

6. CONCLUSION

Twitter due to its popularity has gained attention of users as well as spammers. These spammers not only try to interfere with privacy of users but also damages the whole internet. Therefore it is necessary to protect the privacy of users. Various spam detection techniques are used to detect spamming activities in twitter. Lfun,ASL,Binary Detection Model are some of the spam detection techniques used. These techniques identifies the spam tweets from incoming tweets. Out of them Lfun technique is better than other techniques on detection rate by 4% and F-measure is consistent. Though its accuracy is less than Binary detection model, fun eliminates the problem of Twitter Spam Drift.

ACKNOWLEDGEMENT

Every orientation work has an imprint of many people and it becomes the duty of the author to express the deep gratitude for the same. I feel pleasure to express deep sense of gratitude and indebtedness to my guide Prof. (Ms) Aparna Junnarkar, for constant encouragement and noble guidance. I also express my sincere thanks to the Computer Department as well as Library of my college. Last but not the least, I am thankful to my friends and my parents whose best wishes are always with me.

REFERENCES

1. J. Oliver, P. Pajares, C. Ke, C. Chen, and Y. Xiang, "An in-depth analysis of abuse on twitter," Trend Micro, Irving, TX, USA, Tech. Rep.,Sep. 2014.

2. S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd, "Detecting spam in a twitter network," First Monday, vol. 15, Jan. 2010.

3. H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in Proc. 19th Int. Conf. World Wide Web, 2010.

4. F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detect- ing spammer on twitter," in Proc. 7th Annu. Collaboration, Electron. Messaging, Anti-Abuse Spam Conf., Jul. 2010.

5. G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in Proc. 26th Annu. Comput. Security Appl. Conf., 2010.

6. J. Song, S. Lee, and J. Kim, "Spam filtering in twitter using sender-receiver relationship," in Proc. 14th Int. Conf. Recent Adv. Intrusion Detection, 2011.

7. C. Yang, R. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving twitter spammers," IEEE Trans. Inf. Forensics Security, Aug. 2013.

8. K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and evaluation of a real-time URL spam filtering service," in Proc. IEEE Symp. Security Privacy, 2011.

9. S. Lee and J. Kim, "Warningbird: A near real-time detection system for suspicious URLs in twitter stream," IEEE Trans. Depend. Sec. Comput.,vol. 10, May 2013.

10. M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Compa: Detecting compromised accounts on social networks," in Proc. Annu. Netw. Distrib. Syst. Security Symp., 2013.

11. C. Chen, J. Zhang, Y. Xiang, and W. Zhou, "Asymmetric self-learning for tackling twitter spam drift," in Proc. 3rd Int. Workshop Security Privacy Big Data (BigSecurity), Apr. 2015.

12. Monika Verma,Divya and Sanjeev Sofat,"Techniques to Detect Spammers in Twitter- A Survey",International Journal of Computer Applications Volume 85 – No 10, January 2014

13. Tingmin Wu, Shigang Liu, Jun Zhang and Yang Xiang,"Twitter Spam Detection based on Deep Learning",ACSW '17, January 31-February 03, 2017, Geelong, Australia

14. Abdullah Talha Kabakus and Resul Kara,"A Survey of Spam Detection Methods on Twitter",International Journal of Advanced Computer Science and Applications, Vol. 8, No. 3, 2017