Journal of Engineering Sciences
Vol 11, Issue 4 , April/2020
ISSN NO:0377-9254

# A FRAMEWORK FOR REAL TIME SPAM DETECTION IN TWITTER

S.MAHAMMADRAFI[1], Y.VYSHNAVI[2], D.SHAMEEMAFROSE[3], K.UDAY KUMAR[4], M.TEJASWINI[5], K.SREENATH[6]
[1]Assistant Professor, Dept of CSE, AITS, Rajampet, AP, India
[2,3,4,5,6] Student, Dept of CSE, AITS, Rajampet, AP, India

**ABSTRACT:**
As online social networks growing in popularity, spammers can easily access these sites by posting spam messages that draw users to malicious activities. We took Twitter platform in this work, and found spam tweets. Google SafeBrowsing and the Twitter BotMaker software detect and block spam tweets to avoid spammers. Such tools can block malicious connections, but can not protect the user as early as possible in real-time. Thus, various methods have been introduced by industries and researchers to make social network site spam safe. Some of them are based primarily on user-driven apps while some are based entirely on tweet-related features.
However, along with the user-based tools, there is no comprehensive solution which can incorporate text information from tweet. To solve this problem, we propose a system that takes the user-based and tweet-based features together with the tweet text function to identify tweets.
The benefit by using the tweet text function is that spam tweets can be detected even if the spammer creates a new account with user-based functionality and tweeting that was not just possible. With four separate machine learning algorithms-supporting vector machine, neural network, random forest and gradient boosting-we tested our solution. With Neural Network, we can achieve a 91.65 percent accuracy and have exceeded the current solution[1] by around 18 percent.

## INTRODUCTION

Throughout the past few years, online social networks such as Facebook and Twitter have become increasingly popular outlets that are an integral part of everyday life for communities. People spend plenty of time posting their messages in microblogging websites, sharing their thoughts and making friends around the world. Such websites attract a huge number of users, as well as spammers, to transmit their messages to the word because of this rising phenomenon. Twitter is rated as the most famous teen social network[2]. However, Twitter's rapid development also encourages more unsolicited activity on this website. 200 million users today produce 400 million new tweets daily[3]. This rapid expansion of the Twitter platform encourages more spammers to produce spam tweets that contain malicious links that lead a user to outside sites that contain malware downloads, phishing, drug sales, or scams[4]. Not only do these forms of attacks interfere with the user experience, they also disrupt the entire internet, which can also cause temporary disruption of internet services around the world[5]. As a result, both researchers and Twitter have come up with different spam prevention methods to make the online social network site spam-free.

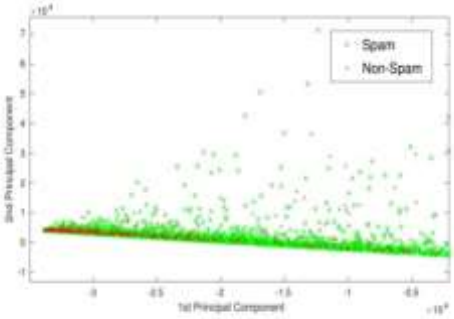


**Fig. 1:** Scatter plot of dataset showing distribution of two classes namely, spam(x) and non-spam(.)

After the introduction of BotMaker, they have seen a 40 per cent reduction in vital spam metrics. Yet one of BotMaker's poor points is that it fails to shield a target from new spam, i.e. it isn't an efficient tool for detecting spam tweets in real time. K. Thomas[7] found that 90 per cent of users could visit a new spam connection before the blacklist blocks it. TingminWu[8] performed identification of spam tweets based on profound research. They used word vector to train their model, but to fix the issue, they did not explore user or tweet-based functionality. At the other hand, Chao Chen[1] used lightweight features (specific app for the user and post) that are ideal for spam post detection in real time. Since Twitter has raised the character cap to 280 characters, scrutinizing the text of the tweet along with the user-specific features is important. Given many current solutions, there are very few robust solutions which can be used to block real-time spam tweets. In this paper, we provide a framework based on different approaches to machine learning that addresses various issues including accuracy shortage, time lag (BotMaker) and high processing time to handle thousands of

tweets in 1 sec. Firstly, we received 400,000 tweets from the dataset of HSpam14[9]. The 150,000 spam tweets and 250,000 non-spam tweets are then further described. We also extracted some lightweight features from the Bag-of-Words model along with the Top-30 words that provide the highest gain of knowledge. This strategy was discussed in detail in section III. The technique is capable of detecting spam in real time.

## MOTIVATION

Spam on Twitter is distinct from spam on other online social networks largely because Twitter reveals developer APIs to promote user interaction. Despite of this limitation spammers learn about Twitter's anti-spam program via the APIs almost all. So we need a reliable framework capable of minimizing the problems of spam detection on Twitter. The next task in real- Twitter spam detection is to pick lightweight apps that should be feasible in a very short time to process a large number of tweets and identify the spam tweets as soon as possible. Since the longer a spam tweet stays in the network, the clearer it is that it impacts users. Chao Chen[10] suggested a novel Lfun algorithm using twelve features to fix a spam-problem on Twitter. In Figure. 1 We present a graphical representation of the Chao Chen dataset[1]. As shown in Fig. 1 The distribution of the two classes, namely spam and non-spam, has overlapped greatly, making it difficult to divide the dataset into two classes. In fact, after Twitter has increased the character limit to 280, we will find the text of tweet as one of the apps.
To overcome these challenges, we integrate information gain from the Bag-of-Words model in Twitter platform along with user-based apps. To summarize, the following are our contributions:

- We collect real-world tweets provided in the HSpam14 dataset from tweet ids. We then remove from 150,000 spam tweets and 250,000 tweets user specific apps.
- We gather about 100,000 unique words from over 400,000 tweets 'text, out of which we classify 30 words that may be good indicators for labeling a tweet as spam or non-spam.
- Using specific machine learning algorithms we train our model on this structured dataset.

## PROPOSED WORK

We are planning our dataset by collecting tweets from HSpam14 that correspond to 400,000 tweet ids[9]. Then we generated the features set out on our dataset in Table I. To get details from the text of tweets, we would like to extract certain terms that can be powerful indicators to identify tweets into one of the classes: spam or non-spam.

### A. Information Gain from Bag-of-Word Model

After characterizing the text of the spam and non-tweets into two different texts, the following sets are constructed:

*US* = Set of single words in the text of Spam message.

*UNS* =Set of single words in the text of Non-spam messages.

The following probability values are determined for each term T in US and UNS:

$$P(T|US) = \frac{\#\text{ of Spam tweets that contain } T}{\text{total \# of Spam tweets}} \quad (1)$$

$$P(T|UNS) = \frac{\#\text{ of Non-Spam tweets that contain } T}{\text{total \# of Non-Spam tweets}} \quad (2)$$

For each term T, we measure the information gain ÿT as follows:

$$Y_T = P(T|U_S) \cdot P(T|U_S) \times \log_{10} P(T|U_S). \quad (3)$$

We sort terms in decreasing order based on the measurement in Equ of their ÿT score. 3. Using the above equation we take the top 15 terms from each of the US and UNS. Table II shows sample top-10 terms. These words are combined to form top-30 words we use in our feature collection. The advantage of using these terms in the feature-set based on their entropy score is that we have been able to minimize uncertainty in the outcome of the prediction because these terms have a different effect on spam and non-spam tweet frequency counting. That's why we plan to find these top 30 terms to help us identify the tweets for each class accurately.

### B. Extracting Lightweight Features

We collected some 350,000 English tweets after collecting 400,000 labelled tweets. Because we receive an arbitrary individual tweet from Twitter API, we have not been able to access the full social graph of users of Twitter. Consequently, we take the feature set from the work of Chao Chen[1] that is best suited for the timely detection of Twitter

spam. We introduce yet another feature, i.e., no of non-ASCII on top of those 12 features. From our study we find that 88% of spam tweets use non-ASCII values to post a text tweet. Table I lists the 13 features derived from the dataset.

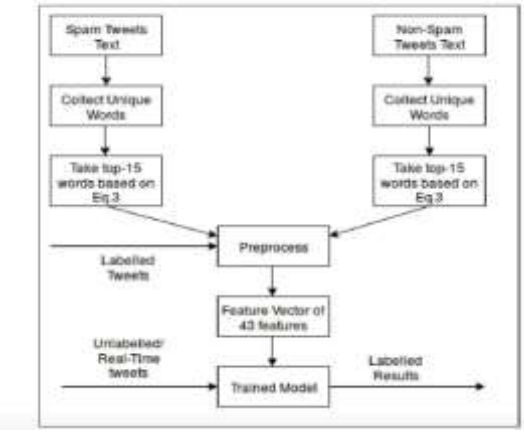


**Fig. 2:** Flow Diagram to preprocess the dataset for Information gain

| Feature Name | Description |
|---|---|
| account age | The age (days) of an account since its creation until the time of sending the most recent tweet |
| no_follower | The number of followers of this Twitter user |
| no_following | The number of followings/friends of this Twitter user |
| no_userfavourites | The number of favourites this Twitter user received |
| no_lists | The number of lists this Twitter user added |
| no_tweets | The number of tweets this Twitter user sent |
| no_retweets | The number of retweets this tweet |
| no_hashtag | The number of hashtags included in this tweet |
| no_usermention | The number of user mentions included in this tweet |
| no_urls | The number of URLs included in this tweet |
| no_char | The number of characters in this tweet |
| no_digits | The number of digits in this tweet |
| no_non-ASCII_characters | The number of non-ASCII characters in this tweet |

Table1:feature set

## C. Scaling of Dataset

We classify our feature sets as shown in Table IV into 3 groups. We investigate that the values of features are not within the same range that will impact our model training in Section IV.
So we are scale our data as follows for Featureset-1:

| Top 5 Words from Spam Tweets | Top 5 Words from Non-Spam Tweets |
|---|---|
| harvested | rain |
| tribez | asleep |
| coins | rather |
| collected | college |
| unfollower | fell |
| openfollow | follback |
| inspi | dinos |
| build | bullshit |
| smurf | child |
| brainy | coach |

*D*1 = Matix representation of Dataset-1 of size of $M*N$,
where*M* = numbers of tweets, *N* = … of features. In our case m = 350,000 and n = 43.
*Dji= jth*feature of *ith*tweet.

We normalize the data in order to represent the data using Feature-set-1, such that each function has zero mean and standard deviation of the variable.

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\bar{\sigma}_j} \quad (4)$$

Where xij is the jth feature value in the ith tweet,
μj is the standardized feature value for the jth feature value in the ith tweet μi is the mean value for the jth feature across all tweets.
Using the feature-set-2 we represent each feature using its representative Bag-of-Word.
We store this representation using libsvm format. Here every attribute is a word and the corresponding meaning is the word frequency in the tweet. Using *l*2-norm., we normalize any tweet-
.

## IV. EXPERIMENTAL SETUP AND RESULTS

In this segment, we will use four machine learning algorithms, support vector machine with kernel, neural network, gradient boosting and random forest to calculate the spam detection output on our dataset. We also equate our findings to the spam detection technique used by Chao Chen on their dataset[1]. For our experiment, we also patterned three distinct feature sets. The dataset is presented in Table IV. We use Recall, Precision, F-measurement and Accuracy to calculate the effectiveness of classifiers to assess the efficiency of our generated classification and make it comparable with current approaches. We consider the spam class to be a plus and the non-spam class to be a negative. Recall, Precision, F-measurement and Accuracy are calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP+FP+FN+TN} \quad (5)$$

| Unit % | Feature-set-1 | | Feature-set-2 | | Feature-set-3 | |
|---|---|---|---|---|---|---|
| Classifier | F-Measure | Accuracy | F-Measure | Accuracy | F-Measure | Accuracy |
| SVM with Kernel | 86.18 | 85.95 | 84.28 | 83.88 | 79.9 | 79.1 |
| Neural Network | 90.56 | 91.65 | - | - | 71.25 | 72.15 |
| Gradient Boosting | 75.81 | 85.84 | - | - | 81.26 | 82.69 |
| Random Forest | 75.39 | 86.25 | - | - | 93.6 | 92.9 |

Recall (Sensitivity) is defined as the ratio of spam properly classified in total real spam, as

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

Precision is characterized as true spam that is projected into classified spam.
It can be obtained from

$$\text{Precision} = \frac{TP}{TP+FP} \quad (7)$$

F- is the harmonic mean of Recall and Precision, which can be measured as follows:

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$
$$= 2TP/2TP+FP+FN$$

Table III demonstrates the comparison of various feature sets for different classificators. From Table III, we can conclude that Neural Network with Feature-set-1 gives the finest accuracy, i.e. 91.65 percent across all classifiers.
Our approach to use top 30 terms for features also outperformed the approach of Chen Chon[10] by 18 percent. Nevertheless, we can not use different classifiers for Feature-set-2 other than Support Vector Machine because it is impractical for other classifiers to offer input vectors having dimensions of 100,000 features. So we are just testing Featureset-2 for Support Vector Machine. Table III shows that Random Forest for Feature- is 2 percent stronger than a Dataset-1 neural network, but Feature- is more user- (e.g. account age, # of followers) so that Twitter spam can not be identified when a spammer creates a new user account. But with top-30 terms, we add user-driven functionality, then we can predict it as spam driven on the text of tweet. Thus, detecting Twitter spam as soon as possible is critical in minimizing the loss caused by spam.

Since of that property our approach makes a convincing contribution to detecting real-time Twitter spam.

## V. CONCLUSION & FUTURE WORK

Within this paper we present within Twitter a novel mechanism for real-time spam detection. We collected a huge number of 400,000 tweets from the public. Based on the text of the tweet, we extract top-30 terms which can provide the highest benefit of knowledge to identify the tweets. We also checked our approach with real- tweet detection that surpassed the current approach[1] by 18%. Spammers can change their actions over time because Twitter API is open to all users.

| Feature-Set | Sampling Method | Ratio (Spam:Non-Spam) |
|---|---|---|
| 1 | Use 43 features to train a model | 1:2 |
| 2 | Use Bag-of-Word to select features in libsvm format | 1:2 |
| 3 | Use Chao Chen's [1] dataset for comparison | 1:2 |

Table 4:sample dataset

In the real world, the function of spam tweet is continuing to change unforeseenly.
This problem is referred to as "Spam Drift." By introducing a self-learning algorithm, we will continue to update our Bag-of-Words model, based on new spam tweets. We also find in our dataset that there is a malicious connection in 79 per cent of spam tweets. So we'll also be using the URL crawl method to detect spam on Twitter. Frequent pattern mining of the text of tweets can also be a critical feature of identifying spam on Twitter in real time. We must combine those three strategies to deal with the issue of spam drift.

## VI. REFERENCES

[1] C. Chen, J. Zhang, X. Chen, Y. Xiang, and W. Zhou, "6 million spam tweets: A large ground truth for timely twitter spam detection," in 2015 IEEE International Conference on Communications (ICC), June 2015, pp. 7065–7070.
[2] A. Greig, "Twitter Overtakes Facebook as the Most Popular Social Network for Teens,According to Study, DailyMail, accessed on Aug. 1, 2015 ," http://www.dailymail.co.uk/news/article-2475591/ Twitter-overtakes-Facebook-popular-socialnetwork-teens-according-study. html, 2015, [Online].

[3] H. Tsukayama, "Twitter turns 7: Users send over 400 million tweets per day," https://tinyurl.com/ybsaq7e7, 2013, [Online].
[4] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in In Collaboration, Electronic messaging, AntiAbuse and Spam Conference (CEAS, 2010.
[5] C. Pash., "The lure of Naked Hollywood Star Photos Sent the Internet into Meltdown in New Zealand, Bus. Insider, accessed on Aug. 1, 2015 ," https://tinyurl.com/yc93ssj4, 2014, [Online].
[6] "BotMaker," https://blog.twitter.com/engineering/en us/a/2014/ fighting-spam-with-botmaker.html, [Online].
[7] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended accounts in retrospect: An analysis of twitter spam," in Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, ser. IMC '11. New York, NY, USA: ACM, 2011, pp. 243–258. [Online]. Available: http://doi.acm.org/10.1145/2068816.2068840
[8] T. Wu, S. Liu, J. Zhang, and Y. Xiang, "Twitter spam detection based on deep learning," in Proceedings of the Australasian Computer Science Week Multiconference, ser. ACSW '17. New York, NY, USA: ACM, 2017, pp. 3:1–3:8. [Online]. Available: http://doi.acm.org/10.1145/3014812.3014815
[9] "HSpam14 Dataset," http://www.ntu.edu.sg/home/axsun/datasets.html, [Online].
[10] C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou, and G. Min, "Statistical features-based real-time detection of drifted twitter spam," IEEE Transactions on Information Forensics and Security, vol. 12, no. 4, pp. 914–925, April 2017.