Volume 5, Issue 6, June 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review on Spam Classification of Twitter Data Using Text Mining and Content Filtering

Sandeep Kumar Rawat
M.Tech, Computer Science & Engineering,
Sri Sai University,
Palampur, India

Assistant Prof. Saurabh Sharma
Computer Science & Engineering,
Sri Sai University,
Palampur, India

Abstract— ***In todays world, people are so much inclined towards Social Networking due to which it has become easy to spread spam contents through them. One can have access to the details of any person easily through these social networking websites. No one is secure inside the social media sites. Social Networking sites are fast becoming popular in the recent years, among which Twitter is one of the fastest growing sites. It plays a double role of Online Social Networking and Micro Blogging. Spammers try to attack the twitter trending topics to harm the useful content. Social spamming is more successful as compared to the email spamming by using social relationship between the users. Spam detection is very important because Twitter is mainly used for the commercial advertisements. The spammers attack the private information of the user and also the reputation of the user is harmed. Spammers can be detected by using the content and user based attributes. Traditional classifiers are required for spam detection.***

Keywords— ***Twitter Spam Detection, Classification, Content based Detection. Natural language processing, tweets, machine learning, URLs.***

I. INTRODUCTION

Social Networking is becoming more and more popular these days. It is a sort of platform through which people can share their ideas and thoughts. The most popular Online social networking sites are Twitter, Facebook, MySpace, Google+ etc. With the increasing popularity of these websites, their vulnerability to be attacked is also increasing. These websites have millions of users and not all of these users have legal accounts. Each of these OSNs have many illegal (or spam) accounts with them.Here we have concentrated more on the spammers in Twitter. Twitter is a microblogging site, in which a user is allowed only upto maximum of 140 characters in each tweet. The four major types of spammers on Twitter that we have considered are Phishers, Malware propagators, Marketers and Adult content propagators. *Phishers* spread URLs that intend to harm through their tweets. Innocent users who click on these links end up in wrong websites and the spammers can steal the passwords, credit card information and so on. *Malware propagators* are the users who tweet malicious links, which when clicked on leads to the downloading of malwares which may prove to be very harmful to the genuine users. *Marketers* are the spammers who concentrate on advertising different products. They are normally harmless because, the only thing they do is, publicize their products. But sometimes these users may also mislead the legitimate users. *Adult Content Propagators* are the spammers who broadcast the links containing adult contents. On clicking these links, the user will be redirected to malicious sites. A tweet may or may not contain a URL. Since the Twitter supports only 140 characters in a tweet, a long URL has to be shortened. So many URL shortening services are available now a days. For example, Google URL shortener, Bitly, Twitter URL shortener etc. These shorteners generate short URLs which has extensions like .gl, bit.ly, t.co etc. Moreover, Twitter has different features like @ mentions, # tags and RT. @ mentions are actually used to address a user. # Tags are used for trending a topic. RT shows that the tweet is retweeted.

II. LITERATURE SURVEY

Previous work has classified spammers with high accuracy, but two critical limitations exist. First, they used the account features such as tweeting interval, content similarity, age, the number of followings and the number of followers. These account features, however, can be manipulated by spammers. For instance, spammers can post both benign and spam tweets at irregular intervals. They can also create several spam accounts and follow each other to raise their reputation in social networks. Moreover, spammers can use accounts created a long time ago to manipulate the age feature. Secondly, previous work is able to detect spammers only after spam has already been sent to legitimate users because user history data is needed to decide whether a user is a spammer or not. To classify a user, previous methods need to know how a user has been tweeting and what a user has been tweeting. Therefore, there is an inevitable delay between spam account creation and its detection. Because of the delay, previous work has been criticized [8]. Even if spammers are detected and removed, they can still create accounts and then send spam again.

Spam appears in email, blogs, Short Message Services (SMS), and Social Networking Sites (SNS). Many researchers have proposed schemes to detect spam. The common feature of spam, as defined by the researchers, is that it is unsolicited one [7]. However, it is difficult to decide whether a message is unsolicited in receivers' side. Thus, content

filtering methods are widely used [13]. In social networking services such as Twitter, however, content filtering approaches are not effective because spam contains only a few words and URLs. Domain and URL blacklisting techniques have also been proposed to filter spam, but Grier et al. showed that the blacklists are too slow to protect users since there is a delay before hostile sites are included in blacklists [16]. Moreover URL shortening services make it more difficult to detect sites in blacklists. Thus, the approach is not effective in Twitter because almost all users use URL shortening services due to limitation of message length. Because of these reasons, traditional spam detection approaches are difficult to apply to Twitter. Therefore, a new approach is needed with a focus on the characteristics of Twitter.

In this section, we make a research about how to detect spam user and how to make topic classification. Currently many papers proposes several methods like using graph-based model (Naïve Bayesian) [12] and some content features to detect spams[11]. Since all these features play important roles when trying to detect a spammer, it's a good idea to try to merge them together to create a better model to further improve the detection accuracy. For topic classification, we mainly make a research about the models which are generally used in topic classification area. After that, we make a comparison about each model: Unigram model, this is a naïve model which ignores the topic of each model and only takes account of the words and documents. Mixture of unigram model, this model conquers the weakness of Unigram model, and takes topic of each document as a parameter. However, this model only gives each document one certain topic, while in the real life, a document can be labeled two or more topics. For example, a document about the death of Michael Jackson can be labeled both as news and entertainment.

C. Grier et al. [1] show that it takes a few weeks for URLs posted in Twitter to be on its blacklist. In addition to the fact that Twitter itself does to prevent spamming, Twitter relies on users to report spam. Once a report is filed, Twitter investigates it to decide to suspend an account or not. Currently, much research is going on to find a method to detect Twitter spamming in an efficient and automated way. After all, it is not very reliable for the Twitter community to depend on users to identify spams manually based on previous spam activities. Thomas et al. [16] analyzed over one million suspended Twitter accounts to characterize the behavior and lifetime of spam accounts, to understand how spammers abuse legitimate web services such as URL shortening services by exploring spam affiliate programs and market place of illegitimate programs run by spammers. They report that 77% of spam accounts are pended within the first day of their first tweet and 92% of accounts within three days. Less than 9% of accounts form follower/following relationships with regular users, 52% of spam accounts use unsolicited mentions, and 17% used hashtags in the messages with unrelated contents for trend topic search. They also report that 89% of spam accounts have fewer than 10 followers. Link is an important feature to detect spams or to infer user opinions in sentiment analysis [12].

The computed rank depends on the user's connectivity in the social graph. The more followers a user has, the more likely his/her tweets are to be ranked high. Spammers attempt to use this ranking score by acquiring links in Twitter-they follow other users and try to get others to follow them as a courtesy of "social etiquette". Links in Twitter and reported that popular users who have many followers are not necessarily influential in terms of spawning retweets or mentions.

Their findings also show that ranking users by the number of followers atches with results computed by PageRank while ranking by retweets differs from PageRank and from the number of followers and any retweeted tweet reached an average 1000 users regardless of the number of followers of the original tweet account. Previous study at U.C. Berkeley shows that 45% of users on a social network site readily click on URLs without doubt. Grier et al. [5] collected over 400 public tweets and reported that 8% of 25 million unique URLs posted to Twitter point to phishing, malware, and scam. They reported that the click through rate is 0.13% which is almost twice higher than the email spam click through rate previously published and 80% of clicks occur within the first day of a spam URL appearing on Twitter.

In previous work, Benevenuto edeveloped a system to detect spammers in Twitter by using the properties of the text of the users' tweets and user behavioural attributes, being able to detect about a 70% of the spammer accounts.

Grier proposed a schema based on URL blacklisting.

Wang proposed a new approach that employed features from the social graph from Twitter users to detect and filter spam. Although these approaches are able to deal with the problem, a new spammer account may emerge to substitute the filtered accounts. Hence, these blacklisting systems, similarly as happens with e-mails, should be complemented with content-based approaches commonly used in e-mail spam filtering. In this context, a recent work by Martinez-Romo and Araujo proposed a method for detecting spam tweets in real time using different language models and measuring the divergences. Also, currently Twitter is not only a communication application, but also a fast way to help users get the latest information they are interested in. More and more people are not satisfied only by using Twitter as a communication application, they also expect Twitter to recommend some useful information which they are interested in. In order to build an automatically recommendation system, one of the most critical task is how to extract interest for each user. To realize this critical task, generally we need to extract tweets posted by each user and classify them into different classes, after that, we can find exactly interest for each user. In conclusion, the necessity of building a model to automatically classify tweets into different topics is self-evident

III. PROBLEM FORMULATION

Twitter spam is different from traditional spam such as email and blog spam, conventional spam filtering methods are inappropriate to detect it. Thus, many researchers have proposed schemes to detect spammers in Twitter. These schemes are based on the features of spam accounts such as content similarity, age and the ratio of URLs. However, there are two significant problems in using account features to detect spam. First, account features can easily be fabricated by spammers. Second, account features cannot be collected until a number of malicious activities have been done by spammers. This means that spammers will be detected only after they send a number of spam messages. In this paper, we

propose a novel spam filtering system that detects spam messages in Twitter. Instead of using account features, we will use relation features, such as the distance and connectivity between a message sender and a message receiver, to decide whether the current message is spam or not.

IV. OBJECTIVE

we analyze an impact of preprocessing of Twitter data for a spam detection task. More specifically the purpose is

- To evaluate the impact of using different attributes of Twitter data on different classifiers.
- To evaluate an impact of using a small number of attributes and a large number of attributes.
- To evaluate if preprocessing steps, such as discretization, normalization and transformation with Twitter data may increase classification rates.
- To evaluate if all these steps are consistently needed for all classifiers used.

V. METHODOLOGY

We propose a spam filtering method in Twitter. Instead of account features, our study considers the relation features between a message sender and a receiver, which are difficult for spammers to manipulate. This means that when a user receives a message from a stranger, our system identifies the sender at once. If the sender is identified as a spammer, the message is filtered. We propose a spam filtering system for Twitter. We classify the messages as spam or benign messages by identifying the sender. Our system identifies spammers in real-time, meaning that service managers or clients can classify the messages as benign or spam when a message is being delivered.

Pre-processing

Tokenization

A token is a block of text which is considered as a useful part of the unstructured text. In most of the cases it is mapped to a word, but a token could be also represented by a paragraph, a sentence, a syllable or a phoneme. Tokenization represents the process of converting a stream of characters into a sequence of tokens. Tokenization is done by removing punctuation, brackets, capitals, etc.

Stemming

Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form. For example, "run", "runs", "ran", and "running", are all forms of the same root, conventionally written as "run" and the role of a stemmer is to attribute all the derived forms to the root of the lexeme. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

Content-based Features

A) Duplicate Tweets

A user may be considered as a suspicious spammer if he post too many duplicate tweets. Duplicate tweets are detected by measuring the Damerau Levenshtein distance (also known as edit distance) between two different tweets posted by the same account.

B) HTTP links

Malicious links is another way to spread advertisements and spams. We will check the tweets data whether contains too many links. Moreover, we will try to go deep into the link to find exactly whether the forwarded webpage is a spam or not.

Replies and Mentions Twitter provides user a function to reply message to anther user in @username + message format where @username is the message receiver. This reply and mention function is designed to help users to track conversation and discover each other on Twitter. However, this service is abused by the spammers to gain other user's attention by sending unsolicited replies and mentions. We will use the number of "@"symbol as a feature in a proper way to detect spammers.

C) Trending Topics

Trending topics are the most-mentioned terms on Twitter at that moment, in this week, or in this month. Users can use the hash tag, which is the # symbol followed by a term to describe a topic in a tweet. If there are too many tweets containing the same term, this will help the term to become a trending topic. Unfortunately, because of how prominent trending topics are, spammers post multiple unrelated tweets that contain the trending topics to lure legitimate users to read their tweets. We will count the number of trending topics in the tweets as a feature for spammer detection.

VI. CONCLUSION

Spammers are the problem in any online social networking sites. Once a spammer is detected it is easy to suspend his/her account or block their IP address. But they try to spread the spam from other account or IP address. Hence it is recommended to check for spam content in a tweet in the server. If any content matches the spam words present in the data set it is prevented from being displayed. Accuracy is being evaluated in classifying the spam content. Many traditional classifiers are present in classifying spammers from legitimate users but many classifiers wrongly classify

non-spammers as spammers.It is efficient to check for spam content in tweets. So we have to propose an integrated approach for the classification of a Twitter user into spam or legitimate.The combined approach, which includes URL analysis,Natural Language processing and Machine Learning techniques, which could successfully do the classification. The combined approach can give more accuracy than each of these methods being applied alone. Also, here we can identify different set of expressions, tweets, words and other features which can show that a user is a spam or legitimate.

REFERENCES

[1] Thomas, K., Grier, C., Song, D., Paxson, V.: *Suspended accounts in retrospect: an analysis of twitter spam*. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, ACM (2011) 243–258

[2] Bratko, A., Filipiˇc, B., Cormack, G., Lynam, T., Zupan, B.: *Spam filtering using statistical data compression models*. The Journal of Machine Learning Research 7 (2006) 2673–2698

[3] Jagatic, T., Johnson, N., Jakobsson, M., Menczer, F.: *Social phishing*. Communications of the ACM 50(10) (2007) 94–100

[4] Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: *Detecting spammers on twitter.* In: Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS). (2010)

[5] Grier, C., Thomas, K., Paxson, V., Zhang, M.: @spam: *The underground on 140 characters or less.* : Proceedings of the 17th ACM conference on Computer and communications security, ACM (2010) 27–37

[6] Wang, A.H.: Don't follow me: *Spam detection in twitter*. In: Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on, IEEE (2010) 1–10

[7] Gao, H., Chen, Y., Lee, K., Palsetia, D., Choudhary, A.: *Towards online spam filtering in social networks.* In: Symposium on Network and Distributed System Security (NDSS). (2012)

[8] Ahmed, F., Abulaish, M.: *A generic statistical approach for spam detection in online social networks.* Computer Communications (2013) in press.

[9] Martinez-Romo, J., Araujo, L.: *Detecting malicious tweets in trending topics using a statistical analysis of language.* Expert Systems with Applications (2012)

[10] Sebastiani, F.: *Machine learning in automated text categorization*. ACM computing surveys (CSUR) 34(1) (2002) 1–47

[11] Schneider, K.: A comparison of event models for Naive Bayes anti-spam e-mail filtering. In: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics. (2003) 307–314

[12] Androutsopoulos, I., Koutsias, J., Chandrinos, K., Spyropoulos, C.*: An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages*. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. (2000) 160–167

[13] Seewald, A.: *An evaluation of naive Bayes variants in content-based learning for spam filtering.* Intelligent Data Analysis 11(5) (2007) 497–524

[14] Wang, A.H.: Don't follow me: *Spam detection in twitter*. In: Proceedings of 5th International Conference on Security and Cryptography (SECRYPT). (2010) 142– 151

[15] Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: *Who is tweeting on twitter: Human, bot, or cyborg?* In: Proceedings of Annual Computer Security Applications Conference (ACSAC). (2010)

[16] Thomas, K., Grier, C., Ma, J., Paxson, V., Song, D.: *Design and evaluation of a real-time url spam filtering service*. In: Proceedings of the IEEE Symposium on Security and Privacy. (2011)

