Spam Detection on Twitter: A Survey

Prabhjot Kaur
Maharaja Surajmal Institute of Technology, C-4, Janakpuri, New Delhi, INDIA
Email Id: thisisprabhjot@gmail.com

Anubha Singhal
Maharaja Surajmal Institute of Technology, C-4, Janakpuri, New Delhi, INDIA
Email id: anubha.ssinghal@gmail.com

Jasleen Kaur
Maharaja Surajmal Institute of Technology, C-4, Janakpuri, New Delhi, INDIA
Email Id: jasleenkk94@gmail.com

Abstract– **With the rapid growth of social networking sites for chatting with friends, meeting new people and keeping informed of what's happening in the world, there are those who saturate user's account with messages that are not of his interest, called spam. Spam is used to misled and deceive user by posting harmful links, posting repeatedly to trending topics to grab attention or by posting advertisements. A lot of research has been done to detect spam on twitter. In this paper, we have reviewed research papers published from 2010-2015. Current study provides techniques used, its type, dataset and accuracy.**

Keywords – Content based; Hybrid; Relation based; Spam; Twitter; User based

I. INTRODUCTION

Twitter is a micro-blogging service that enables users to connect with friends and other fascinating people by posting short messages, called tweets. Each tweet is restricted to 140 short characters and text. HTTP links, photos and videos can also be included in a tweet. Twitter acts as a platform for discussion, breaking news, entertainment. Twitter provides users with certain features such as retweet, hashtag, mention, reply, following and followers. Retweet is used when a user wants to share a tweet of another person to his followers. Hashtag symbol # is used in front of a keyword to categorise messages and all the messages related to the topic can be read clicking on that keyword. Mention allows user to mention another user in a tweet by using @username .A reply is answer to another user's tweet and it begins with @username with the name of a person you are replying to. This can be used by a spammer to gain user's attention. Following someone means subscribing to their tweets and getting all the updates of following.

Till May 2015, Twitter has more than 500 million users, among which more than 302 million are active users [15]. So, the more appealing the mechanism to spread news, to discuss new events or to post a status, the more it opens opportunities for the new types of spam. Spam is defined by Twitter as unsought, recurred actions that have a negative impact on other users. It includes many forms of automated account interactions and behaviours to mislead or deceive users. Behaviours that constitute "spamming" on Twitter will continue to evolve. [16] For instance, trending topics, which are the most talked about event on Twitter at any point of time, are seen as an opportunity to generate traffic and revenue. When any notable event occurs, thousands of users tweet about it and rapidly make it trending topic. These trending topics become the target of spammers who post tweets consisting of some characteristic words of the trending topic with URL links that lead users to completely unrelated websites. As tweets usually include shortened URL links, it becomes difficult for the users to identify the content of the URL without loading the website. This type of spam can affect real time search services. So the mechanism to fight and stop spammers need to be established.[1]

Spammers can have several motives behind spamming, such as to propagate advertisements to generate high return on sales, to disperse pornography, to diffuse viruses or phishing. Spammers not only pollute real time search environment, but they also interfere with the statistics provided by tweet mining tools[2].So , spam detection is the most critical step in fighting spam. Spam can be detected by user or content or relation based techniques.

Twitter allows user to report spam by visiting the spam account's profile then click or tap the gear icon. Select "Report" from the menu and select "They're posting spam" to submit a report. It also allows users to report individual tweets. But, this manual method requires users' effort and there can be many fake reports.

Many research papers have been published to detect spam on twitter. In this research paper, we have done a survey of research papers from 2010 – 2015. Our paper aims to define the various types of techniques used to detect spam in twitter and the attributes used by the techniques to detect spam. This paper also aims to give a review of how these techniques have been implemented by various researchers. This paper is structured as follows: (II) Types of detection techniques (III) Related work (IV) Conclusion

II. TYPES OF DETECTION TECHNIQUES

Based on our survey of research papers from 2010-2015, we classified the various spam detection techniques into following four categories –

A. User Based Techniques

Every tweet besides its text includes several other attributes that give additional useful information of the user and his behaviour. Based on these attributes we can classify a user into

978-9-3805-4421-2/16/$31.00 ©2016 IEEE

spammer and non-spammer. Few of Twitter's inherent attributes can be listed as –

- Followers count
- Followings count
- Mentions count
- Reply count
- date of creation
- time of tweet

These inherent attributes together with derived attributes can efficiently help in detecting spammers. Few such derived attributes are-

- Reputation (R) - R is the relative amount of friends' count to followers' count [3]. Spammers have been found to have lower R than non-spammers.
- Age of the user – It can be calculated from date of creation of user account. Spammers have relatively lesser age [6].
- Following Rate (FR) – The rate at which the user follows other users. [8]
- Tweet Frequency (TF) – how frequent a user tweets? TF is higher in spammers.

Various works has been done using these user based attributes, some of which are enlisted in the table in following section.

B. Content Based Techniques

The user based techniques have a disadvantage that they can easily be manipulated by the user. For instance, a user can create multiple accounts and may follow each other back to increase his reputation [5]. Also the suspended user can create a new account.

An alternative to this is to use content based features that use probabilistic models and analyze the linguistic features of the text of tweet and use the result to classify spam and non-spam tweets rather than spam and non-spam users.

The content based techniques (mentioned in [1]) use the linguistic properties of tweet such as –

- Count of hash-tags in a tweet in relation to word count,
- Count of URLs in a tweet
- Count of words in tweet
- Count of characters in tweet
- Count of numerals that appear in the text part of the tweet
- Number of users enlisted in each tweet
- Count of times the tweet is retweeted (measured by existence of "RT @username" in the text).

This methodology allows locating spam without any know about of the user who tweets the tweet and analysis of the speech in the tweet. It facilitates analysis of content of the tweet rather than user account.

C. Hybrid Based Techniques

As mentioned earlier, use of user based techniques is not effective as the user based attributes can be simply manipulated by the spammer. Similarly, using content based features has its own limitations as the spam contains only few words and URLs. URL shortening services extensively used in order to cope with 140 character limit of Twitter makes it even tougher. Thus, several techniques came up that uses both user based as well as content based features to identify the spam.

D. Relation Based Techniques

Relation based technique is used because in user based technique, spammer is detected after the spam has been sent to the valid user so there is a unavoidable delay between spam account creation and spam account detection [12].So, relation features are used as they are difficult to manipulate. It is used to find spam in real-time since history data of user is not required. For instance, if user receives a tweet from an unknown person then it recognizes the sender immediately. Two types of relation features are:-

- Distance – It is the length of the shortest path between users. For example, the distance between the users is one if they are connected by a single edge. It means that the two users must be friends. When some users have distance two or three, they are not friends themselves but they have mutual friends. But if it is more than three then it can be a spam[12]
- Connectivity – The connectivity resembles the relationship strength. The connectivity between a valid user and a spammer is usually weaker than the connectivity between non-spammers, even if the distance is same [12].

III. RELATED WORK

TABLE I. OUTLINE OF RESEARCH IN DETECTING SPAM

No.	TITLE	METHODOLOGY	TYPE	DATASET	ACCURACY/ RESULT
1	"The Social Honey pot Project: Protecting Online Communities from Spammer" [13]	Deployment of social honey pots for harvesting deceptive spam profiles	User Based	Validated on 500 user set	88.98%
2	"Detecting Spammers on Twitter" [1]	Support Vector Machine (SVM) classifier	User Based and Content Based	Validated on 8207 user set	70% for spammers ; 96% for non spammers
3	"Spam Detection on Twitter Using Traditional Classifiers" [2]	Compared Random Forest, SVM, Naïve Bayesian, KNN	User Based and Content Based	Validated on 1000 user set	95.7% - random forest(highest)

4	"Don't follow me -spam detection in twitter" [3]	Compared Naïve Bayesian , Neural network, SVM and Decision tree	User Based and Content Based	Validated on 25K users, 500K tweets	89%
5	"Detecting Spammers on Social Networks" [4]	Creation of honey profiles	User Based	Validated on 135,834 user set	90.93%
6	"Machine Learning Techniques applied to Twitter Spammers Detection" [6]	SVM, Extreme Learning Machine (ELM), Random Forests (RFs)	User Based and Content Based	Validated on 1065 users set	For 20,10,5 parameters for spammers detected as spam RF- 76.9%,75.2%, 77.7% SVM- 68.6%,69.4%,67.7% ELM- 76.8%,70.2%,71.9%
7	"Making the Most of Tweet-Inherent Features for Social Spam Detection on Twitter" [8]	Naïve Bayes, k-Nearest Neighbors, SVM, Decision tree and Random Forest	User Based, Content Based , N-grams and Sentiments Based	Validated on 2 data sets First with 20,707 spam & 19,249 other tweets and second with 1,000 spam & 9,828 other tweets	Random Forest proves to be best classifier with precision of 83.1% for 1KS-10KN Dataset and 94.6% for Social honey pot Dataset
8	"Detecting malicious tweets in trending topics using a statistical analysis of language" [14]	Decision tree, Naïve Bayes, Logistic Regression, SVM, Decorate, Random Forest	Content Based	Validated on set of 34 K trending top topics & 20 million total tweets	Able to locate 94.5% of actual spam and resulted in false positive rate of 5.4%.
9	"Spam Filtering in Twitter using Sender-Receiver Relationship" [12]	Sender-receiver relationship measuring distance and connectivity	Relation based	Validated on 148,371 profiles and 267,551 tweets	Bagging- 93.3% (TP) 8.5%(FP) LibSVM- 93.2%(TP) 8.3%(FP) FT- 93.1%(TP) 7.7%(FP) J48- 92.3%(TP) 8.7%(FP) BayesNet- 92%(TP)8%(FP)

IV. CONCLUSION

From the research papers reviewed it can be concluded that the spam detecting techniques can be broadly divided into four categories based on what attributes they use and whether they tend to identify the spam user or spam tweet. It has also been observed that a hybrid of these techniques yields better resultsas it can identify both spam tweet as well as spam user.

REFERENCES

Journal References

[1]. Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida, "*Detecting Spammers on Twitter*", CEAS 2010 Seventh annual Collaboration,Electronic messaging, Anti Abuse and Spam Conference, July 2010, Washington, US.

[2]. M. McCord, M. Chuah, "*Spam Detection on Twitter Using Traditional Classifiers*", ATC'11, Banff, Canada, Sept 2-4, 2011, IEEEn on Twitter Using Traditional Classifiers, ATC'11, Banff, Canada, Sept 2-4, 2011, IEEE

[3]. A.H. Wang, "*Don't follow me: spam detection in Twitter*", Security and Cryptography (SECRYPT), in: Proceedings of the 2010 International Conference on IEEE, 2010, pp. 1–10.

[4]. Xianghan Zheng , Zhipeng Zeng , Zheyi Chen , Yuanlong Yu , Chunming , *"Detecting Spammers on Social Networks*", www.elsevier.com/locate/neucomputing

[5]. Sandeep Kumar Rawat , Assistant Prof. Saurabh Sharma ,"*A Review on Spam Classification of Twitter Data Using Text Mining and Content Filtering*", International Journal of Advanced Research in Computer Science and Software Engineering 5(6), June- 2015, pp. 485-488

[6]. Claudia Meda et al."*Machine Learning Techniques applied to Twitter Spammers Detection*", www.wseas.us/elibrary/conferences/2014/Florence/CSCCA/CSCC A-23

[7]. Ms. Monali kakde, Amol Muley "*Designing a Framework To Detect Twitter Spammers Using Forensic Approach*" ISSN (Online): 2347 - 2812, Volume-2, Issue -11,12 2014,pp- 81-85

[8]. Bo Wang et al. "*Making the Most of Tweet-Inherent Features for Social Spam Detection on Twitter*", published at a part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395, May 18,2015

[9]. A Jenefa, Dr. R. Ravi, "*Classifier: A Real-Time Detection system for suspicious URLs in Twitter Stream*", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 2, Issue 2, February 2014 , pp-53-57

[10]. Amrita Mathur ,Prachi Gharpure, "*Spam Detection Techniques: Issues and Challenges*", Foundation of Computer Science FCS, New York, USA International Conference & workshop on Advanced Computing 2013

[11]. Girisha Khurana, Mr Marish Kumar,"*Review: Efficient Spam Detection on Social Network*", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 3 Issue VI, June 2015,pp 76-80

[12]. Song, J., Lee, S., & Kim, J. (2011). "*Spam filtering in twitter using sender–receiver relationship*". In R. Sommer, D. Balzarotti, & G. Maier (Eds.), Recent advances in intrusion detection. Lecture notes in computer science (Vol. 6961, pp. 301–317).Berlin/Heidelberg: Springer.

[13]. Kyumin Lee, James Caverlee, Steve Webb, "*The Social Honeypot Project: Protecting Online Communities from Spammers∗*", WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA

[14]. Juan Martinez-Romo , Lourdes Araujo, "*Detecting malicious tweets in trending topics using a statistical analysis of language*" , NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Madrid 28040, Spain

Website References

[15]. https://en.wikipedia.org/wiki/Twitter

[16]. Twitter Support https://support.twitter.com/articles/64986