Wanchun Zhao. Spam Detection on Twitter: A Comparison between Content-based and Graph-based Features. A Master's Paper for the M.S. in I.S degree. March, 2016. 44 pages. Advisor: Bradley M. Hemminger

The popularity of social media has triggered the development of spammers, which produces useless information and costs normal user more time in information seeking process. In this paper, Twitter is studied as an example of spam detection in social media. Using Twitter APIs, content-based and graph-based features were extracted from datasets and analyzed with users' level of spam. Combining two kinds of features with J48, NaïveBayes and SVM classifiers, content-based features with J48 have the best performance in evaluation.

Headings:

Spam Detection

Twitter

Content-based

Graph-based

Machine Learning

SPAM DETECTION ON TWITTER: A COMPARISON BETWEEN CONTENT-BASED AND GRAPH-BASED FEATURES

by
Wanchun Zhao

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2016

Approved by

Bradley M. Hemminger

Table of Contents

INTRODUCTION

With the development of Internet and Information technology, social media services have become an irreplaceable part in people's lives. Billions of users post messages, share pictures and connect with other users through websites and mobile applications like Facebook, Twitter or Instagram. The popularity of social media services has reflected a new way for people to communicate with each other, but it also triggers a problem, which is the rapid proliferation of spam among social media sites.

According to Wikipedia, spamming is the use of electronic messaging systems to send unsolicited messages, especially advertising, as well as sending messages repeatedly on the same site (Spamming, 2016). The most widely recognized form of spam is email spam, but the flexibility and popularity of social network service has provided spammers another way to spread spam messages.

There are several types of spammers existing in social media sites. The most common ones attach a URL in their posts or messages. The links in the posts redirect users to unrelated websites, illegal contents, or even computer viruses and phishing websites. There are also spammers posting advertisements or inappropriate images for publication or spreading rumors for attention. The existence of these spammers in social network sites produces lots of useless information, exposes users to content they do not wish to see, costing users more time in information seeking process and sometimes even get users into financial loss and identity security issues. Therefore it is important to come up with a way to filter those spammers and spam information to create clearer environment for users.

Social media services like Twitter have already been working on the spam problem but more work is needed to find effective spam filters. In addition, scholars have also focused on this issue and tried to extract features to identify spam accounts utilizing machine learning and data mining methods. However, there are few studies concentrating on summarizing proposed approaches and compare the strengths and weaknesses of each algorithm. Therefore this paper is aimed at finding and analyzing features that is able to identify spam accounts, and also comparing two prediction methods employed by researchers: content-based and graph-based, by using real Twitter data collection.

The remainder of the paper is organized as follows. In Section II, past studies on spam analysis and spam detection are reviewed. How the data sets are collected and cleaned is described in Section III. Then in Section IV and Section V, the results of descriptive analysis of extracted features, the correlation test between features and spam level and the accuracy of prediction are displayed. Finally, the Section VI and Section VII will demonstrate the findings of this paper and some limitations in this research.

Literature Review

This section provides an overview of spammers' activities on Twitter, discusses the results of spam analysis as well as approaches, mechanisms and systems to detect spammers proposed by previous studies.

A. **Twitter Spam**

One of the most successful and popular social networking services in United States is Twitter, which is an online service that enables users to send and read short 140-character messages (Twitter, 2015). The twitter platform's main functions include:

(1) Tweet: users post short messages to let followers or sometime strangers see and comment on. URL links and images are allowed to be included in tweets.

(2) Follow: a relationship that users maintain with which followers could see tweets of the user he follows in his own twitter home page.

(3) Mention: users mention other users in the tweets so that either followings or non-followings could see the contents of that certain tweet.

(4) Retweet: users repost of other users' tweet.

(5) Direct message: users send messages to the user he follows privately and the following user would get notification of message.

Based on twitter's main functions, there are mainly four types of strategies that spammers employ in twitter:

(1) Including malicious URLs in tweets. This type of spammers usually post a link in the tweet. Some URLs redirect users to unrelated websites to gain website visits while other URLs might get computer infected with virus and get users into identity theft.

(2) Posting advertisements. This type of spammers usually post pictures or videos of commercial products in tweets.

(3) Including inappropriate contents in tweets. This kind of spam constantly includes inappropriate contents in tweets, like fake news, rumors or pornographic content etc.

(4) Sending disturbing messages. Spammers send direct messages to users to advertise their products or other disturbing contents.

Usually, spammers combine several strategies together in their daily activities.

B. Spam Analysis

Spam analysis usually focuses on the features of spam accounts and the comparison between spam and normal users. In spam analysis studies, researchers extract features and employ descriptive analysis or statistical analysis to study spam accounts' activities, show the difference and determine if one feature or a combination of features is able to differentiate spammers with non-spammers.

Wang, Navathe et al. (Wang, et al., 2013) collected short URLs from Twitter and retrieved click traffic data from Bitly. After analyzing and comparing the click traffic generated and determining the top click sources for spam and non-spam short URLs, the results show that the majority of the clicks are from direct sources and that the spammers utilize popular websites to attract more attention by cross-posting the links. Similarly, Lin and Huang (Lin & Huang, 2013) evaluated the common features to see how effective they are to detect Twitter spam accounts with collected datasets and have found that features like number of words per tweet do not show significant difference between spammers and regular users while the URL rate and the interaction rate features are effective in detecting spam. Song, Lee and Kim (Song, Lee, & Kim, 2011) proposed a novel spam filtering system using relation features, such as the distance and connectivity between a message sender and a receiver to decide whether the current message is spam or not, because account features can easily be fabricated by spammers.

Some studies directly analyze the behavior of spammers, studying how they behave and exist in Twitter. Thomas, Grier et al. studied over 1.1 million accounts suspended by Twitter and observed the difference among human, bot, and cyborg in terms of tweeting behavior, tweet content, and account properties (Thomas, Grier, Song, & Paxson, 2011). The results showed that 77% of spam accounts identified by Twitter are suspended within on day of their first tweet but new fraudulent accounts are created to take their places. Less than 9% of spam accounts form social relationships with regular Twitter users. 17% of spam accounts rely on hijacking trends, while 52% of accounts use unsolicited mentions to reach an audience.

Stringhini, Kruegel and Vigna used another way to study spammers. They created a number of honeypot-profiles in Facebook, MySpace and Twitter to attract spammers in order to study how spammers operate (Stringhini, Kruegel, & Vigna, 2010). They periodically connected to those accounts and collected spammers' behavior data. After analyzing anomalous behavior of spammers, they developed six features to identify spam account, including FF ratio (ratio of followers over followings), URL ratio, message similarity, choices of friends, messages sent and number of friends.

C. Spam Detection

Spam detection studies proposed methods to identify or predict spam among social networking sites, which are usually based on the analysis of spam account features. Most of related studies extracted features to create user profile and apply to machine learning or data mining methods to distinguish spammers with normal users.

Common features used in the models include user behavior features, content-based features and graph-based features. User behavior features capture user activities on Twitter network, like posting frequency, timeline of user activities and social interactions. While content-based features focus more on the text of tweets submitted by users, including URLs, keywords, mentions, hashtags etc. Graph-based features depict the following/followed relationship between users in twitter and sometimes also classified as user behavior features. Researchers usually combine multiple types of features to predict spam.

Most of the studies employed supervised learning methods, usually classification. Benevenuto, Magno et al. (Benevenuto, Magno, Rodrigues, & Almeida, 2010) picked three trending topics in twitter and crawled relevant tweet and user information, manually classifying spammer and non-spammer accounts in datasets. Then they proposed a SVM classifier with content attributes like number of hashtags per number of words on each tweet, number of URLs per words, number of words of each tweet etc. and user behavior attributes like number of tweets, age of the user account, number of times the user was mentioned, number of times the user was replied to etc. for spam detection. Approximately 70% of spammers and 96% of non-spammers were correctly classified in their experiment.

Similarly, McCord and Chuah (McCord & Chuah, 2011) discussed some features that differentiate spammers ad non-spammers, like number of followings and followers, distribution of tweets over 24-hour period, replies/mentions, keywords/wordweight etc. and used Twitter API methods to crawl active Twitter users, their followers/following information and their most recent 100 tweets. Then they employed Random Forest, Naïve Bayesian, Support Vector Machine and K-nearest neighbor four classifiers to identify spammers with datasets and compared accuracy of each classifier. Their results show that among the four classifiers, the Random Forest classifier produces the best results, which can achieve 95.7% precision and 95.7% F- measure using the Random Forest classifier.

Some researchers emphasized more on graph-based features to create a network model among users and detect spam. (Wang A. H., 2010) established a social graph model with four kinds of relationships (follower, friend, mutual friend and stranger) between accounts in Twitter, viewing each account as a node and relationship as edge. Then he used Decision Tree, Neural Networks, Support Vector Machines and Naïve Bayesian classifier to classify labelled accounts and evaluate each machine learning method.

Besides classification, some studies applied unsupervised learning methods like clustering. Miller et al. (Miller, Dickinson, Deitrick, Hu, & Wang, 2014) viewed spam detection as an anomaly detection problem. It introduced 95 one-gram features from tweet text alongside the user information analyzed in previous studies and used two stream clustering algorithms: StreamKM++ and DenStream to cluster normal Twitter users and treat outliers as spammers. Each of these algorithms performed well individually and the conjunction reached 100% recall and a 2.8% false positive rate. Tan and Guo et al. (Tan, Guo, Chen, Zhang, & Zhao, 2013) designed an unsupervised spam detection scheme which works by deliberately removing non-spammers from the network, leveraging both the social graph and the user-link graph. The underpinning of the system is that while spammers constantly change their patterns to evade detection, non-spammers do not have to do so and thus have a relatively non-volatile pattern, which outperforms existing schemes.

The studies mentioned above have all come up with methods using features to detect spam among social media sites but there are still not enough studies digging in the strengths and weaknesses of each feature and method as well as comparison analysis of existing algorithms. Therefore this paper will focus on comparing two main models used in spam detection: content-based and graph-based and explained relative strengths and weaknesses of the approaches in particular situations.

RESEARCH METHOD

This section describes what methods will be used to compare two algorithms, how the experiment data sets were collected from twitter and the preliminary analysis of the data sets.

A. Research Method

In order to study spam detection among social networking services, this paper employs experimental methods to use Twitter as an example and collects user accounts and interaction data from Twitter public API as datasets for analysis.

The datasets include user account information, tweet information, timeline, relationship between users etc. Each account in datasets would be manually judged as several levels of spam, from non-spam to total spam.

After data cleaning process, each feature from the datasets is analyzed to see if there is significant association between the feature and if the user is spam or not, and why the feature show/don’t show the differences between spammers and non-spammers. After that, the experiment will implement two existing classification algorithms using content-based and graph-based features accordingly, and combined with different classifiers provided by machine learning tool weka to predict whether an account is spam or not.

The metrics that evaluate the performance of each algorithm are the precision and recall of predicting results compared with human judgments.

B. Data Collection

Twitter has several public APIs for developers to access authorized users' data on Twitter. Among those APIs, the REST APIs provide programmatic access to read and write Twitter data, including authoring a new tweet, reading author profile and following data etc. (REST APIs, 2016). The Streaming APIs give developers low latency access to Twitter's global stream of Tweet data (Streaming APIs, 2016).

Due to the data sets needed in the experiment, I first used Streaming APIs to collect a list of Twitter users' id, which is unique to each user, and then selected samples from collected list randomly. Then I employed REST APIs to extract sample users' name, tweets, tweet creation time, platform used to post tweets and the number of users' followings and followers. In this process, I wrote Python scripts to connect to APIs and automatically extract data. Specifically, StreamListener instance and tweepy.api's user_timeline function, friends_ids function and followers_ids function in tweepy package were used to extract needed features. During the data acquiring process, the extracted data was stored in text files and then used MySQL Bulk Loader to save in MySQL database.

The Streaming process was conducted on December 25th, 2015 and extracted 646,032 user ids. Using python to generate random numbers, I selected 516 users as samples from the data sets. The sample users account information and tweet information were extracted between January 19^{th}, 2016 to January 21^{st}, 2016. Extracted datasets include user id, user name, 20 tweets of each user, tweet creation time and tweeting platform, total 10320 tweets. The following and follower numbers were extracted from APIs between February 7^{th}, 2016 and February 8^{th}, 2016. As some sampled accounts were suspended by Twitter during the process, only 501 users' following and follower information were acquired.

C. Spam Label

As the boundary of spammers and non-spammers is ambiguous, it is sometimes difficult to judge if an account is spammer or not. In order to manually label each account, I have divided sample accounts in several spam level. Each level corresponds to several situations and the higher level the account belongs to, the more it is likely to be spammer. The level is defined based on the possible harm one account could do other accounts on twitter. Descriptions of spam level are listed below:

(1) Level 0(Not Spam): normal twitter user accounts. Accounts only include normal and regular activities of twitter users.

(2) Level 1(Slightly Spam): twitter accounts that contain meaningless/repeated contents, but do not disturb other users' activities. Or official publication accounts post promotional contents.

Figure 1 Examples of Level 1 Accounts

(3) Level 2(Likely Spam): twitter accounts that contain promotion contents but not official account of one company or personal brand. Or accounts that contain URLs linking to another website, trying to sell things to other users.

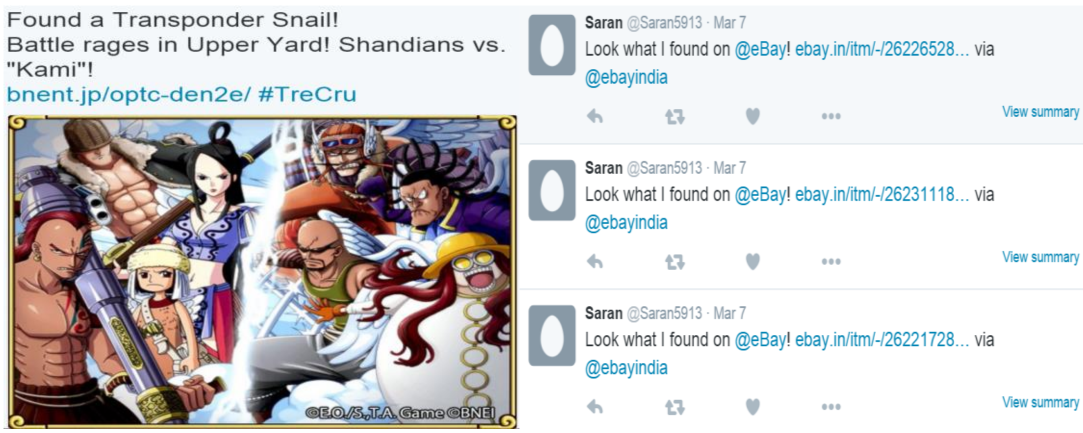

Figure 2 Examples of Level 2 Accounts

(4) Level 3(Spam): twitter accounts that post inappropriate contents, including pornographic and violent images, or URLs link to viruses/dangerous/phishing websites.

In summary, 242 users (46.90%) from samples were labelled as level 0, 167 users (32.36%) users were level 1, 77 users (14.92%) were labelled as level 2, and 30 users (5.82%) were level 3.

FEATURE ANALYSIS

Based on previous studies and extracted data sets, 9 features were used to detect spam on Twitter. Among all features, 5 features belong to content-based features, including URL rate, mention rate, hashtag rate, word count and spam word rate. 3 features belong to graph-based features, including number of followings, number of followers and reputation, and also the platform feature.

A. Content-based Features

a. URL Rate

URL Rate is the average number of URLs contained in each user's tweets. In the datasets extracted, URL is formed as a string which begins with "http". Therefore to calculate this metric, I used python to sum up the total number of "http" string in tweet texts of each user and divide this number by number of tweets. In addition, as some users used third-party platform to post or share tweets, like Facebook, Youtube or Instagram, which will automatically attach a URL linking to the original post, those URLs were deducted from the total number of links appeared in tweets.

The results are listed below. According to Table 1, the average URL rate of users who belong to level 0(Not Spam) and level 1(Slightly Spam) are relatively low compared with users in spam level 2 and 3. The URL Rate of level 2 is closed to level 3.

Table 1 Average URL Rate of Different Spam Level(%)

SPAM_LEVEL		N	Mean
Not Spam	URL_RATE_100	242	27.42
Slightly Spam	URL_RATE_100	167	66.17
Likely Spam	URL_RATE_100	77	127.79
Spam	URL_RATE_100	30	114.17

Figure 3 displays users' distribution by URL Rate. It is seen from the figure that regular users aggregate at low URL rate level and most of slightly spam users have no URLs while some of them attach one link on average. Most of likely spam users and spammers attach one to two URLs in their posts.

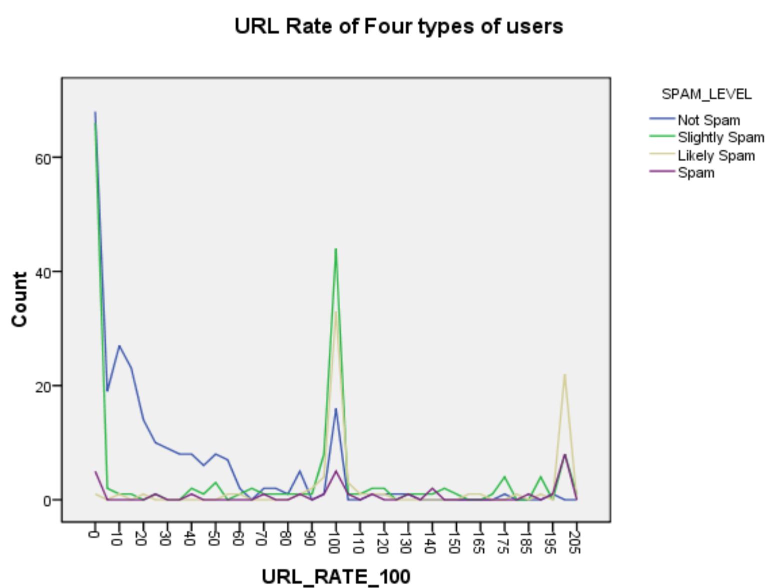

Figure 3 URL Rate of Four Different Types of Users

The graphs above indicate that spammers are more likely to attach URL in their tweets compared with regular users. The average number of URLs in their tweets are almost twice as many as normal users.

In order to see if the association in URL rate is statistically significant, a Chi-Square Test was employed between URL Rate and Spam Level(See Table 2). According results shown in the table, Pearson Chi-Square's asymptotic significance is .000, less than .05, which demonstrates that URL Rate is statistically significant associated with spam level. The higher URL Rate is, the more likely tested user is spam. This might because spammers on Twitter usually employ URLs to attract users to other websites or products in order to generate traffic or revenue.

Table 2 Chi-Square Tests: URL Rate and Spam Label

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.186E2[a]	120	.000
Likelihood Ratio	430.247	120	.000
Linear-by-Linear Association	160.290	1	.000
N of Valid Cases	516		

a. 142 cells (86.6%) have expected count less than 5. The minimum expected count is .06.

b. Mention Rate

Similarly, mention rate is the average number of mentioning contained in each user's tweets. As mention always appears with symbol "@", the metric was calculated by number of "@" in the tweet texts and the result is shown in Table 3. According to the mean value of mention rate, regular users' mention rate is close to spammers while slightly spam and likely spam users have lower mention rate compared with regular users and spammers. However, based on the median value of mention rate, regular users have the highest mention rate among all users and the remaining three categories of users' mention rate is closed to 0. It possibly results from that spammers not usually use mentioning as tactic on Twitter because Twitter does not support massive mentioning in tweets. But regular users use mention to share their thoughts with friends or followers.

Table 3 Average and Median Mention Rate of Different Spam Level

Not Spam	Mean	.4780991736
	Median	.3000000000
Slightly Spam	Mean	.1377245509
	Median	.0000000000
Likely Spam	Mean	.1922077922
	Median	.0000000000
Spam	Mean	.4716666667
	Median	.0000000000

The Chi-Square Test shows that mention rate is statistically associated with spam level. And the Goodman and Kruskal's gamma coefficient indicates that mention rate and spam level have negative correlation.

Table 4 Chi-Square Tests and Gamma Coefficent between Mention Rate and Spam Level

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.393E2[a]	129	.000
Likelihood Ratio	244.238	129	.000
Linear-by-Linear Association	10.631	1	.001
N of Valid Cases	516		

a. 165 cells (93.8%) have expected count less than 5. The minimum expected count is .06.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	-.520	.054	-10.154	.000
N of Valid Cases		516			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Hashtag Rate

Hashtag rate is the average number of hashtags contained in each user's tweets and is calculated by number of pound sign in tweets. The average hashtag rate of regular user is 16.22%, while the rest three categories are 79.58%, 74.42% and 149.67%. The average number illustrates that regular users are likely to have low hashtag rates and spammers probably use hashtags (trending topics) to attract normal users, which results in high hashtag rate.

The Chi-Square Test shows a statistically significant association between hashtag rate and spam level.

Table 5 Chi-Square Tests between Hashtag Rate and Spam Label

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	3.192E2[a]	171	.000
Likelihood Ratio	260.576	171	.000
Linear-by-Linear Association	44.452	1	.000
N of Valid Cases	516		

a. 220 cells (94.8%) have expected count less than 5. The minimum expected count is .06.

d. Word Count

Word count is the average number of words in each users' tweets. From calculation, not spam and spam users have the least number of words in their tweets, while likely spam users are more likely to write more words in their tweets due to most of likely spam users are unofficial promotion accounts.

Table 6 Average and Median Word Count of Different Spam Level

Not Spam	Mean	10.4917355
	Median	9.85000000
Slightly Spam	Mean	10.7362275
	Median	10.0000000
Likely Spam	Mean	13.1331168
	Median	14.3000000
Spam	Mean	9.5983333
	Median	9.7750000

The Chi-Square test shows a significant association between word count and spam level and the Gamma coefficient value is 0.133, displaying a positive correlation between two factors.

Table 7 Chi-Square Tests between Word Count and Spam Label

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8.762E2[a]	723	.000
Likelihood Ratio	782.949	723	.060
Linear-by-Linear Association	5.371	1	.020
N of Valid Cases	516		

a. 965 cells (99.7%) have expected count less than 5. The minimum expected count is .06.

e. Spam Word Rate

Spam word rate measures the ratio of number of spam words in each tweet and the tweet length. Based on (Stop Spammers with a Custom Comment Blacklist, 2016), (wordpress-blacklist-words, 2016) and (The Ultimate List of Email SPAM Trigger Words, 2016), I created a list of words that are likely to be used in spam messages on Twitter (See Appendix). The list of spam words contains 423 words and phrases, most of which are promotional words or words involved with inappropriate contents. Then I calculated spam word numbers in each tweet by tweet length in the light of this list.

The results are shown in Table 8. The average spam word rate of not spam and slightly spam users are 0.92% and 0.89%. In contrast, likely spam users and spam users' spam word rate is 1.56% and 5.43%, which are much higher than spam and slightly spam users.

Table 8 Average Spam Word Rate of Different Spam Level(%)

SPAM_LEVEL		N	Mean
Not Spam	SPAMWORD_RATE_100	242	.9275120073
Slightly Spam	SPAMWORD_RATE_100	167	.8856839462
Likely Spam	SPAMWORD_RATE_100	77	1.5567443210
Spam	SPAMWORD_RATE_100	30	5.4322253696

According to Chi-Square Test, the asymptotic significance is .000, less than .05. Therefore spam word rate and spam level have statistically significant association.

Table 9 Chi-Square Test between Spam Word Rate and Spam Level

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8.418E2[a]	678	.000
Likelihood Ratio	621.573	678	.940
Linear-by-Linear Association	46.270	1	.000
N of Valid Cases	516		

a. 904 cells (99.6%) have expected count less than 5. The minimum expected count is .06.

Based on all findings listed above, five content-based features: URL Rate, Mention Rate, Hashtag Rate, Word Count and Spam Word Rate could significantly differentiate users from different spam level so that those five features could be used in spam detection process.

B. Graph-based Features

a. Number of Followings

Number of Followings stands for the number of accounts that testing user follows. The data could be directly extracted from Twitter API. According previous studies, some spammers employ the strategy to follow other users in order to spread spam messages, therefore number of followings is proposed to be a feature to detect spam. However, based on results of sample data, regular users have 871.21 followings on average and likely spam users have 799.91 followings while slightly spam and spam users have more followings on average: 1727.1 and 1954.7. The abnormal results might be caused by some outliers so I also calculated the median of each level. Slightly spam and likely spam's spam is less than not spam and spam users and spammers have the highest median of followings.

Table 10 Average and Median Following of Different Spam Level

Not Spam	Mean	871.21
	Median	232.00
Slightly Spam	Mean	1727.10
	Median	35.00
Likely Spam	Mean	799.91
	Median	59.50
Spam	Mean	1954.70
	Median	458.50

The Chi-Square Test displays a statistically significant association between followings and spam level. The gamma coefficient value is -0.192, indicating that number of followings is negatively correlated with spam level. But this outcome is likely to result from the first three levels since spammers have the highest number of followings measured with both median and mean value.

Table 11 Chi-Square Test between Followings and Spam Level

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.100E3[a]	981	.005
Likelihood Ratio	828.230	981	1.000
Linear-by-Linear Association	.431	1	.512
N of Valid Cases	501		

a. 1307 cells (99.6%) have expected count less than 5. The minimum expected count is .04.

b. Number of Followers

Number of followers is the number that users follow the testing user's account, which could be extracted from Twitter datasets directly. Based on the Chi-Square Test, number of followers is independent with spam level.

Table 12 The Chi-Square Test between Follower and Spam Level

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.069E3[a]	1014	.113
Likelihood Ratio	850.701	1014	1.000
Linear-by-Linear Association	.604	1	.437
N of Valid Cases	501		

a. 1352 cells (99.7%) have expected count less than 5. The minimum expected count is .04.

c. Reputation

Reputation is a metric generated from the number of followings and number of followers. It is defined as follows:

$$Reputation = \frac{1 + Number\ of\ Followers}{1 + Number\ of\ Followings + Number\ of\ Followers}$$

, which is the ratio of followers by total number of followings and followers. As some users have no followings and no followers, therefore the numerator and denominator all plus 1.

The average reputation of not spam users and slightly spam users are 0.54 and 0.58, higher than likely spam users' reputation: 0.497. However, spammers have gotten the highest reputation score: 0.604. The Chi-Square Test also demonstrates that there is no statistically significant association between Reputation and Spam Level. This result might due to that some spammers have large number of followers and do not need to attract additional followers in order to attract users, like some accounts spread links of porn movies.

Table 13 The Chi-Square Test between Reputation and Spam Level

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.274E3[a]	1239	.237
Likelihood Ratio	966.824	1239	1.000
Linear-by-Linear Association	.177	1	.674
N of Valid Cases	501		

a. 1654 cells (99.9%) have expected count less than 5. The minimum expected count is .04.

Unlike content-based features, graph-based features do not have significant association with spammers' behaviors. It is likely that graph-based features of spammers on Twitter do not follow the traditional patterns of spammers, or they have employed strategies to alter their following/follower structure.

C. Platform

In the sample datasets, user have used 219 kinds of platforms to post their tweets. Specifically, 44% of tweets in sample sets are posted from Twitter's web or mobile clients. 13% of tweets are from Certified Third-party Application, like Facebook, Google, Instagram or Yelp etc. And the rest 43% are from other third-party applications or websites.

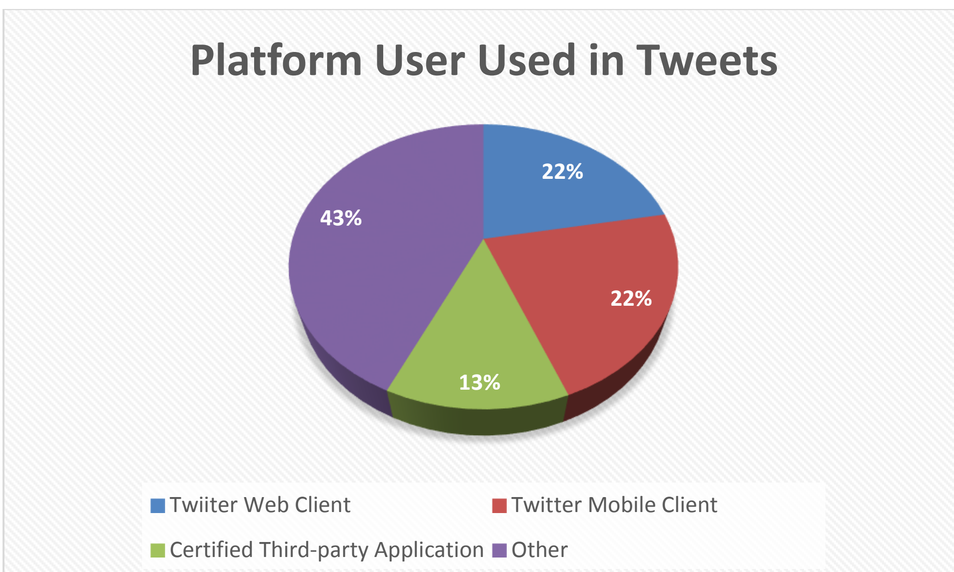

Figure 4 Platform User Used in Tweets

There is no significant association between the platform user used and user's spam level. But based from the sample sets, promotion accounts tend to use third-party applications, usually sharing from other websites or mobile apps.

EXPERIMENT

In the experiment section, I used content-based features and graph-based features combined with machine learning algorithms: J48 classification, NaïveBayes and SVM provided by weka. The classification process employed 10-folds cross-validation to reduce overfitting effect.

A. Content-based Features

The weighted average classification results based on content-based features are listed below,

Table 14 Predicting Result of Cotent-based Features

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
J48	0.711	0.142	0.699	0.711	0.704	0.792
NaïveBayes	0.595	0.244	0.567	0.595	0.567	0.775
SVM	0.585	0.273	0.561	0.585	0.542	0.705

From Table 14, J48 Classification algorithm has a precision of 0.699, a recall of 0.711 and the F-measure reaches 0.704. The precision, recall and F-Measure of NaïveBayes and SVM are lower than J48.

In order to know the reliability of the results, I used weka's Experimenter to compare different classifiers with Paired T-Tester. As Figure 5 suggests, NaïveBayes(58.90%) and SVM(58.76%) are significantly worse than J48(71.70%) at the 5% level of statistical significance. Therefore J48 outperforms the other two algorithms with content-based feature datasets.

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       3/24/16 6:51 PM


Dataset                   (1) trees.J4 | (2) bayes (3) funct
------------------------------------------------------------
'user_behavior_feature_su(100)   71.70 |   58.90 *   58.76 *
------------------------------------------------------------
                               (v/ /*) |   (0/0/1)   (0/0/1)
```

Figure 5 Classifier Comparison Results of Content-based Datasets

Looking into the details of prediction results of J48 Classification algorithm(Figure 6), the performance of predicting level 0(not spam user) and level 1(slightly spam user) is better than detecting likely spam and spam users in level 2 and 3. The former F-measure is 0.842 and 0.663, and the performance of detecting spammers are 0.553 and 0.204.

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.872     0.175     0.815     0.872     0.842      0.856    0
                 0.641     0.14      0.686     0.641     0.663      0.746    1
                 0.571     0.087     0.537     0.571     0.553      0.765    2
                 0.167     0.029     0.263     0.167     0.204      0.601    3
Weighted Avg.    0.711     0.142     0.699     0.711     0.704      0.792
```

Figure 6 Predicting Results of J48 Classification

In addition, NaïveBayes and SVM outperforms J48 in detecting level 3 users. NaïveBayes’s precision is 0.455 and recall is 0.333, resulting in a 0.385 F-Measure.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.868	0.383	0.667	0.868	0.754	0.83	0
	0.287	0.155	0.471	0.287	0.357	0.657	1
	0.506	0.087	0.506	0.506	0.506	0.855	2
	0.333	0.025	0.455	0.333	0.385	0.782	3
Weighted Avg.	0.595	0.244	0.567	0.595	0.567	0.775	

Figure 7 Predicting Results of NaiveBayes

SVM, on the other hand, does not perform well in detecting all spammers, but have a high precision: 0.75, which indicates that SVM is relatively accurate in detecting spammers.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.913	0.401	0.668	0.913	0.771	0.779	0
	0.377	0.241	0.429	0.377	0.401	0.567	1
	0.195	0.043	0.441	0.195	0.27	0.805	2
	0.1	0.002	0.75	0.1	0.176	0.613	3
Weighted Avg.	0.585	0.273	0.561	0.585	0.542	0.705	

Figure 8 Predicting Results of SVM

B. Graph-based Features

For three graph-based features, two features showed no association with spam level of users. Due to the lack of features, SVM would definitely have bad prediction performance. Therefore in Graph-based algorithm, only J48 and NaïveBayes will be used for experiment. As there are only three features for graph-based algorithms, therefore the experiment will be conducted with three features (Following, Follower and Reputation) and with one feature (Following).

Table 15 Prediction Results of Graph-based Features

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
J48- 3 features	0.571	0.313	0.526	0.571	0.53	0.648
NaiveBayes-3 features	0.473	0.469	0.336	0.473	0.332	0.529
J48-1 feature	0.569	0.354	0.468	0.569	0.494	0.605
NaiveBayes-1 feature	0.479	0.47	0.348	0.479	0.331	0.514

From Table 15, J48 and NaïveBayes's performance is similar when using three features or 1 feature. The best performance is J48 with 3 features, which has a 0.526 precision, 0.571 recall and 0.53 F-Measure.

According to the results of t test, either one feature or three features, J48 decision tree's results are significantly better than NaïveBayes' result at the 5% level of statistical significance.

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   2
Resultsets: 2
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       3/26/16 3:58 PM


Dataset                     (1) trees.J4 | (2) bayes
----------------------------------------------------
user_follow_spam_1f-weka.(100)   57.32 |   47.79 *
user_follow_spam_3f-weka.(100)   56.31 |   47.29 *
----------------------------------------------------
                                 (v/ /*) |   (0/0/2)
```

Figure 9 Classifier Comparison Results of Graph-based Datasets

The result of graph-based feature prediction is worse than the result of content-based features. One of the reason might be that the number of features are less than content-based features. The other reason is that it is likely that graph-based features are not accurate and sensitive to detect spammers on Twitter compared with content-based features.

The detailed predicting result of J48 classification also shows that with graph-based features, the performance to classify regular and slightly spam users are better than detecting real spammers. The performance of NaïveBayes in predicting different categories is similar to J48.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.833	0.521	0.595	0.833	0.694	0.675	0
	0.455	0.173	0.564	0.455	0.503	0.643	1
	0.145	0.045	0.367	0.145	0.208	0.585	2
	0	0.004	0	0	0	0.59	3
Weighted Avg.	0.571	0.313	0.526	0.571	0.53	0.648	

Figure 10 Predicting Results of J48 Classification with Graph-based Features

The experiment results show that with sample datasets, algorithms based on content-based features outperform algorithms based on graph-based features. One reason is that the number of content-based features are more than the number of graph-based features so that content based classification has more information to use. The other reason is that the graph-based features used in the experiments might not accurately indicate spammers. Number of followers and reputation features are not associated with spam level. And there is no patterns that could be found in spammers' relationship structures. Some spammers have high following and high followers, while some spammers do not follow other users but have a great amount of followers. Therefore following, follower and reputation those graph-based features may not perform well in spam prediction experiments.

In addition, J48 classification algorithm performs better than NaïveBayes and SVM with both content-based and graph-based features. But when detecting if users belong to spam level 2 and 3 with content-based features, NaïveBayes and SVM have better performance, SVM's precision is relatively high in particular.

With either content-based features or graph-based features, three classifiers all have better performance in classifying regular and slightly spam users. The reason might be that regular users usually have constant patterns in their information behavior, while spammers employ different strategies to spread spam messages, which is difficult to summarize and used for detection. Therefore, ruling out normal users repeatedly from datasets is likely to be an effective way to target spammers existing in social networking services.

CONCLUSION

This work employs experiment method, using datasets extracted from Twitter to compare two different kinds of features on how they differentiate normal users and spammers as well as how well they could perform to detect spammers.

The results show that content features URL rate, mention rate, hashtag rate, word count and spam word rate have statistically significant association with users' level of spam. And those content-based features combined with J48 classifier perform best in detecting spammers, which achieves a 0.699 in precision, a 0.711 in recall is 0.711 and a 0.704 in F-measure.

On the other hand, among graph-based features, only number of followings is significantly associated with users' level of spam. Number of followers and reputation of user are independent with users' level of spam. Algorithms based on graph features' performance are not as good as content-based features.

And finally, all algorithms combined with either content-based or graph-based features perform well in classifying normal users.

LIMITATIONS AND FUTURE WORK

This study has several limitations that could be improved in future work:

(1) Did not extract enough graph-based features for analysis.

In this study, I only extracted the number of followings and followers of each sampled users. Whereas part of the features did not work well indicating spammers, which affected the performance of graph-based algorithms. In the future work, some other features could be included in as well, like the reply, retweet or like features which showing interaction between users.

(2) Sample size is limited.

Due to the hard work to manually label each user as spam or not, I only sampled around 500 users as sample for analysis. With this limited size of sample, only 30 users were categorized as level 3, the real spammers, which is difficult to summarize patterns from the small sample. Therefore in the future work, I will try to find ways to include more users in the sample as well as labelling users automatically to reduce manual work.

(3) Lack deep analysis on how each factor works in machine learning algorithm.

The study only compares the performance of algorithms based on two kinds of features while there lacks deeper analysis on how each feature or factor performs in detecting spammers, like how much each feature contribute in the precision and recall etc., which could be improved in the future.

BIBLIOGRAPHY

Ahmed, F., & Abulaish, M. (2012). An MCL-based approach for spam profile detection in online social networks. *Proc. of the 11th IEEE Int. Conference on Trust, Security and Privacy in Computing and Communications, TrustCom-2012 - 11th IEEE Int. Conference on Ubiquitous Computing and Communications, IUCC-2012*, 602–608.

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). Detecting spammers on twitter. *Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS).*

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811–824.

Chu, Z., Widjaja, I., & Wang, H. (2012). Detecting social spam campaigns on Twitter. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7341 LNCS*, 455–472.

Delany, S. J., Buckley, M., & Greene, D. (2012). SMS spam filtering: Methods and data. *Expert Systems with Applications*, *39*(10), 9899–9908.

Grier, C., Thomas, K., Paxson, V., & Zhang, M. (2010). @ spam : The Underground on 140 Characters or Less ∗ Categories and Subject Descriptors. *Proceedings of the 17th ACM Conference on Computer and Communications Security*, 27–37.

Lin, P.-C., & Huang, P.-M. (2013). A study of effective features for detecting long-surviving Twitter spam accounts. *Advanced Communication Technology (ICACT), 2013 15th International Conference on*, 841-846.

Lumezanu, C., & Feamster, N. (2012). Observing common spam in tweets and email. *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, 441–466.

McCord, M., & Chuah, M. (2011). Spam detection on twitter using traditional classifiers. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 175-186.

Miller, Z., Dickinson, B., Deitrick, W., Hu, W., & Wang, A. H. (2014). Twitter spammer detection using data stream clustering. *Information Sciences,* 64-73.

REST APIs. (2016, March 15). Retrieved from Twitter: https://dev.twitter.com/rest/public

S. Abu-Nimeh, T. Chen, and O. Alzubi, "Malicious and spam posts in online social networks." *IEEE Computer Society,* 44(9), Sep. 2011

Song, J., Lee, S., & Kim, J. (2011). Spam filtering in twitter using sender-receiver relationship. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 301–317). 6961 LNCS.

Spamming. (2016, March 14). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Spamming

Stop Spammers with a Custom Comment Blacklist. (2016, March 7). Retrieved from Digging Into WordPress: https://digwp.com/2010/02/stop-spammers-custom-blacklist/

Streaming APIs. (2016, March 15). Retrieved from Twitter: https://dev.twitter.com/streaming/overview

Stringhini, G., Kruegel, C., & Vigna, G. (2010). Detecting spammers on social networks. *Annual Computer Security Applications Conference(ACSAC)*, 1-9.

Tan, E., Guo, L., Chen, S., Zhang, X., & Zhao, Y. (2013). UNIK: unsupervised social network spam detection. *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*, 479–488.

The Ultimate List of Email SPAM Trigger Words. (2016, March 7). Retrieved from HubSpot: http://blog.hubspot.com/blog/tabid/6307/bid/30684/The-Ultimate-List-of-Email-SPAM-Trigger-Words.aspx

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

Thomas, K., Paxson, V., Mccoy, D., & Grier, C. (2013). Trafficking Fraudulent Accounts : The Role of the Underground Market in Twitter Spam and Abuse Trafficking Fraudulent Accounts : *USENIX Security Symposium*, 195–210.

Twitter. (2015, September 14). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Twitter

Wang, A. (2010). Detecting spam bots in online social networking sites: a machine learning approach. *Data and Applications Security and Privacy XXIV*, 335–342.

Wang, A. H. (2010). Don't follow me: Spam detection in Twitter. *Security and Cryptography (SECRYPT)* (pp. 1-10). Proceedings of the 2010 International Conference.

Wang, D., Navathe, S. B., Liu, L., Irani, D., Tamersoy, A., & Pu, C. (2013). Click traffic analysis of short URL spam on Twitter. *Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom)* (pp. 250-259). 9th International Conference Conference.

wordpress-blacklist-words. (2016, March 7). Retrieved from NotaGrouch: http://notagrouch.com/wp-content/uploads/2009/12/wordpress-blacklist-words.txt

Yang, C., Harkreader, R. C., & Gu, G. (2011). Die free or live hard? Empirical evaluation and new design for fighting evolving twitter spammers. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6961 LNCS*, 318–337.

Zhang, C. M., & Paxson, V. (2011, January). Detecting and analyzing automated activity on twitter. In *Passive and Active Measurement* (pp. 102-111). Springer Berlin Heidelberg.

APPENDICES

Appendix A: Python Scripts Used for Twitter API

(1) Streaming API

```
import tweepy
import codecs
import sys
from time import clock

#OAuth Authentication

auth=tweepy.OAuthHandler(consumer_key,consumer_secret)
auth.set_access_token(access_token,access_token_secret)
api = tweepy.API(auth)

file = open("data.txt",'ab')

print api.me().name

start=clock()
print start

class StreamListener(tweepy.StreamListener):
    def on_status(self,status):
        if(status.lang=="en"):
            try:
                userid=status.author.id
                print >> file, "%s" % (userid)

            except Exception,e:
                print >> sys.stderr, 'Encountered Exception:',e
                pass

    def on_error(self,status_code):
        print 'Error:' + repr(status_code)
        return True

    def on_timeout(self):
        print >> sys.stderr, "Timeout..."
        time.sleep(10)
        return True


public_stream=tweepy.Stream(auth=auth,listener=StreamListener())
public_stream.sample()
```

```
file.close()
pass
```

(2) REST APIs

```
import tweepy
import sys

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth)

file = open("data.txt",'a')
id_file=open("IDS.txt","r")
ids=id_file.readlines()

for id in ids:
    id=id.rstrip()
    try:
        user_timeline = api.user_timeline(id)

        for status in user_timeline:
            try:
                tweet=status.text.encode('utf-8')
                tweet=tweet.replace('\n',' ')
                user=status.author.screen_name.encode('utf-8')
                userid=status.author.id
                time=status.created_at
                source=status.source
                tweetid=status.id

                # print tweet


                print >> file, "%s|%s|%s|%s|%s|%s" % (userid, user, time,
tweetid, tweet, source)

            except Exception,e:
                print >> sys.stderr, 'Encountered Exception:',e
                pass


    except Exception,e:
            print >> sys.stderr, 'Encountered Exception:',e
            pass

id_file.close()
file.close()
pass
```

```
import tweepy
import sys
import time


auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth)

file = open("follow.txt",'ab')
id_file=open("user_list.txt","r")
ids=id_file.readlines()

for id in ids:
    id=id.rstrip()
    try:
        followed = api.friends_ids(id)
        following=api.followers_ids(id)
        count_followed=str(len(followed))
        count_following=str(len(following))
        record=id+"|"+count_followed+"|"+count_following
        print record
        file.write(record)
        file.write("\n")
        time.sleep(180  )


    except Exception,e:
            print >> sys.stderr, 'Encountered Exception:',e
            pass

id_file.close()
file.close()
pass
```

Appendix B: Spam words list

$$$	[/url]	[url=	100% free	100% Satisfied
4u	50% off	Accept Credit Cards	Access	aceteminophen
Act Now!	Ad	adderall	Additional Income	Addresses on CD
adipex	advicer	Affordable	All natural	All new
Amazing	ambien	anime	Apply now	Apply Online
As seen on	ass	augmentation	Auto email removal	Avoid bankruptcy
baccarat	baccarrat	Bargain	bdsm	Be your own boss
Being a member	Beneficiary	Best price	Beverage	Big bucks
Billing address	Billion dollars	bitch	blackjack	bllogspot
Bonus	booker	Brand new pager	breast	Bulk email
Buy direct	Buying judgments	byob	Cable converter	Call free
Call now	Calling creditors	Cannot be combined with any other offer	Can't live without	Cards accepted
carisoprodol	car-rental-e-site	car-rentals-e-site	Cash	Casino
casinos	Celebrity	Cents on the dollar	cephalaxin	Certified
chatroom	Cheap	Check	money order	cialis
citalopram	Claims	Clearance	Click	clomid
cock	Collect	Compare rates	Compete for your business	Confidentially on all orders
Congratulations	Consolidate debt and credit	Consolidate your debt	coolcoolhu	coolhu
Copy DVDs	Cost	Credit	cumshot	Cures baldness
cwas	cyclen	cyclobenzaprin e	cymbalta	dating
dating-e-site	day-trading	Deal	debt	debt-consolidation
Diagnostics	dick	Dig up dirt on friends	Direct email	Direct marketing
Discount	discreetorderin g	Do it today	Don't delete	Don't hesitate
Double your	doxycycline	Drastically reduced	dutyfree	duty-free
Earn	Easy terms	Eliminate bad credit	Eliminate debt	Email harvest
Email marketing	enhancement	ephedra	equityloans	Expect to earn
Explode your business	Extra income	facial	Fantastic deal	Fast cash
Fast Viagra delivery	femdom	fetish	finance	Financial freedom
Financially independent	Financially independent	fioricet	flowers-leading-site	For free
For instant access	For just $XXX	For Only	For you	Form

Free	freenet	fuck	Full refund	gambling
gdf	gds	Get it now	Get out of debt	Get paid
Get started now	Gift certificate	Giving away	Great offer	Guarantee
hair-loss	Have you been turned down?	Hidden assets	hidden charges	holdem
Home based	Home employment	Homebased business	homeequityloans	homefinance
hotel	hqtube	Human growth hormone	hydrocodone	If only it were that easy
Important information regarding	In accordance with laws	incest	Income	Increase sales
Increase traffic	Increase your sales	Incredible deal	Info you requested	Information you requested
Instant	Insurance	Internet market	Investment	It's effective
Join millions	jrcreations	Laser printer	leading-site	Legal
levitra	lexapro	Life Insurance	limited time	lipitor
loan	Long distance phone offer	lorazepam	Lose weight	Lower interest rate
Lower monthly payment	Lower your mortgage rate	Lowest insurance rates	Lowest price	lunestra
Luxury car	macinstruct	Mail in order form	Make $	Make money
male	Marketing	Mass email	Medicine	Meet singles
Member	meridia	Message contains	Million dollars	Money back
Money making	Month trial offer	More Internet Traffic	mortgage	Multi level marketing
naked	Name brand	New customers only	New domain extensions	No age restrictions
No catch	No claim forms	No cost	No credit check	No disappointment
No experience	No fees	No gimmick	No hidden Costs	No inventory
No investment	No medical exams	No middleman	No obligation	No purchase necessary
No questions asked	No selling	No strings attached	No-obligation	Not intended
Notspam	Now only	nude	Obligation	Off shore
Offer	Once in lifetime	One hundred percent free	One hundred percent guaranteed	One time
One time mailing	Online biz opportunity	Online degree	Online marketing	Online pharmacy
online-gambling	Only $	Opportunity	Opt in	Order now
Order status	Order today	Orders shipped by	ottawavalleyag	Outstanding values
ownsthis	oxycodone	oxycontin	palm-texas-holdem-game	paxil
payday	penis	Pennies a day	Per day	Per week
percocet	Performance	pharmacy	phentermine	pills

Please read	poker	porn	Potential earnings	poze
Pre-approved	Price	Priority mail	Prize	Produced and sent out
Profits	Promise you	propecia	Pure profit	pussy
Quote	Real thing	Refinance	Removal instructions	Removes wrinkles
rental	rental-car-e-site	Requires initial investment	Reserves the right	Reverses aging
ringtone	Risk free	Rolex	roulette	Sale
Satisfaction guaranteed	Save $	Save big money	Save up to	Score with babes
Search engine listings	Search engines	Sent in compliance	Serious cash	sex
shemale	shit	shoes	shopper	Shopping spree
slot-machine	Social security number	soma	Special promotion	Stainless steel
Stock alert	Stock disclaimer statement	Stock pick	Stop snoring	Stuff on sale
Subject to credit	Subscribe	Supplies are limited	Take action now	Terms and conditions
texas holdem	texas-holdem	The best rates	The following form	They keep your money -- no refund!
They’re just giving it away	This isn't junk	This isn't spam	thorcarlson	Time limited
tits	titties	top-e-site	top-site	trading
tramadol	Trial	trim-spa	ultram	Undisclosed recipient
University diplomas	unlimited	Unsecured credit	Unsecured debt	Unsolicited
Unsubscribe	Urgent	US dollars	Vacation	valeofglamorganconservative s
valium	valtrex	viagra	vicodin	vicoprofen
vioxx	visa	Visit our website	Warranty	We hate spam
We honor all	Web traffic	Weekend getaway	Weight loss	What are you waiting for?
While supplies last	While you sleep	Why pay more	Will not believe your eyes	Win
won	Work at home	Work from home	xanax	xenical
You have been selected	Your income	zolus	Б	д
ж	и	Ч		