Proceedings of the International Conference on Inventive Research in Computing Applications (ICIRCA 2018)
IEEE Xplore Compliant Part Number:CFP18N67-ART; ISBN:978-1-5386-2456-2

An Efficient Approach of Spam Detection in Twitter

Rutuja Katpatal
Department of Computer Engineering
P. E. S. Modern College of Engineering, Pune-05
India
mailmeritu89@gmail.com

Aparna Junnarkar
Department of Computer Engineering
P. E. S. Modern College of Engineering, Pune-05
India
Aparna.junnarkar@gmail.com

***Abstract*—Twitter spam has turned into a basic issue these days. Late work concentrate on applying machine learning systems for twitter spam discovery which make utilisation of factual components of tweets.In our marked tweets informational collection, we watch that the measurable properties of spam tweets fluctuate after some time and along these lines,the execution of existing machine learning based classifier diminishes. This issue is called Twitter spam drift. With a specific end goal to handle this issue,a scheme called Lfun scheme is used which can find changed spam tweets from unlabelled tweets and fuse them into classifiers training process.The new training dataset is used to trained another dataset containing unlabelled tweets which will result in finding of spam tweets.Our proposed scheme will adjust training data such as dropping too old samples after certain time which will eliminate unuseful information saving space.**

***Keywords*—Twitter Spam Drift,Lfun,Classifier,Detection Rate, False Measure,Accuracy,Naive Bayes**

I. Introduction

With the advancement in technology,Online Social Network has become very popular among people of all age groups.People share there views, connect with people of same interest and interact with each other through OSN. OSN provides them platform to build professional and social networks. Nowadays various OSN like LinkedIn,Twitter,Facebook have come up with different ways to connect people and people can find relevant information.Due to its popularity it is also attracting spammers.Spammers spread wrong information, rumours,virus links. Sometimes they direct normal users to different website where malware is downloaded. Due to these spammers, normal functioning of users are hampered.They interfere with their work and damages it. It also hampers the reputation of OSN.

Twitter is a social networking site where people interact with each other through tweets. Only the registered users can post the tweets but unregistered users can read it.These tweets are restricted to 140 characters. An unwanted content appearing in twitter can be said as spam. It is necessary to save users and system from such spammers. 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 hashtag 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 nonspam. They first extract these statistical properties of tweets. These properties helps us to differentiate between spam and nonspam. 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.

II. Literature Review

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 [2]. 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[3] filtered spam on three rules: suspicious URL searching, keyword detection and username pattern matching.To eliminate impact of spam, References[4] removed all tweets which has more than three hashtag.

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[5] 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[6] and then used this classifier to detect spam on social networking sites like Twitter, Facebook and MySpace.

Features discussed in [5] and [6] 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[7] where the distance and connectivity between tweet sender and receiver was extracted to find out whether it is spam or nonspam. 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[8] 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[9]. In [10] ,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 [11] 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.In [12] and [14],authors have reviewed various techniques of detecting spammers in Twitter.

Lfun [1] deals with detection spam tweets even if its statistical properties like account age,no. of followers,no. of followings,no.of user favourites,no. of list added,no. of tweets sent etc changes. This changing of properties which is termed as "Twitter Spam Drift" is done by spammers to avoid being detected.It has good detection rate and over period of time its F-measure is consistent.

III. Architecture Of System

It is necessary to detect spam tweets so that user can securely use internet. Lfun(**L**earning **F**rom **Un**labelled) has two components:LDT and LHL. LDT is to learn from detected tweets and LHL is to learn from human labelling. At last test component is introduced to give final classification result.Lfun addresses the problem of spam drift. As the properties of tweets changes, the old model is not updated with changed samples and as a result accuracy decreases. Thus changed samples can be obtained from above two components.After we get enough changed samples, Random Forest is used for classification.

A. *Learning from Unlabelled Tweets(LDT)*

LDT is to learn from detected tweets. In LDT, training sample is updated from given input unlabelled tweets. Initially,

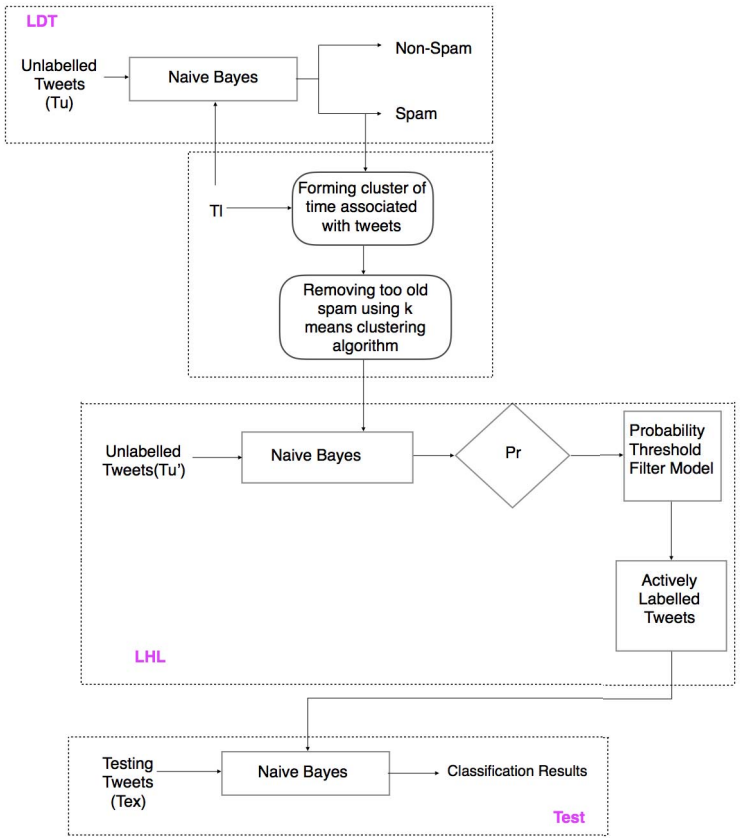

Fig. 1. System Architecture of Lfun

training sample contains set of labelled tweets. A classification model,which classifies tweet as spam or non spam, are updated based on information obtained by LDT.

First two inputs are considered:labelled set of data(T_l) and unlabelled set of data(T_u). Both Tl and Tu contains x_i and y_i as its components. x_i is the vector which represents feature of tweet whereas y_i represents label to tweet.A labelled dataset(T_l) trains the classifier. T_u is given as input to classifier which predicts whether the given tweets is spam or non spam based on training given by T_l. Spam tweets obtained from above classification result will be added to training data sample which in turn gives training to classifier for further classification. Thus a label for each tweet is obtained.

B. *Learning from Human Labelling(LHL)*

In a supervised spam detection system a classifier must be trained by labelled data so that results can be obtained accurately.But these labelled data are expensive. In this case LHL is used. In LHL there are number of unlabelled data and

human annotator is used which label the data. This data is then used to train classifier.The main aim of LHL is to decrease the cost of labelling the unlabelled data by using various models which selects informative samples.

In LHL, a labelled training data set T_t which contains m labelled tweets in which x_i is the vector which represents feature of tweet and y_i represent label of a tweet.Probability Threshold Filter Model serves as selection criteria.To tackle Spam Drift problem,this model select informative tweet from all incoming tweet.The probability threshold value is set between 0.4 and 0.7. After filtering some candidate tweets to be labelled using this model large number of tweets are obtained.Then tweets are selected randomly from the candidate tweets (in this scheme,it is 100) and are manually labelled.

After a training data set is obtained in LDT and LHL, a test component is executed.

C. *Test*

T_e,which is testing sample, along with tweets which are labelled manually trains a classifier.This classifier in turn tackle Spam Drift problem. The old sample technique can also be applied here before training the classifier.Thus this will give more accurate trained data.

It has been observed that as classifier classifies the incoming tweets, size of training data increases as samples are added to it. Hence more time is required to train the classifier which in turn increases response time.So it is necessary to delete old samples from training data set.It can be deleted by:

D. *Old Samples*

After spam is detected it is added to training data set to train the classifier which can further do classification more accurately. These spam tweets can be used to detect too old samples in training data set which can be deleted to save space. Also since training data is reduced it becomes faster to train model.An algorithm known as k means clustering algorithm is implemented which clusters the tweets based on there time stamp.Clusters which are too old are deleted. Thus training data set is reduced which in turn trains the classifier faster.

IV. Experimental Setup

All the tweets which are posted by twitter user is considered as input.These tweets are in the form of numbers representing characteristics of that tweet. These characteristics include account age,no. of followers,no. of followings,no.of user favourites,no. of list added,no. of tweets sent,no. of retweets,no. of hashtags, no. of urls etc.

A. *Hardware and Software Requirements*

The above system is implemented using corei7 processor with speed of 1.3GHz and RAM of 8GB. 80 GB Hard disk is used for this purpose.An operating system Macintosh OS X 10.11.6 with Java 1.8 is used.Netbeans 8.0.2 is considered as the Development Environment. MySQL is used as database for implementation.

B. *Performance Parameters*

For a given system, the detection rate and False measure is calculated and impact of Spam Drift problem is evaluated. Though it evaluates all classes performance,we consider detection rate and F-measure of only spam class.

Detection Rate :

Detection Rate is the ratio of tweets which are classified as spam to the total spam tweets.It can be given as

$$DetectionRate = \frac{TP}{TP + FN} \tag{1}$$

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

$$Falsemeasure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2}$$

It is used in many detection algorithms to measure class performance. Its value as 1 indicates it has best value and 0 indicates worst value. It is used in information retrieval to classify various documents,classification of query, searching of particular document.

Precision:

Precision is defined as the fraction of spam tweets among the retrieved tweets.

Recall:

Recall is defined as the fraction of spam tweets among the relevant tweets.

V. Results

The result of both the system is compared and shown in the form of table and graph
The graphical and tabular result shows the Detection rate and F-measure of Lfun and the improved Lfun approach.

TABLE I. Comparison of Measures

Method	Detection Rate	F-measure
Lfun	91%	88%
Improved Lfun	93%	89%

Fig shows the detection rate of LDT, LHL and Test for both Lfun and improved Lfun.

Fig shows F-measure of LDT, LHL and Test for both Lfun and improved Lfun .

Below fig shows the execution time of Lfun and Improved Lfun

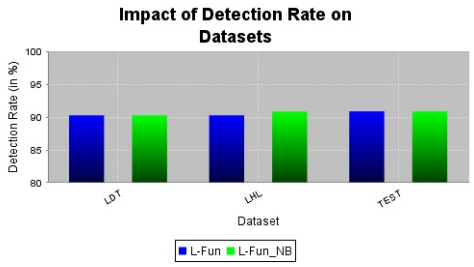

Fig. 2. Detection rate for LDT, LHL and Test

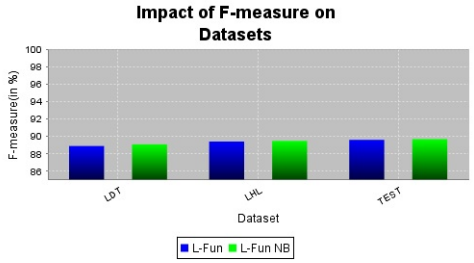

Fig. 3. F-measure for LDT, LHL and Test

Fig. 4. Execution Time for Lfun and Improved Lfun

It can be clearly seen that the Improved Lfun takes less time to execute as we have deleted old spam from its training samples.

VI. 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 deals with detecting spam in twitter even though its statistical properties changes.In Lfun, classifier are trained by training data which classifies tweet as spam and non spam.Detected spam tweets are added to training data set.Over a period of time number of samples in training data set increases. Thus more time is required to train the classifier.It has been observed that too old samples becomes less effective over a period of time. Thus deleting these samples will fasten the process of training as well as space will be saved.

We have applied k means clustering to delete old samples from training data set. Also, we have used Naive Bayes classifier in improved Lfun approach. Results show that Naive Bayes classifier gives better result in terms of Detection Rate and F-measure by 2%. Even the processing time is reduced.In future, we will apply this method on other social networking sites.

References

[1] Chao Chen, Yu Wang, Jun Zhang, Yang Xiang, Wanlei Zhou,Geyong Min, Statistical Features-Based Real-Time Detection of Drifted Twitter Spam, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 12, NO. 4, APRIL 2017.

[2] 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.

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

[4] 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.

[5] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, Detecting spammer on twitter, in Proc. 7th Annu. Collaboration, Electron. Messaging, Anti-Abuse Spam Conf., Jul. 2010.

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

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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

[14] 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

[15] 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