DON'T FOLLOW ME
Spam Detection in Twitter

Alex Hai Wang
College of Information Sciences and Technology, The Pennsylvania State University, PA 18512, Dunmore, U.S.A.
hwang@psu.edu

Keywords: Social network security, Spam detection, Machine learning, Classification.

Abstract: The rapidly growing social network Twitter has been infiltrated by large amount of spam. In this paper, a spam detection prototype system is proposed to identify suspicious users on Twitter. A directed social graph model is proposed to explore the "follower" and "friend" relationships among Twitter. Based on Twitter's spam policy, novel content-based features and graph-based features are also proposed to facilitate spam detection. A Web crawler is developed relying on API methods provided by Twitter. Around 25K users, 500K tweets, and 49M follower/friend relationships in total are collected from public available data on Twitter. Bayesian classification algorithm is applied to distinguish the suspicious behaviors from normal ones. I analyze the data set and evaluate the performance of the detection system. Classic evaluation metrics are used to compare the performance of various traditional classification methods. Experiment results show that the Bayesian classifier has the best overall performance in term of F-measure. The trained classifier is also applied to the entire data set. The result shows that the spam detection system can achieve 89% precision.

1 INTRODUCTION

Online social networking sites, such as Facebook, LinkedIN, and Twitter, are one of the most popular applications of Web 2.0. Millions of users use online social networking sites to stay in touch with friends, meet new people, make work-related connections and more. Among all these sites, Twitter is the fastest growing one than any other social networking sites, surging more than 2,800% in 2009 according to the report (**?**). Founded in 2006, Twitter is a social networking and micro-blogging service that allows users to post their latest updates, called *tweets*. Users can only post text and HTTP links in their tweets. The length of a tweet is limited by 140 characters.

The goal of Twitter is to allow friends communicate and stay connected through the exchange of short messages. Unfortunately, spammers also use Twitter as a tool to post malicious links, send unsolicited messages to legitimate users, and hijack trending topics. Spam is becoming an increasing problem on Twitter as other online social networking sites are. A study shows that more than 3% messages are spam on Twitter (**?**). The trending topics are also often abused by the spammers. The trending topics, which displays on Twitter homepage, are the most tweeted-about topics of the minute, day, and week on Twitter. The attack reported in (**?**) forced Twitter to temporarily disable the trending topic and remove the offensive terms. I also observed one attack on February 20, 2010 as shown in Figure **??**.

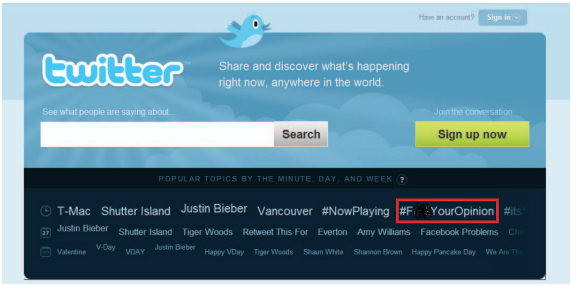


Figure 1: A Twitter trending topic attack on February 20, 2010 (The offensive term is shown in the red rectangle and is censor blurred).

Twitter provides several methods for users to report spam. A user can report a spam by clicking on the "report as spam" link in the their homepage on Twitter. The reports are investigated by Twitter and the accounts being reported will be suspended if they are spam. Another simple and public available method is to post a tweet in the *"@spam @username"* format where *@username* is to mention the spam ac-

count. I tested this service by searching "@spam" on Twitter. Surprisingly the query results show that this report service is also abused by both hoaxes and spam. Only a few tweets report real malicious accounts. Some Twitter applications also allow users to flag possible spam. However, all these ad hoc methods require users to identify spam manually and depend on their own experience.

Twitter also puts effort into cleaning up suspicious accounts and filtering out malicious tweets. Meanwhile, legitimate Twitter users complain that their accounts are mistakenly suspended by Twitter's anti-spam action. Twitter recently admitted to accidentally suspending accounts as a result of a spam clean-up effort (**?**).

In this paper, the suspicious behaviors of spam accounts on Twitter is studied. The goal is to apply machine learing methods to automatically distinguish spam accounts from normal ones. The major contributions of this paper are as follows:

1. To the best of our knowledge, this is the first effort to automatically detect spam on Twitter;
2. A directed graph model is proposed to explore the unique "follower" and "friend" relationships on Twitter;
3. Based on Twitter's spam policy, novel graph-based features and content-based features are proposed to facilitate the spam detection;
4. A series of classification methods are compared and applied to distinguish suspicious behaviors from normal ones;
5. A Web crawler is developed relying on the API methods provided by Twitter to extract public available data on Twitter website. A data set of around 25K users, 500K tweets, and 49M follower/friend relationships are collected;
6. Finally, a prototype system is established to evaluate the detection method. Experiments are conducted to analyze the data set and evaluate the performance of the system. The result shows that the spam detection system has a 89% precision.

The rest of the paper is organized as follows. In Section **??** the related work is discussed. A directed social graph model is proposed in Section **??**. The unique friend and follower relationships are also defined in this part. In Section **??**, novel graph-based and content-based features are proposed based on Twitter's spam policy. Section **??** introduces the method in which I collect the data set. Bayesian classification methods are adopted in Section **??** to detect spam accounts in Twitter. Experiments are conducted in Section **??** to analyze the labeled data. Traditional classification methods are compared to evaluate the performance of the detection ssystem. The conclusion is in Section **??**.

2 RELATED WORK

Spam detection has been studied for a long time. The exsiting work mainly focuses on email spam detection and Web spam detection. In (**?**), the authors are the first to apply a Bayesian approach to filter spam emails. Experiment results show that the classifier achieves a better performance by considering domain-specific features in addition to the raw text of E-mail messages. Currently spam email filtering is a fairly mature technique. Bayesian spam email filters are widely implemented both on modern email clients and servers.

Web is massive and changes more rapidly compared with email system. It is a significant challenge to detect Web spam. (**?**) first formalized the Web spam detection problem and proposed a comprehensive solution to detect Web spam. The TrustRank algorithm is proposed to compute the trust scores of a Web graph. Based on computed scores where good pages are given higher scores, spam pages can be filtered in the search engine results. In (**?**), the authors based on the link structure of the Web proposed a measurement Spam Mass to identify link spamming. A directed graph model of the Web is proposed in (**?**). The authors apply classification algorithms for directed graphs to detect real-world link spam. In (**?**), both link-based features and content-based features are proposed. A basic decision tree classifier is implemented to detect spam. Semi-supervised learning algorithms are proposed to boost the performance of a classifier which only needs small amount of labeled samples in (**?**).

For spam detection in other applications, the authors in (**?**) present an approach for detection of spam calls over IP telephony called SPIT in VoIP system. Based on the popular semi-supervised learning methods, an improved algorithm called MPCK-Means is proposed. In (**?**), the authors study the video spammers and promoters on YouTube. A supervised classification algorithm is proposed to detect spammers and promoters. In (**?**), a machine learning approach is proposed to study spam bots detection in online social networking sites using Twitter as an example. In (**?**), the authors collected three datasets of the Twitter network. The Twitter users' behaviors, geographic growth pattern, and current size of the network are studied.

3 SOCIAL GRAPH MODEL

In this work, Twitter is modeled as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ which consists of a set $\mathcal{V}$ of nodes (vertices) representing user accounts and a set $\mathcal{A}$ of arcs (directed edges) that connect nodes. Each arc is an ordered pair of distinct nodes. An arc $a = (i, j)$ is directed from v_i to v_j which stands for the user i is *following* user j. Following is one of the unique features of Twitter. Unlike most other online social networking sites, following on Twitter is not a mutual relationship. Any user can follow you and you do not have to approve or follow back. In this way, Twitter is modeled as a directed graph.

Since there may or may not be an arc in either direction for a pair of nodes, there are four possible states for each dyad. Four types of relationships on Twitter are defined as follows:

First, followers represent the people who are following you on Twitter. In this paper, follower in Twitter's graph model is defined as:

Definition 1 ***(Follower).*** *Node v_j is a* follower *of node v_i if the arc $a = (j,i)$ is contained in the set of arcs, $\mathcal{A}$.*

Based on the definition, followers are the incoming links, or *inlinks*, of a node. Let the set $\mathcal{A}_i^I$ denote the inlinks of node v_i, or the followers of user i.

Second, Twitter defines friends as the people whose updates you are subscribed to. In other words, friends are the people whom you are following. I give a formal definition of the friend relationship in graph model as follows:

Definition 2 ***(Friend).*** *Node v_j is a* friend *of node v_i if the arc $a = (i, j)$ is contained in the set of arcs, $\mathcal{A}$.*

Friends are the outgoing links, or *outlinks*, of a node. Let the set $\mathcal{A}_i^O$ denote the outlinks of node v_i, or the friends of user i.

Third, a novel relationship on Twitter, mutual friend, is proposed. If two users are following each other, or are the friends of each other, the relationship between these two users is mutual friend. A formal definition of the mutual friend relationship on Twitter is defined as follows:

Definition 3 ***(Mutual Friend).*** *Node v_i and node v_j are* mutual friends *if both arcs $a = (i, j)$ and $a = (j,i)$ are contained in the set of arcs, $\mathcal{A}$.*

Since a mutual friend is your follower and friend at the same time, the set of mutual friends is the intersection of the set of friends and the set of followers. If let $\mathcal{A}_i^M$ denote the set of mutual friends of node v_i, the following holds: $\mathcal{A}_i^M = \mathcal{A}_i^I \cap \mathcal{A}_i^O$.

Finally, two users are strangers if there is no connection between them. A formal definition is as follows:

Definition 4 ***(Stranger).*** *Node v_i and node v_j are* strangers *if neither arcs $a = (i, j)$ nor $a = (j,i)$ is contained in the set of arcs, $\mathcal{A}$.*

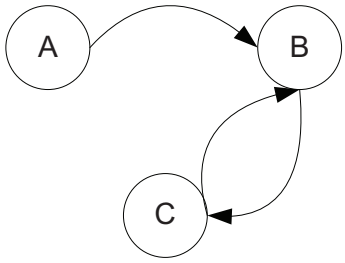


Figure 2: A simple Twitter graph.

A simple Twitter social graph example is shown in Figure **??** where user A is following user B, and user B and user C are following each other. Based on our definitions, user A is a follower of user B. User B is a friend of user A. User B and User C are mutual friends. User A and user C are strangers.

Based the directed social graph model proposed above, a real Twitter social graph is shown in Figure **??**. 20 random users and their followers and friends are collected from Twitter's public timeline and the figure is drawn using Pajek software (**?**).

4 FEATURES

In this section, the features extracted from each Twitter user account for the purpose of spam detection are introduced. The features are extracted from different aspects which include graph-based features and content-based features. Based on the unique characteristics of Twitter, novel features are also proposed in this section.

4.1 Graph-based Features

One important function of twitter is that you can build your own social network by following friends and allowing others to follow you. Spam accounts try to follow large mount of users to gain their attention. The twitter's spam and abuse policy (**?**) says that, "if you have a small number of followers compared to the amount of people you are following", it may be considered as a spam account.

Three features, which are the number of friends, the number of followers, and the reputation of a user,

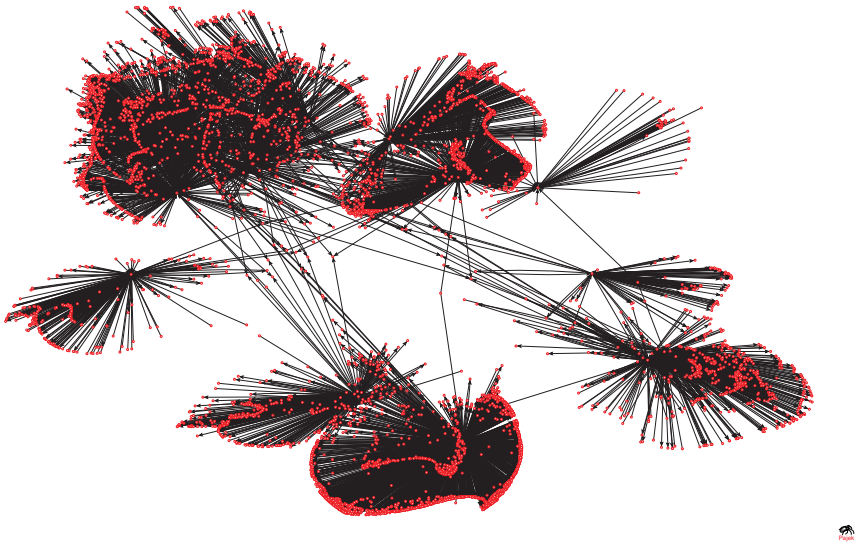


Figure 3: Twitter social graph.

are computed to detect spam from this aspect. According to the social graph model proposed in Section **??**, the *indegree* $d_I(v_i)$ of a node v_i, which is the number of nodes that are adjacent to node v_i, stands for the feature of the number of followers. The feature of the number of friends is represented by the *outdegree* $d_O(v_i)$ of a node v_i, which is the number of nodes that are adjacent to v_i.

Furthermore, a novel feature, *reputation*, is proposed to measure the relative importance of a user on Twitter. The reputation R is defined as the ratio between the number of friends and the number of followers as:

$$R(v_i) = \frac{d_I(v_i)}{d_I(v_i) + d_O(v_i)} \quad (1)$$

Obviously if the number of followers is relatively small compared to the amount of people you are following, the reputation is small and close to zero. At the same time the probability that the associated account is spam is high.

4.2 Content-based Features

4.2.1 Duplicate Tweets

An account may be considered as a spam if you post duplicate content on one account. Usually legitimate users will not post duplicate updates.

Duplicate tweets are detected by measuring the Levenshtein distance (**?**) (also known as edit distance) between two different tweets posted by the same account. The Levenshtein distance is defined as the minimum cost of transforming one string into another through a sequence of edit operations, including the deletion, insertion, and substitution of individual symbols. The distance is zero if and only if the two tweets are identical.

Figure 4: A Twitter spam page (Duplicate tweets are circled in the same color rectangles).

A typical Twitter spam page is shown in Figure **??**.

As can be seen, spammers often include different *@usernames* in their duplicate tweets to avoid being detected. This is also an efficient way to spam legitimate users, since Twitter automatically collects all tweets containing your *@username* for you. The example in Figure **??** also shows that spammers include different *#topics* and "http" links in their duplicate tweets. Because of the URL shortening service, such as bit.ly, the different "http" links may have the same destination. Based on these reasons, when the Levenshtein distance is calculated between different tweets, I clean the data by stopping the words containing "@", "#", "http://", and "www." in the tweets. In other words, the username information, topic information, and links are ignored. Instead only the content of the tweets is considered. As shown in Figure **??**, the duplicate tweets are circled in the same color rectangles, although they have different *@username*, *#topic*, and links.

After cleaning the data, the pairwise Levenshtein distance is calculated in the user's 20 most recent tweets. If the distance is smaller than a certain threshold, it is counted as one duplicate. In this paper, the threshold is set to zero, which means two tweets are considered as duplicate only when they are exactly the same.

4.2.2 HTTP Links

Malicious links can spread more quickly than ever before because of Twitters lightning-fast communications platform. Twitter filters out the URLs linked to known malicious sites. However, a great vulnerability is the presence of shorten URLs. Twitter only allows users to post a short message within 140 characters. URL shortening services and applications, such as bit.ly, become popular to meet the requirements. Shorten URLs can hide the source URLs and obscure the malicious sites behind them. As a result it provides an opportunity for attackers to prank, phish, and spam. While Twitter does not check these shorten URLs for malware, it is considered as spam if your updates consist mainly of links, and not personal updates according to Twitter's policy.

The number of links in one account is measured by the number of tweets containing HTTP links in the user's 20 most recent tweets. If a tweet contains the sequence of characters "http://" or "www.", this tweet is considered containing a HTTP link.

4.2.3 Replies and Mentions

A user is identified by the unique username and can be referred in the *@username* format on Twitter. The *@username* creates a link to the user's profile automatically. You can send a reply message to another user in *@username+message* format where *@username* is the message receiver. You can reply to anyone on Twitter no matter they are your friends/followers or not. You can also mention another *@username* anywhere in the tweet, rather than just the beginning. Twitter automatically collects all tweets containing your username in the *@username* format in your replies tab. You can see all replies made to you, and mentions of your username.

The reply and mention are 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. Twitter also considers this as a factor to determine spamming. The number of replies and mentions in one account is measured by the number of tweets containing the "@" symbol in the user's 20 most recent tweets.

4.2.4 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 hashtag, which is the # symbol followed by a term describing or naming the topic, to a tweet. If there are many tweets containing the same term, that helps the term to become a trending topic. The topic shows up as a link on the home page of Twitter as shown in Figure **??**.

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. Twitter also considers an account as spam "if you post multiple unrelated updates to a topic using the # symbol". The number of tweets which contains the hashtag # in a user's 20 most recent tweets is measured as a content-based feature.

5 DATA SET

To evaluate the spam detection methods, a real data set is collected from Twitter website. First I use Twitter's API methods to collect user's detailed information. Second, because no Twitter API method could provide information of a specific unauthorized user's recent tweets, a Web crawler is developed to extra a specific unauthorized user's 20 most recent tweets.

5.1 Twitter API

First I use the *public_timeline* API method provided by Twitter to collect information about the non-protected users who have set a custom user icon in real time. This method can randomly pick 20 non-protected users who updated their status recently on Twitter. Details of the user, such as IDs, screen name, location, and etc, are extracted. At the same time, I also use social graph API methods *friends* and *followers* to collect detailed information about user's friends and followers, such as the number of friends, the number of followers, list of friend IDs, list of follower IDs, and etc. The *friends* and *followers* API methods can return maximum 5,000 users. If a user has more than 5,000 friends or followers, only a partial list of friends or followers can be extracted. Based on the observation, the medians of the number of friends and followers are around 300, so this constraint does not affect the method significantly.

Another constraint of Twitter API methods is the number of queries per hour. Currently the rate limit for calls to the API is 150 requests per hour. This constrain affects the detection system significantly. To collect data from different time and avoid congesting Twitter, I crawl Twitter continuously and limit 120 requests per hour per host.

5.2 Web Crawler

Although Twitter provides neat API methods for us, there is no method that allows us to collect a specific unauthorized user's recent tweets. The *public_timeline* API method can only return the most recent update from different non-protected users (one update per user). The *user_timeline* API method can return the 20 most recent tweets posted from an authenticating user. The recent tweets posted by a user are important to extract content-based features, such as duplicate tweets. To solve this problem, a Web crawler is developed to collect the 20 most recent tweets of a specific non-protected user based on the user's ID on Twitter.

The *public_timeline* API method is first used to collect the user's IDs of 20 non-protected users who updated their status recently. Based on the user's IDs, the Web crawler extracts the user's 20 most recent tweets and saves it as a XML file.

A prototype system structure is shown in Figure **??**. Currently 3 Web crawlers extract detailed user information from Twitter website. The raw user tweets are stored as XML files. Other user information, such as IDs, list of friends and followers, are saved in a relational database. The graph-based features are calculated at the same time and stored in the relational database. The XML files are parsed and the content-based features are calculated. The results are saved in the relational database.

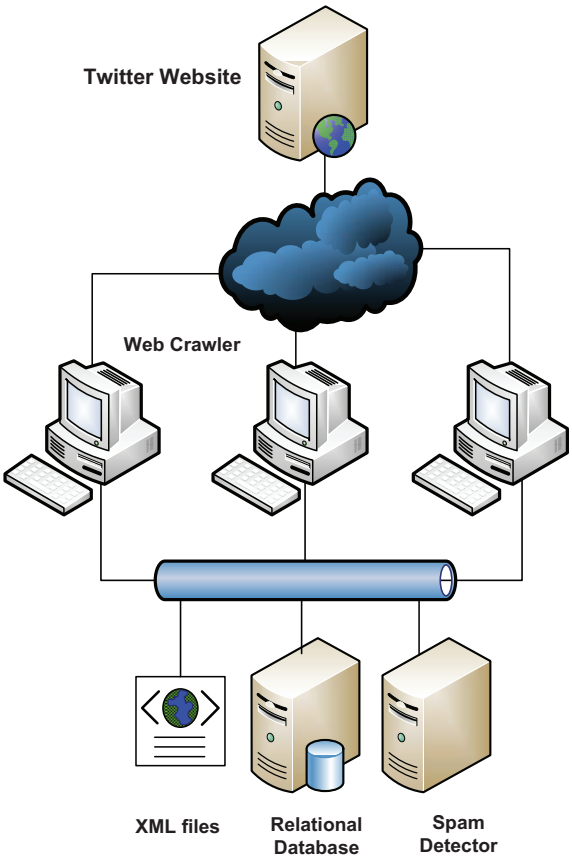


Figure 5: Twitter spam detection system.

Finally, I collect the data set for 3 weeks from January 3 to January 24, 2010. Totally 25,847 users, around 500K tweets, and around 49M follower/friend relationships are collected from the public available data on Twitter.

6 SPAM DETECTION

Several classic classification algorithms, such as decision tree, neural network, support vector machines, and k-nearest neighbors are compared. The naïve Bayesian classifier outperforms all other methods for several reasons. First, Bayesian classifier is noise robust. On Twitter, the relationship between the feature set and the spam is non-deterministic as discussed in Section **??**. An account cannot be predicted as spam with certainty even though some of its features are identical to the training examples. Bayesian classi-

fier treats the non-deterministic relationship between class variables and features as random variables and captures their relationship using posterior probability. While other methods cannot tolerate this kind of noisy data or confounding factors, such as decision tree.

Another reason that Bayesian classifier has a better performance is that the class label is predicted based on user's specific pattern. A spam probability is calculated for each individual user based its behaviors, instead of giving a general rule. Also, naïve Bayesian classifier is a simple and very efficient classification algorithm.

The naïve Bayesian classifier is based on the well-known Bayes theorem:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (2)$$

The conditional probability of $P(Y|X)$ is also known as the posterior probability for Y, as opposed to its prior probability $P(Y)$.

Each Twitter account is considered as a vector X with feature values. The goal is to assign each account to one of two classes Y: spam and non-spam. The big assumption of naïve Bayesian classifier is that the features are conditionally independent, although research shows that it is "is surprisingly effective in practice" without the unrealistic independence assumption (**?**). With the conditional independence assumption, we can only estimate each conditional probability independently, instead of trying every combination of X.

To classify a data record, the posterior probability is computed for each class:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(X)} \quad (3)$$

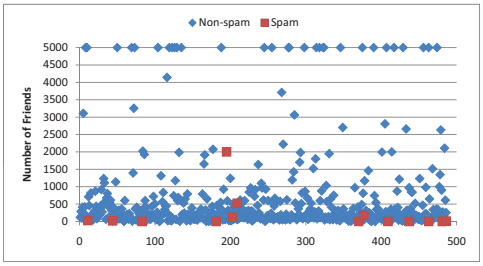
Since $P(X)$ is a normalizing factor which is equal for all classes, we need only maximize the numerator $P(Y)\prod_{i=1}^{d} P(X_i|Y)$ in order to do the classification.

7 EXPERIMENTS

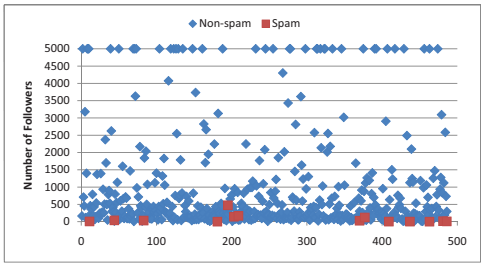
To evaluate the detection method, 500 Twitter user accounts are labeled manually to two classes: spam and non-spam. Each user account is manually evaluated by reading the 20 most recent tweets posted by the user and checking the friends and followers of the user. The results show that there are around 1% spam accounts in the data set. The study in (**?**) shows that there is probably 3% spam on Twitter. To simulate the reality and avoid the bias in the crawling and label methods, additional spam data are added to the data set. I search "@spam" on Twitter to collect additional spam data. Only a small percentage of results report real spam. I clean the query results by manually evaluating each spam report. Finally the data set is mixed to contain around 3% spam data.

7.1 Data Analysis

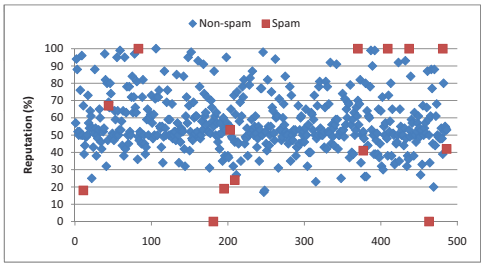
Graph-based Features. Figure **??** show the graph-based features proposed in Section **??**. The number of friends for each Twitter account is shown in Figure **??**. Twitter spam policy says that "if you have a small number of followers compared to the amount of people you are following", you may be considered as a spam account. As can be seen, not all spam accounts



(a) The number of friends (the maximum number of friends is 5,000 which is the maximum return value of Twitter *friends* API method).



(b) The number of followers (the maximum number of followers is 5,000 which is the maximum return value of Twitter *followers* API method).



(c) The reputation.

Figure 6: Graph-based features.

follow a large amount of user as we expected, instead only 30% of spam accounts do that. The reason is that

Twitter allows users to *mention* or *reply* any other user in their tweets. In other words, the spammers do not need to follow legitimate user accounts to draw their attention. The spammers can simply post spam tweets and *mention* or *reply* another user in the @username format in the tweets. These tweets will appear on the user's replies tab whose username is mentioned. In this way, the spam tweets are sent out without actually following a legitimate user. The results show that this is an efficient and common way to spam other users as shown in Figure **??**.

Figure **??** shows the number of followers for each Twitter account. As we expected, usually the spam accounts do not have a large amount of followers. But still I can find there are some spam accounts having a relatively large amount of followers. They may achieve that by letting other spam accounts to follow them collusively or lure legitimate users to follow them.

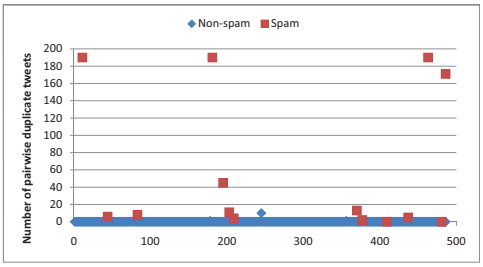
The reputation for each Twitter account is shown in Figure **??**. Surprisingly I find that some spam accounts have a 100% reputation. The reason is as mentioned above that the spam accounts do not have to follow a legitimate user to send malicious tweets. Because of this, some spam accounts do not have a friend ($d_O(v_i) = 0$). However, the reputation feature shows the abnormal behaviors of spam accounts. Most of them either have a 100% reputation or a very low reputation. The reputation of most legitimate users is between 30% to 90%.

Content-based Features. The content-based features proposed in Section **??** are shown in Figure **??**. Twitter spam policy indicates that "multiple duplicate updates on one account" is factor to detect spam. The number of pairwise duplication in a user's 20 most recent tweets is shown in Figure **??**. As expected, most spam accounts have multiple duplicate tweets. This is an important feature to detect spam. However, as shown in the figure, not all spam accounts post multiple duplicate tweets. So we can not only depend on this feature to detect spam.

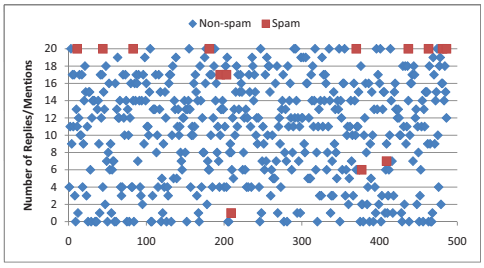
The number of mentions and replies is shown in Figure **??**. As expected, most spam accounts have the maximum 20 "@" symbol in their 20 most recent tweets. This indicates that the spammers intend to mention or reply legitimate users in their tweets to gain attention. This will lure legitimate users to either read their spam messages or even click the malicious links in their tweets.

Figure **??** shows the number of links in each user's 20 most recent tweets. The results show that most spam accounts have the maximum 20 links in their 20 most recent tweets. In other words, each tweet contains a link for most spam accounts. However, the results also show that some legitimate users also include links in all tweets. The reason is that some companies join Twitter to promote their own web sites. Usually they will include a link to their own web page in each of their tweets.

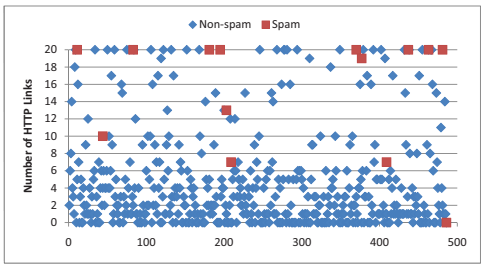
Finally, Figure **??** shows the number of "#" tag signs in each user's 20 most recent tweets. Although spamming Twitter trend topics is reported in news, I cannot find that spammers attack trend topics in the dataset. The reason is that this kind of attack usu-



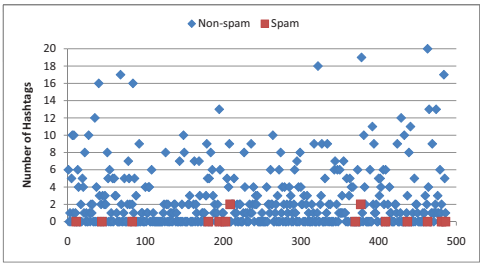
(a) The number of pairwise duplications.



(b) The number of mention and replies.



(c) The number of links.



(d) The number of hashtags.

Figure 7: Content-based features.

Table 1: Classification evaluation.

Classifier	Precision	Recall	F-measure
Decision Tree	0.667	0.333	0.444
Neural Networks	1	0.417	0.588
Support Vector Machines	1	0.25	0.4
Naïve Bayesian	0.917	0.917	0.917

ally occur in a very short period of time and does not happen constantly on Twitter. It is difficult for us to capture their trace. It does not mean that this kind of attack is not common or not even exist.

7.2 Evaluation

The evaluation of the overall process is based on a set of measures commonly used in machine learning and information retrieval. Given a classification algorithm, I consider its confusion matrix:

		Prediction	
		Spam	Not Spam
True	Spam	a	b
	Not Spam	c	d

where a represents the number of spam examples that were correctly classified, b represents the spam examples that were falsely classified as non-spam , c represents the number of non-spam examples that were falsely classified as spam, and d represents the number of non-spam examples that were correctly classified. I consider the following measures: precision, recall, and F-measure where the precision is $P = a/(a+c)$, the recall is $R = a/(a+b)$, and the F-measure is defined as $F = 2PR/(P+R)$. For evaluating the classification algorithms, I focus on the F-measure F as it is a standard way of summarizing both precision and recall.

All the predictions reported in this paper are computed using 10-fold cross validation. For each classifier, the precision, recall, and F-measure are reported. Each classifier is trained 10 times, each time using the 9 out of the 10 partitions as training data and computing the confusion matrix using the tenth partition as test data. I then average the resulting ten confusion matrices and estimate the evaluation metrics on the average confusion matrix. The evaluation results are shown in Table **??**. The naïve Bayesian classifier has the best overall performance compared with other algorithms, since it has the highest F score.

Finally, the Bayesian classifier learned from the labeled data is applied to the entire data set. As mentioned in Section **??**, information about totally 25,817 users was collected. It is nearly impossible for us to label all the data. Instead I only manually check the users who are classified as spam by the Bayesian classifier. 392 users are classified as spam by the detection system. I check the spam data by manually reading their tweets and checking their friends and followers. The results show that 348 users are real spam accounts and 44 users are false alarms. This means that the precision of the spam detection system is $89\% = 348/392$.

8 CONCLUSIONS

In this paper, I study the spam behaviors in a popular online social networking site, Twitter. To formalize the problem, a directed social graph model is proposed. The "follower" and "friend" relationships are defined in this paper. Based on the spam policy of Twitter, novel content-based and graph-based features are proposed to facilitate spam detection. Traditional classification algorithms are applied to detect suspicious behaviors of spam accounts. A Web crawler using Twitter API methods is also developed to collect real data set from public available information on Twitter. Finally, I analyze the data set and evaluate the performance of the detection system.

The results show that among the graph-based features, the proposed reputation feature has the best performance of detecting abnormal behaviors. No many spam accounts follow large amount of users as we expected. Also some spammers have many followers.

For the content-based features, most spam accounts have multiple duplicate tweets. This is an important feature to detect spam. However, not all spam account post multiple duplicate tweets and some legitimate users also post duplicate tweets. In this way we can not only rely on this feature. The results also show that almost all spam tweets contain links and reply sign "@".

Finally, several popular classification algorithms are studied and evaluated. The results show that the Bayesian classifier has a better overall performance with the highest F score. The learned classifier is applied to large amount of data and achieve a 89% precision.

REFERENCES

Analytics, P. (2009). Twitter study. http://www.pearanalytics.com/wp-content/uploads/2009/08/Twitter-Study-August-2009.pdf.

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., and Gonçalves, M. (2009). Detecting spammers and content promoters in online video social networks. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 620–627, New York, NY, USA. ACM.

Castillo, C., Donato, D., Gionis, A., Murdock, V., and Silvestri, F. (2007). Know your neighbors: web spam detection using the web topology. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 423–430, New York, NY, USA. ACM.

CNET (2009). 4chan may be behind attack on twitter. http://news.cnet.com/8301-13515_3-10279618-26.html.

Geng, G.-G., Li, Q., and Zhang, X. (2009). Link based small sample learning for web spam detection. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 1185–1186, New York, NY, USA. ACM.

Gyongyi, Z., Berkhin, P., Garcia-Molina, H., and Pedersen, J. (2006). Link spam detection based on mass estimation. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*, pages 439–450. VLDB Endowment.

Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. (2004). Combating web spam with trustrank. In *VLDB '04: Proceedings of the Thirtieth international conference on Very large data bases*, pages 576–587. VLDB Endowment.

Krishnamurthy, B., Gill, P., and Arlitt, M. (2008). A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA. ACM.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Nooy, W. d., Mrvar, A., and Batagelj, V. (2004). *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, New York, NY, USA.

Opera (2009). State of the mobile web. http://www.opera.com/smw/2009/12/.

Rish, I. (2005). An empirical study of the naive bayes classifier. In *IJCAI workshop on Empirical Methods in AI*.

Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. (1998). A bayesian approach to filtering junk e-mail. In *AAAI Workshop on Learning for Text Categorization.*

Twitter (2009a). Restoring accidentally suspended accounts. http://status.twitter.com/post/136164828/restoring-accidentally-suspended-accounts.

Twitter (2009b). The twitter rules. http://help.twitter.com/forums/26257/entries/18311.

Wang, A. H. (2010). Detecting spam bots in online social networking websites: A machine learning approach. In *24th Annual IFIP WG 11.3 Working Conference on Data and Applications Security*.

Yu-Sung, W., Bagchi, S., Singh, N., and Wita, R. (2009). Spam detection in voice-over-ip calls through semi-supervised clustering. In *DSN '09: Proceedings of the 2009 Dependable Systems Networks*, pages 307 –316.

Zhou, D., Burges, C. J. C., and Tao, T. (2007). Transductive link spam detection. In *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 21–28, New York, NY, USA. ACM.