Spam Detection in Twitter - A Review

C. Divya Gowri and Professor V. Mohanraj
Sona College of Technology, Salem

ABSTRACT
Social Networking sites have become popular in recent years, among these sites Twitter is one of the fastest growing site. It plays a dual role of Online Social Networking (OSN) and Micro Blogging. Spammers invade twitter trending topics (popular topics discussed by Twitter user) to pollute the useful content. Social spamming is more successful compared to email spamming by using social relationship between the users. Spam detection is important because Twitter is mainly used for commercial advertisement and spammers invade the privacy information of the user and also the reputation of the user is damaged. Spammers can be detected using content and user based attributes. Traditional classifiers are required for spam detection. This paper focuses on study of detecting spam in twitter.

Keywords: Social Network Security, Spam Detection, Classification, Content Based Detection.

1. INTRODUCTION
Web-based social networking services connect people to share interests and activities across political, economic, and geographic borders. Online Social Networking sites like Twitter, Facebook, and MySpace have become popular in recent years. It allows users to meet new people, stay in touch with friends, and discuss about everything including jokes, politics, news, etc., Using Social networking sites marketers can directly reach customers this is not only benefit for the marketers but it also benefits the users as they get more information about the organization and the product. Twitter [1] is one among these social networking sites. Twitter provides a micro blogging (Exchange small elements of content such as short sentences, individual images, or video links) service to users where users can post their messages called tweets. Tweet can be limited to 140 characters only HTTP links and text are allowed. Twitter user is identified by their user name optionally by their real name. The user ‘A’ starts following other users and their tweets will appear on A’s page. User A can be followed back if other user desires. Trending topics in Twitter can be identified with hash tags (‘#’). When a user likes a tweet he/she can ‘retweet’ that message. Tweets are visible publically by default, but senders can deliver message only to their
followers. The ‘@’ sign followed by username is a reply to other user. The most common type of spamming in Twitter is through Tweets. Sometimes it is via posting suspicious links.

Spam [14] can arrive in the form of direct tweets to your Twitter inbox. Unfortunately spammers use twitter as a tool to post malicious link, send spam messages to legitimate users. Also they spread viruses, or simply compromise the system reputation. Twitter is mainly used for commercial advertisement, and spammers invade the privacy information of the user and also the reputation of the user is damaged. The attackers advertise on the Twitter for selling products by offering huge discount and free products. When users try to purchase these products, they are asked to provide account information which is retrieved by attackers and they misuse the information. Therefore spam detection in any social networking sites are important.

2. RELATED WORKS
McCord et al., [1] has proposed user based and content based features to facilitate spam detection.

User Based Features
The user based features considered are number of friends, number of followers, user behaviors e.g. the time periods and the frequencies when a user tweets and the reputation (Based on the followers and friends) of the user. Reputation of a user is given by the equation,

\[ R(j) = n_t(j)/(n_t(j) + n_0(j)) \] (2.1)

Where \( n_t(j) \) represents the number of followers of user ‘j’ and \( n_0(j) \) represents the number of friends the user ‘j’ has. According to Twitter Spam and Abuse Policy ‘if the users have small number of followers compared to the amount of people the user following then it may be considered as a spam account’. Spammers tend to be most active during the early morning hours while regular users will tweet much less.

Content Based Features
The content based features [11] considered in this approach are number of Uniform Resource Locator’s (URL), replies/mentions, keywords/word weight, retweet, hash tags. Retweet is a reposting someone’s post, it is like a normal post with author’s name. It helps to share the entire tweet with all the followers. The ‘#’ containing tweets are the popular topics being discussed by the users.
Secondly, they compare four traditional classifiers namely Random forest, Support Vector Machine (SVM), Naïve Bayesian and K-nearest neighbor classifiers which are used to detect Spammers. Among these classifiers Random Forest is found to be effective but this classifier can deal with only imbalanced data set (data set with more regular users than spammers). Alex HaiWang [2] considered ‘Follower – Friend’ relationship in his paper. A ‘Direct Social Graph’ is modeled. The author considers content based and graph based features to facilitate spam detection.

**Graph Based Features**

A social graph is modeled as a direct graph G= (V, A) where set of V nodes representing user accounts and set A that connect the nodes. An arc a = (i, j) represents user i is following user j. Follower is considered as the Incoming links or in links of a node i.e., People following you not necessary that you should follow back. A Friend is an Outgoing links or out links. i.e., People you are following. A Mutual Friend is a Follower and Friend at the same time. When there is no connection between two users then they are considered to be strangers.

![Fig 2.1 A Simple Twitter Graph](image)

In the above figure user A is following user B, user B and user C are following each other. i.e., User B and user C are mutual friends, and User A and User C are strangers. The graph based features considered are number of followers, number of friends, and the reputation of a user.

The classifier used in this paper to detect spam is Naïve Bayesian classifier [10]. It is based on Bayes theorem which is given by the equation,

\[
P(Y|X) = \frac{(P(X|Y)P(Y))}{P(X)}
\]

The twitter account is considered as vector X and each account is assigned with two classes Y spam and non-spam, the assumption is that the features are considered to be conditionally independent. This classifier is easy to implement, it requires small amount of training data set. But, conditionally independence may lead to loss of accuracy. This classifier cannot model independency.
Twitter Account Features:
Zi Chu et al., [13] review some of the classification features to detect spammers. These include tweet level features and account level features. The tweet level features include Spam Content Proposition i.e. tweet text checked with spam word list and the final landing URL is checked. The account level features include account Profile which is the self description of short description text and homepage URL and check whether the description contains spam words.

Fabricio Benevento et al., [3] have considered the problem of detecting spammers. In this paper approximately 96% legitimate users and 70% spammers were correctly classified. Like [1] user based and content attributes are considered. To detect spammers with accuracy, confusion matrix is introduced.

<table>
<thead>
<tr>
<th></th>
<th>Spam</th>
<th>Not Spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Not Spam</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Fig 2.2 An Example of Confusion Matrix

In the above matrix, ‘a’ is the number of spam correctly classified, ‘b’ is the number of spam wrongly classified as non-spam, ‘c’ is the number of non-spam wrongly classified as spam and ‘d’ is the number of non-spam correctly classified. For effective classification some of evaluation metrics are considered. They are Precision, Recall, F-measure (Micro-F1, Macro-F1).

**Evaluation Metrics:**

**Precision:** It is defined as the ratio of the number of users classified correctly to the total predicted users and is given by the equation,

\[
Precision, p = \frac{a}{a + c}
\]  (2.3)

**Recall:** It is defined the ratio of number of users correctly classified to the number of users and is given by the equation,

\[
Recall, r = \frac{a}{a + b}
\]  (2.4)

**F-measure:** It is the harmonic mean between precision and recall and it is given by the equation,

\[
F\_measure = \frac{2pr}{p + r}
\]  (2.5)
The classifier used to detect spam is SVM. It is a state of the art method in classification and in this approach they use non-linear SVM with the Radial Basic Function kernel that allow SVM to perform with complex boundaries. The biggest limitation of the support vector approach lies in choice of the kernel and high algorithmic complexity. This approach mainly focuses on detecting spam instead of spammers so that it can be useful in filtering spam. Once a spammer is detected it is easy to suspend that account and block the IP address but spammers continue their work from other new account.

Puneeta Sharma and Sampat Biswas [4] proposed two key components (1) identifying timestamp gap between two successive tweets and (2) identifying tweet content similarity. They found two common techniques used by spammers (1) Posting duplicate content by modifying small content of the tweet; (2) Post spam within short intervals. Spam Identification approach included BOT Activity Detection and Tweet Similarity Index. Twitter data can be filtered in various ways by user id, by keyword, many spammers post spam messages using BOT (computer program), reducing the frequency between consecutive tweets. To calculate timestamp between tweets, they first cluster tweets based on user id and sort by increasing timestamp.

![Fig 2.3 BOT activity detection](image-url)
Spammers can be classified as (1) desperate spammers and (2) sophisticated spammers. Desperate spammers use automatic programs to post multiple tweets with small time difference between posts. Sophisticated spammers create time gap between each tweet. Spammers mostly post duplicate tweets in trending topics such as jumbling the words between tweets, using set of words, including numbers in the topic or appending the topic with commercial advertisement. Tweet similarity index approach determines the behavior of spammers and filters spam.

They first cluster tweets based on user id and then process each user’s set of tweets independently. They create buckets of similar tweets by calculating Jaccard and Levenshtein similarity coefficient. As a result, they have buckets containing most similar tweets together resulting in clusters of similar text. Once all the tweets are collected they check the size of each bucket and if it is greater than one then they considered it as spam.

![Tweet Similarity Index Diagram](image-url)
Levenshtein distance

It is a string metric measurement for calculating the difference between two sequences or text. Informally, the Levenshtein distance between two words is the minimum number of single-character edits required to change one word into the other including insertion, deletion, and substitution. The edit distance phrase is often used to refer Levenshtein distance. The distance is zero if the strings are equal. For example, the Levenshtein distance between "sitter" and "sitting" is 3

- Sitter → sitt (substitution of "i" for "e")
- Sitter → sittin (substitution of "r" for "n")
- Sittin → sitting (insertion of "g" in the end).

This Levenshtein distance is used to find out the duplicate tweets i.e. if the tweets are duplicates then the distance is zero.

Jaccard Index

It is also called Jaccard similarity coefficient. It is used for comparing diversity and similarity of sample sets.

\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} \]  \hspace{1cm} (2.6)

Jaccard Distance measures dissimilarity between sample sets which is obtained by subtracting the jaccard coefficient from 1.

\[ d_j(A, B) = 1 - J(A, B) \]  \hspace{1cm} (2.7)

Dolvara Gunatilaka [9] discusses about two privacy issues. First is user’s identity or user’s anonymity. The second issue is about user profile or personal information leakage.

User anonymity

It is that in many social networking sites users use their real name to represent their account. There are two methods to expose user’s anonymity: (1) de-anonymization attack and (2) neighborhood attack [15]. In the first one, the user’s anonymity can be revealed by history stealing and group membership information while in the second one, the attacker finds the neighbors of the victim node. Based on user’s profile and personal information, attackers are attracted by user’s personal information like their name, date of birth, contact information, relationship status, current work and education background.

There can be leakage of information because of poor privacy settings. Many profiles are made public to others i.e. anyone can view their profile. Next is
leakage of information through third party application. Social networking sites provide an Application Programming Interface (API) for third party developers to create applications. Once users access these applications the third party can access their information automatically.

**Social Worms**

It discuss about some of the social worms. Among those worms Twitter worm is one of the popular worms.

**Twitter worm:** It is a term to describe worms that are spreading through twitter. There are many versions and two worms that are discussed in this paper are:

**Profile Spy worm:** This worm spreads by posting a link that downloads a third party application called “Profile Spy” (a fake application). When users try to download the application they need to fill some personal information which allows attacker to obtain user’s information. Once account is infected, it will continuously tweet malicious messages to their followers. Next twitter worm is Google worm which uses shortened Google URL that tricks the users to click the link. The fake link will redirect users to a fake antivirus website. The website will provide a warning saying that computer got affected and allows user to download the fake antivirus which is actually a malicious code.

**Sender Receiver Relationship**

Jonghyuk Song et al., [7] propose a spam filtering techniques based on sender receiver relationship. This paper addresses two problems in detecting spam. First is account features can be fabricated by spammers. Second is account features cannot be collected until number of malicious messages are reported in that account. The spam filtering does not consider account features rather than it uses relational features i.e. the connectivity and the distance between the sender and receiver. Relational features are difficult to manipulate by spammers. Since twitter limits the tweet to 140 characters spammers cannot put enough information in that. For this reason spammers go for posting URL containing spam. They classify the messages as spam based on the sender. Content filtering is not effective in twitter because it contains small amount of text.

**Restrictions in Twitter**

Some of the restrictions considered in twitter [9] are: The user must not follow large number of users in a short time.

a. Unfollowing and following someone repeatedly.

b. Small number of followers when compared to the amount of following.
c. Duplicate tweets or updates.
d. Update consisting of only links.

The distance between two users is calculated as follows [5][6] when two users are directly connected by an edge, the distance is one. This means that the two users are friends. When the distance is greater than one, they have common friends but not friends themselves. Next the connectivity represents the strength of the relationship. The way to measure connectivity is counting the number of paths. Hence, the connectivity between a spammer and a legitimate user is weaker. The problem of this system is that it identifies messages as normal if it comes from infected friends. Sometimes attackers may send spam messages from legitimate accounts by stealing passwords.

D. Karthika Renuka and T. Hamsapriya [8] an unsolicited email is also called spam and it is one of the fastest growing problems associated with Internet. Among many proposed techniques Bayesian filtering is considered as an effective one against spam. It works based on probability of words occurring in spam and legitimate mails. But keywords are used to detect spam mails in many spam detection system.

In that case misspellings may arise and hence it needs to be constantly updated. But it is difficult to constantly update the blacklist. For this purpose Word Stemming or Hashing Technique is proposed. This improves the efficiency of content based filter. These types of filters are useless if they don’t understand the meaning of the meaning of the words. They have employed two techniques to find out the spam content

**Content based spam filter [10]** This filter works on words and phrases of the email content i.e. associates the word with a numeric value. If this value crosses certain threshold it is considered as spam. This can detect only valid words with correct spellings. This filter uses bayes theorem for detecting the spam content.

**Word stemming or word hashing technique** this filter [12] extracts the stem of the modified word so that efficiency of detecting spam content can be improved. Rule-based word stemming algorithm is used for spam detection. Stemming is an algorithm that converts a word into related form. One such transformation is conversion of plurals into singular, removing suffixes.
3. CONCLUSIONS

Spammers are the major 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. Hence it is efficient to check for spam content in tweets.

REFERENCES


This paper may be cited as: