ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection

Shinde Asha Ashokrao[1], Shital Y. Gaikwad[2]

M.E Student, Dept. of Computer Science and Engineering, Matoshri Pratishthan Group of Institutions Vishnupuri, Nanded. (M.S), India[1]

Asst.Prof.(B.E.M.Tech.), Dept. of Computer Science and Engineering, Matoshri Pratishthan Group of Institutions Vishnupuri, Nanded. (M.S), India[2]

ABSTRACT: Twitter being popular among masses is definitely a major source of attraction for more and more spammers. Spammers send unwanted tweets to Twitter users to promote websites or services that are harmful to normal users. In order to stop spammers, the researchers proposed mechanisms. Recent work focuses on the application of automatic learning techniques in Twitter spam detection. However, the tweets are recovered in streaming and Twitter provides the Streaming API for developers and researchers to access public tweets in real time. There is no evaluation of the performance of continuous learning methods based on machine learning. In this article, we filled the gap by conducting a performance evaluation, which were three different aspects of the data, characteristics and model. A great truth of more than 600 million Public Tweets was created using a security tool based on commercial URLs. For spam detection in real time, we also extracted light features for the tweet representation. The detection of spam was then transformed into a binary classification problem in the feature space and can be solved by conventional machine learning algorithms. We evaluated the impact of different factors on spam detection performance, which included non-spam spam, discretization of functionality, size of learning data, data sampling, data related to Time and semi supervised learning algorithms.

KEYWORDS: Hashtags; Linguistics; Opinion Mining; Spam Detection; Tweets; Stanford NLP; SVM; Naive Bayes; Tagging

I. Introduction

Online social networking sites such as Facebook, LinkedIn and Twitter enable millions of users to meet new people, stay in touch with friends, establish professional links and more. According to the report in [4], Twitter is the fastest growing social networking site among all social networking sites. Twitter provides a microblogging service to users where users can send messages across Twitter and these messages are called tweets. Each tweet is limited to 140 characters. Text and images can also be included along with HTTP links.

Twitter users have different levels of awareness regarding hidden security threats in social networking sites. For example, a previous study showed that 45% of users on a social networking site easily click on the links posted by any friend in the accounts of their friends even though they may not know that person in life [12]. Thus, spammers are attracted to the use of Twitter as a tool for sending unsolicited messages to legitimate users, publishing malicious links and hijacking trend topics. Spam has become a growing problem on Twitter as well as on other online social networking sites. A study shows that more than 3% of messages are spam on Twitter [1,2,14]. Even trend topics, which are the most tweeted-about-topics on Twitter, have been attacked by spammers. An attack by themes was reported in [3], forcing Twitter to temporarily disable subjects tending to delete offensive terms.

DOI: 10.15680/IJIRCCE.2017. 0501036

To cope with the growing threats of spammers, Twitter provides several ways for users to report spam. A user can report a spam by clicking on the link "Report as spam" in his home page on Twitter. Reports are analyzed by Twitter and reported accounts will be suspended if they are considered spam. Another known and commonly available method is to display a tweet in the following format @spam @username wherein @username specifies a spam account. However, even this service is also abused by spammers. Some Twitter apps also allow users to report possible spammers. Some methods and applications to lower Twitter spam are described in [9] which can also be used to some effective strength. Twitter also puts efforts to close suspicious accounts and filter malicious tweets. However, some legitimate Twitter users complain that their accounts have been suspended by mistake by Twitter's cleaning efforts [18]. All of these ad hoc methods largely depend upon users to identify spam. We need some tools to automatically identify spammers. In addition, we need more accurate but effective methods of detecting junk e-mail to avoid causing inconvenience to legitimate users.

In this paper, we first study the differences between tweets published by spammers and legitimate users. Our goal is to identify useful features that can be used in automatic learning systems to automatically distinguish between spam accounts and legitimate accounts. The main contributions of this document are as follows:

• We propose the use of user-based features and content-based features to facilitate the detection of spam

• We propose a hybrid model that uses Vector Machine and Naïve Bayesian Classifiers, in their ability to distinguish suspect users from normal ones.

II. RELATED WORK

A. *THE TWITTER SOCIAL NETWORK:*

Twitter is a social networking site like Facebook and MySpace, which provides a microblogging service where users can send short messages (called tweets) that appear on the pages of their friends. In addition, the Twitter user is majorly identified by his username and sometimes possibly identified by a real name. A Twitter user can start "tracking" another X user. Therefore, this user receives tweets from X on his own page. User X, which is "tracked", can follow if desired. Tweets can be grouped using hashtags that are popular words, starting with a "#" symbol. If someone likes a tweet that is sent through twitter, then the other user can "retweet" this message. This retweeted message is then displayed to all its followers on Twitter. A user may decide to protect their profile. This process allows any user who wants to track a private user to ask for permission. These Hashtags are always used to allow users to effectively search for tweets on Twitter based on topics of interest. Twitter is the fastest growing social networking site with a growth rate of 660% in 2009 [4].

B. *RELATED WORK:*

Since social networks rely heavily on the concept of trusted network, exploiting this trust can have important consequences. In 2008, experience showed that 41% of Facebook users who were contacted acknowledged a friendship request from a random person [15]. L. Bilge et al. [12] show that after an attacker enters the trusted network of a victim, the victim will likely click on any link contained in the posted messages, that she knows the aggressor in real life or not. Another interesting discovery of the researchers [17] is that phishing attempts are more likely to succeed if the attacker uses stolen information from the friends of the victims in social networks to develop their phishing emails. For example, shoppybag phishing emails have often been sent from a user's buddy list and as a result, a user is often misled by believing that these emails come from trusted friends and voluntarily provides login information from his personal email account. In [16], the authors created a popular hashtags on Twitter and observed how spammers began using it in their posts. They discuss some features that could distinguish spammers from legitimate users, the degree of node and frequency of messages. However, using simple features such as degree of nodes and frequency of messages may not be sufficient because there are young Twitter users or TV anchors that display many messages.

III LITERATURE SURVEY

A larger study of Spam has been reported in [7]. The authors [7] generate honey profiles to encourage spammers to interact with them. They create 300 profiles on popular social networking sites like Facebook, Twitter and MySpace. Their 900 honey profiles attract 4250 friends asks (mainly on Facebook) but 361 out of 397 requests from friends on Twitter were spammers. They then suggested using features such as percentage of tweets with URLs, message similarity, total sent messages, number of friends for spam detection. Their detection system based on the random forest classifier can produce a false positive rate of 2.5% and a false negative rate of 3% on their Twitter dataset.

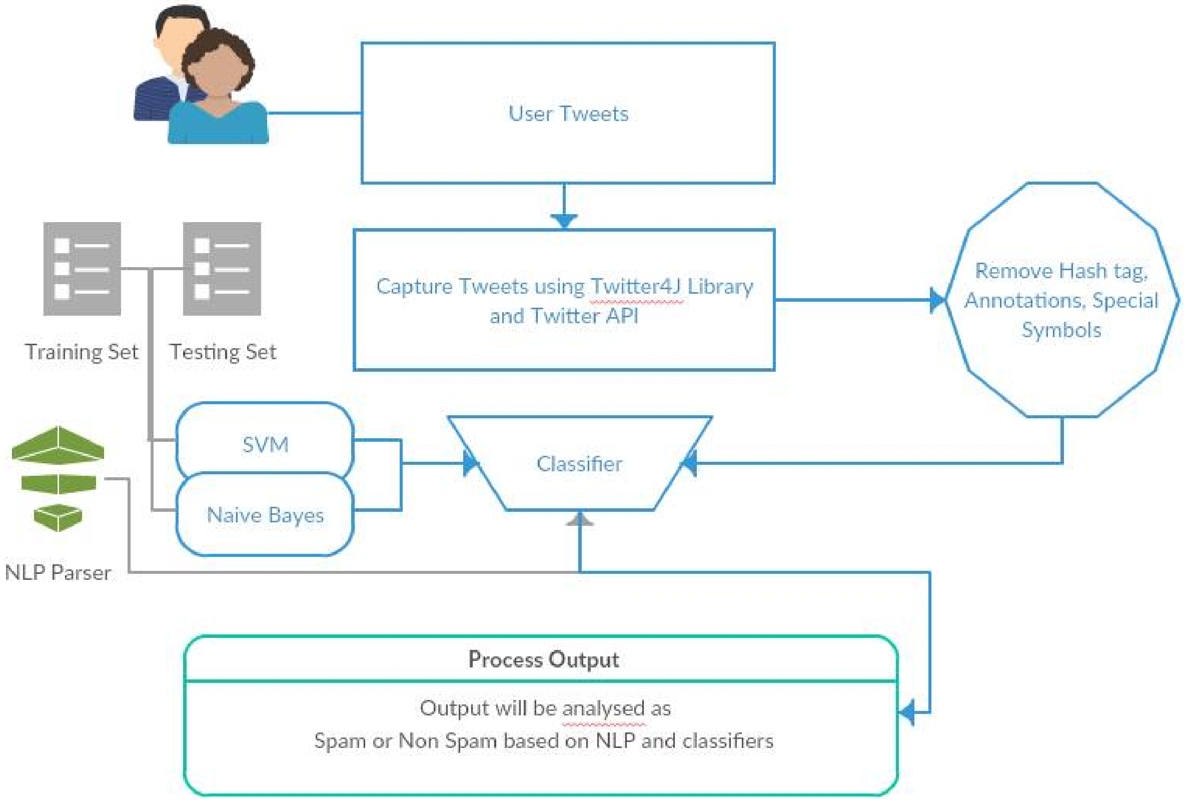
In [1], authors propose the use of graphs and content-based functionality to detect spammers. The graphic features used include the number of followers, the number of friends (the number of people you follow), and a reputation score that is defined as the ratio of the number of followers to the total number of followers The number of people a user follows. The guess is that if the number of followers is small compared to the amount of people you follow, the reputation is low and therefore the likelihood is high that the associated account is spam. The features based on the content they use include
(a) content similarity,
(b) the number of tweets containing HTTP links ,
(c) the number of tweets containing the symbols "@ "
(d) the number of tweets containing the hashtag symbol" # ".

The automatic classification of communicative constructions in short texts has become a subject widely studied in recent years. Large amounts of comments, status post and personal quotes are updated on social media web portals such as Twitter. The process of automatic labeling for the polarity associated with the text as positive or negative can reveal, aggregate or track over time how the general public thinks about certain things. Previous work describes the classification of irony [9]. The work of Tsur et al [20] collect a corpus of irony based on tweets that consist of the #irony hashtag to form classifiers on different types of features (signatures, unexpected, style and emotional scenarios) and try to distinguish # irony-tweets tweets Containing hashtags #education, #humor or #politics, reaching F1 scores of about 70. Tsur et al. [20] focus on product reviews on the World Wide Web, and try to identify sarcastic sentences from them in a semi-supervised way. Training data is collected by manually annotating sarcastic sentences and retrieving additional training data based on sentences annotated as queries.

IV. USER-BASED AND CONTENT-BASED CAPABILITIES

In this section, we discuss the features we extract from each Twitter user account for spam detection purposes. The extracted features can be categorized into (A) user-based functionality and (B) content-based functionality. User-based functionality is based on the relationships of a user, Those that the user follows (called friends) and those that follow a user (called followers) or user behaviours, Time periods and frequencies when a user tweet.

DOI: 10.15680/IJIRCCE.2017. 0501036


User Tweets
Capture Tweets using Twitter4J Library and Twitter API
Remove Hash tag, Annotations, Special Symbols
Training Set
Testing Set
SVM
Naive Bayes
Classifier
NLP Parser
Process Output
Output will be analysed as
Spam or Non Spam based on NLP and classifiers

Proposed Architecture

A. *USER-BASED FEATURES:*

When using Twitter we can build your own social tree by following friends and allowing others to follow you. Spam accounts try to track a large number of users to get their attention. The spam and abuse policy of Twitter [19] says that "if you have a small number of followers versus the amount of people you follow," it can be considered a spam account. There are three features that are user based namely the number of friends, the number of followers and the reputation of a user are calculated for the detection of spam in [15]. The reputation of a user is defined in [1] as

$$R(j) = n_i(j)/n_i(j) + n_o(j)$$

Where n_i (j) represents the number of followers that the user j a and not (j) represents the number of friends ("next") the user j a. However, in our work, we only use the number of followers and the number of "next" as part of our user-based functionality.

B. *CONTENT-BASED CAPABILITIES:*

For content-based functionality, we use some obvious functionality, for example, the average length of a tweet. Additional features based on content are described in the following subsections.

1. Number of URLs:
Because Twitter only allows a message with a maximum length of 140 characters, many URLs included in the tweets are shortened URLs. Spammers often include shortened URLs in their tweets to encourage legitimate users to access them.
Twitter filters URLs related to known malicious sites. However, shortened URLs can hide source URLs and obscure malicious sites behind them. While Twitter does not check these shortcut URLs for malware, updates to any user that consist mainly of links are considered spam according to Twitter policy. In [1], authors use the percentage of tweets containing HTTP links in the 20 most recent tweets of the user. If a tweet contains the sequence of characters "http, // or www. This tweet is considered to contain an HTTP link." In our work we use the number of HTTP links contained in a user's 100 most recent tweets.

2. Answers / Mentions:
A user is identified by a unique user name and can be referenced using the format @username in tweets on Twitter. Each user can send a reply message to another user using the @ username + message format where @ username is the message receiver. Each user can respond to anyone on Twitter if they are his / her friends / followers or not. It can also mention another @username anywhere in its tweet, rather than just at the beginning. Twitter automatically collects all tweets containing a user name in the @username format in its answers tab. The response and mentioning features are designed to help users follow the conversation and discover themselves on Twitter.
However, spammers often abuse this feature by including many @numbers as unsolicited responses or mentions in their tweets. If a user includes too many responses / mentions in their tweets, Twitter will consider this account as suspicious. The number of responses and mentions in a user account is measured by the number of tweets containing the @symbol in the user's 20 most recent tweets in [1]. However, we used a function that measures the total number of responses / mentions in the last 100 recent tweets for each user.

3. Key words / Weight:
Since we observe that the content of spam tweets contains similar words, we define two parameters to help identify spammers. First we have created a list of spam words that are often found in tweets of spammers and the associated probabilities of these words and a list of popular words in the legitimate tweets and associated probabilities of these words. Our two metrics defined using this information are: (a) the keyword metrics that counts the average number of spam words found in the 100 most recent tweets. For example, if we find a total of 50 spam words in the 100 most recent tweets, the keyword metric for that user will be 50/100, (b) the word weight metric which is defined as the difference between The sum of weighted probabilities Spam words and the sum of weighted probabilities of legitimate words found in a user's tweets. Suppose the word "hello" appears in a user's tweet and the weight of the word "hello" in the list of spam words is 0.2 while the weight of the word "hello" in the list of Regular words is 0.1. This word "hello" will be 0.2-0.1 = 0.1.

4. Hashtags:
Popular topics are the most mentioned terms on Twitter at this time, this week or this month. Users can use the hashtag, which is the # symbol followed by a term describing or naming subjects, to a tweet. If there are many tweets containing the same term, the term will become a trend. Spammers often display many unrelated tweets that contain popular topics to attract legitimate users to read their tweets. Twitter considers an account as spam "if a user publishes multiple updates that are not related to a topic using the # symbol." The number of tweets that contain the "#" symbol in the 100 most recent tweets of a user is used as one of the content-based features in [1]. However, in our work, we count the total number of hashtags in the 100 most recent tweets of each user.

C. ***SVM CLASSIFICATION :***

SVM[7] defined the input and output format. The input is a vector space and the output is 0 or 1 (Sarcastic/ Non Sarcastic). The text document in its original form is not suitable for learning. They are transformed into a format that corresponds to the input of the machine learning algorithm input. For this pre-processing on text documents is

performed. Then we do the transformation. Each word will correspond to a dimension and words identical to the same dimension.
SVM Evaluation Text categorization systems can make mistakes. To compare the different text classifiers to decide which is the best, performance measures are used. Some measure performance on a binary category, other aggregates by measurement category, to give overall performance. TP, FP, TN, FN are the number of true / false positives / negatives.

D. *NAIVE BAYES CLASSIFIER :*

A naive bayes classifier [1] is a simple probabilistic model based on the Bayes rule with a strong hypothesis of independence. The Naïve Bayes model implies a simplified conditional independence hypothesis. This is given a class (positive or negative), the words are conditionally independent of each other. This assumption does not significantly affect the accuracy of the text classification, but makes the classification algorithms very fast applicable to the problem. In our case, the probability of maximum likelihood of a word belonging to a given class is given by the expression:

$$P(x|c) = \frac{countofxintweetsofclassc}{totalnumberofwordsinclassc}$$

Here, the xi s are the individual words of the post tweet. The classifier delivers the class with the maximum a posteriori probability. We also remove duplicate words from tweets, they do not add any additional information; this type of naive bayes algorithm is called Bernoulli Naïve Bayes. The inclusion of the presence of a word instead of the count has been found to improve performance marginally, when there are a large number of training examples.

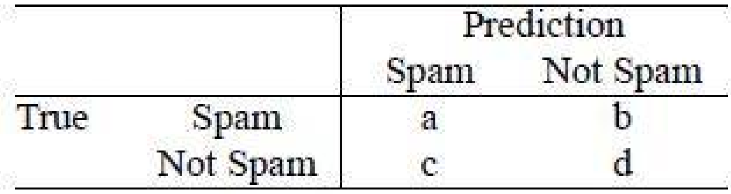
V. KEY INDEX PARAMETERS FOR RESULT CLASSIFICATION

In information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, whereas recall (also called sensitivity) is the fraction of the relevant instances which are recovered. Accuracy and recall are therefore based on an understanding and measurement of relevance.
In simple terms, high accuracy means that an algorithm returned results significantly more relevant than irrelevant, while high recall means that an algorithm returned most relevant results.

The most important category measurements for binary categories are:
Accuracy:
$$P = TP/(TP + FP)$$
Recall:
$$R = TP/(TP + FN)$$

VI. EVALUATIONS MAIN PARAMETERS OF THE INDEX

We used standard measures to measure the usefulness of our detection system that uses the chosen user and features based on content. The typical confusion matrix for our spam detection system is shown below



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

Where a is the number of spam that has been correctly ranked, b is the number of spam that has been falsely classified as non-spam, c is the number of spam messages that have been spuriously classified as spam, and d is Number of spam -spam that have been classified correctly. The following measurements are used: precision, recall and measurement F, where the precision is P = a / (a + c), the recall is R = a / (a + b) and the measurement F is defined as F = 2PR / (P + R).

VII. CONCLUSION

In social networks, traditional methods of filtering spam are not effective due to the characteristics of social networks. We propose a method for filtering spam for social networks using automatic learning and classification techniques. We use spam and non-spam as two basic classes to categorize test data and connectivity as features that are difficult to manipulate by spammers and effective in classifying spammers. In addition, our system identifies spam in real time because it does not require user history data. Service managers or customers can decide whether messages are spam or not. We hope our system contributes to quarantining a suspicious message in the anti-spam message box in social networking services. In addition, we have shown that the user relationship concept can be reflected in the user account profile to detect spam accounts. We evaluated the system using Twitter data, but the system is also effective for other social networking services because all of these services contain relationship functionality.

REFERENCES

[1] A. H. Wang, “Don’t Follow me: Twitter Spam Detection”, Proceedings of 5th International Conference on Security and Cryptography, July, 2010.
[2] Analytics, P., “Twitter study- August 2009”, http://www.peranalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf
[3] CNET (2009). 4 chan may be behind attack on twitter. http://news.cnet.com/8301-13515_3-10279618-26.html.
[4] Compete site comparison http://siteanalytics.compete.com/ facebook.com+myspace.com+twitter.com/
[5] D. Aha, D. Kibler, “Instance-based Learning Algorithms”, Machine Learning, Vol 6, pp 37-66.
[6] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.
[7] G. Stringhini, C. Kruegel, G. Vigna, “Detecting Spammers on Social Networks”, Proceedings of ACM ACSAS’10, Dec, 2010.
[8] H. Berger, M. Kohle, D. Merkl, “On the impact of document representation on classifier performance in email categorization”, Proceedings of the 4th International Conference on Information Systems Technology and IST Applications, May, 2005.
[9] How to; 5 Top methods & applications to reduce Twitter Spam http://blog.thoughtpick.com/2009/07/how-to-5-top-methods-applications-to-reduce-twitter-spam.html
[10] I. Rish. “An empirical study of the naïve bayes classifier”. Proeedings of IJCAI workshop on Empirical Methods in Artificial Intelligence, 2005.
[11] J. Platt, “Sequential Minimal Optimization: A fast algorithm for training support vector machines”, Advanced in Kernel Methods – Support Vector learning, B. Schoelkopf et al, eds, MIT Press.
[12] L. Bilge et al, “All your contacts are belong to us: automated identifty theft attacks on social networks”, Proceedings of ACM World Wide Web Conference, 2009.
[13] L. Breiman, “Random Forests”, Machine Learning, Vol 45, Issue 1, Oct, 2001.
[14] M. Mowbray, “The Twittering Machine”, Proceedings of the 6th International Conference on Web Information and Technologies, April 2010.
[15] Sophos facebook id probe, http://www.sophos.com/pressoffice/news/articles/2007/08/facebo ok.html, 2008.
[16] S. Yardi et al, “Detecting Spam in a Twitter Network”, First Monday, Vol 15(1), 2010.
[17] T.N. Jagatic et al, “Social Phishing”, Communications of ACM, Vol 50(10):94-100, 2007.
[18] Twitter (2009a), “Restoring accidentally suspended accounts.”,http://status.twitter.com/post/136164828/restoring -accidentally-suspended-accounts.
[19] Twitter (2009b). The twitter rules. http://status.twitter.com/post/136164828/restoring-accidentally-suspended-accounts.
[20] Davidov, D., Tsur, O., and Rappoport, A. 2010. SemiSupervised Recognition of Sarcastic Sentences in Twitter and Amazon, Dmitry Proceeding of Computational Natural Language Learning (ACL-CoNLL).

BIOGRAPHY

Shinde Asha Ashokrao is a second year student of master of engineering, Dept. of Computer Science and Engineering, Matoshri Pratishthan Group of Institutions Vishnupuri, Nanded.(M.S),India.

Ms Shital Y Gaikwad is a Assistant Professor, department of computer science and engineering, Matoshri Pratishthan Group of Institutions Vishnupuri, Nanded.(M.S),India. She has received Master of technology degree from SRTMU, Nanded.