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

Sai Chandra Sekhar Reddy Dwarampudi (40189233)

Abstract— **In today's world, online social networking is a very large and rapidly expanding industry, yet attacks on it are more frequent; one such incident is the Twitter attack. Spammers disseminate a variety of malicious tweets that are too destructive to actual users and may take the shape of links or hash tags. Current research has focused on the use of machine learning approaches to the detection of Twitter spam. While Twitter's Streaming API allows developers and academics to view public tweets in real time, tweets are really gathered in a streaming form. Watch the labelled tweets dataset to see if the statistical characteristics of spam tweets change over time and how well the current machine learning-based classifiers perform. This issue is indicating to as Twitter Spam Drift. This problem is known as Twitter Spam Drift. There is currently no performance evaluation available for machine learning-based techniques for identifying streaming spam. By performing a performance review from three different perspectives—data, feature, and model. We bridged the gap in this article. Next, using the feature space, spam detection was reduced to a binary classification problem that could be handled using common machine learning methods. We evaluated the effects of a variety of factors on the efficiency of spam detection, including the spam to non-spam ratio, feature discretization, training data size, data sampling, time-related data, and machine learning approaches. The results show that spotting spam in live twitter streams is still a challenge and still requires a good detection technique with the consideration of the data, feature, and model.**

Index Terms— **URL Based, Streaming API, Spam Drift, Performance metrics, Domain Name System (DNS), algorithms, Twitter Spam Detection, Feature extraction.**

I. Introduction

Social networking websites like Twitter, Facebook, Instagram, and several other online social network businesses have skyrocketed in popularity in recent years. Making friends with individuals they know or are interested in occupies a significant amount of time in OSNs. One of the most well-known microblogging services is Twitter, which was established in 2006. Individuals spend a lot of time making friends with people on OSNs that they know or are interested in. The popularity of Twitter has increased since its establishment as a microblogging platform in 2006. 200 million people use Twitter, and today they produce more than 400 million new tweets per day.[2]

Characterizing Twitter Spam and Detecting Twitter Spam are the primary works. According to several Twitter regulations, Twitter will suspend accounts for acting inappropriately. Twitter will suspend any accounts that repeatedly add new friends, send identical messages, reference other users, or provide material that is merely links. The official @spam account is another place where Twitter users may report spammers. A tweet's information that was obtained via Twitter's Streaming API can be classified. We gave a fundamental assessment in this study to help readers better grasp the effectiveness of ML systems in identifying streaming spam tweets. More than 600 million tweets were included in the data, of which 6.5 million were classified as spam using Trend Micro's Web Reputation Service. Also, we extracted some simple attributes for each tweet and looked at how well various ML systems performed at identifying spam from various angles.[1]

1) Developed a significant ground truth for the study on tweet spam detection. We discussed the effects of data-related aspects on detection performance, including the spam to nonspam ratio, training data size, and data sampling.
2) To detecting spam in streaming tweets, 12 lightweight features were extracted, and it was discovered that the efficiency of the feature discretization is crucial. A recent discovery is that the characteristics of spam tweets change over time.
3) To develop the twitter spam detection model, six machine learning methods were investigated. The behaviour of these models was reported under various experiment conditions.

II. Related Works

Researchers have previously noticed Twitter's major spam issue. When several academics researched the characteristics of spam, some important works to identify Twitter spam were put out. As a result, we classify earlier relevant efforts into two groups: describing and identifying spam on Twitter.[1]

A. Characterizing Twitter Spam

Researchers examined a sample of 2060 criminal accounts and discovered that the community's inner social relationships were like those of a tiny world, with criminal hubs at the centre of the social graph more likely to follow criminal accounts. They have put out a criminal accounts inference algorithm based on social relationships that can identify undiscovered spammers by using a list of well-known spammers. 2 million URLs, or 8% of all crawling unique URLs, were discovered to be spam when 25 million URLs from 200 million public tweets were analysed. They also discovered that, with a click-through rate of 0.13% compared to email spam's far lower rate (0.0003%–0.0006%), Twitter spam was significantly more harmful than email spam.

B. Detecting Twitter Spam

Most anti-spam approaches use machine learning algorithms to distinguish between spam and non-spam. To differentiate between spammers and non-spammers, several early attempts used account and content attributes such account age, number of followers or followings, URL ratio, and tweet length. dependable characteristics that depend on the social graph to prevent feature fabrication.[1] To evaluate whether a tweet is spam or not, Song et al. retrieved the connection and distance between the sender and the recipient. Nevertheless, because to the size of the Twitter social graph, gathering these information takes a lot of time and resources. However, collecting such attributes while tweets are arriving in a stream is impossible. Together with various elements from the landing page, domain name system (DNS) information, and domain information, several URL-based features were employed, including the domain tokens, path tokens, and query parameters of the URL. Hey, developed several models for each user, such a language model and a posting time model. After the model began acting strangely, this account would be compromised and might potentially be exploited by attackers to send spam. This approach can identify if an account has been compromised or not, but it cannot identify accounts that have been mistakenly created by spammers.

A performance review of the current machine learning-based streaming spam detection algorithms is lacking, even though there are a few works that are adequate for detecting streaming spam tweets. By conducting a performance evaluation from three separate angles data, feature, and model can reduce the gap.

III. Large Dataset of streaming spam tweets

A dataset with ground truth is required to complete a variety of difficult machine learning-based streaming spam tweets detection tasks, gather live tweets, and provide the truth. We will also make this dataset accessible to other scholars for use. This section will detail our large dataset, which contains more than 600 million tweets, including more than 6.5 million spam tweets.[1]

A. Collecting Spam Data

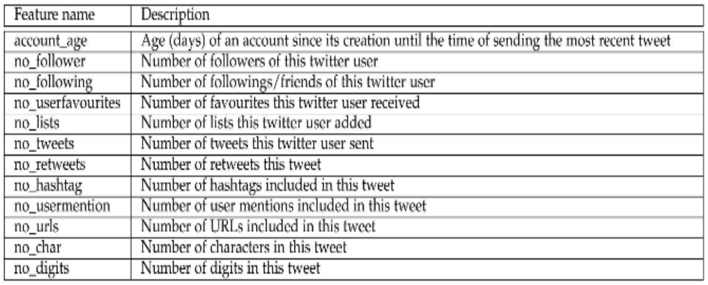
With Twitter's Streaming API, tweets with URLs may be gathered. One percent of all public tweets may be accessed in real-time using the public Streaming API, however neither direct communications nor tweets from protected accounts are accessible. JSON format is used to retrieve a tweet. The content of the tweets, the number of retweets, any included hashtags, URLs, and information about the linked Twitter user, such as the number of tweets, when the account was created, and the number of friends, are all included in the streaming API.

B. Ground Truth

Manual inspection and blacklists filtering are currently the two methods used by researchers to generate ground truth. A modest quantity of training data can be labelled manually, but it takes a lot of effort and resources. Although websites that offer human intelligence task (HIT) assistance can assist in labelling tweets, doing so is expensive and the results are not always reliable. The URLs that were harmful tweets were found using Trend Micro's Web Reputation Service. The WRS maintenance team uses a variety of cutting-edge technologies to analyse and classify URLs. If required, they will even visit the URL manually. The protection rate of define Router is 99.8%, according to recent research by a third party. 6.5 million fraudulent tweets were found in our collection of 600 million tweets accounting for 1% of all the tweets.[1]

C. Features

Firstly, the tweets are identified as spam, and then additional information is extracted from them. Since Twitter's Public Streaming API only produced random public tweets. We are concentrating on finding streaming spam tweets, thus characteristics that can be easily calculated from the tweet itself are desirable. Our dataset has 12 characteristics in all that we have extracted. The flowing table list the 12 characteristics



Feature name	Description
account_age	Age (days) of an account since its creation until the time of sending the most recent tweet
no_follower	Number of followers of this twitter user
no_following	Number of followings/friends of this twitter user
no_userfavourites	Number of favourites this twitter user received
no_lists	Number of lists this twitter user added
no_tweets	Number of tweets this twitter user sent
no_retweets	Number of retweets this tweet
no_hashtag	Number of hashtags included in this tweet
no_usermention	Number of user mentions included in this tweet
no_urls	Number of URLs included in this tweet
no_char	Number of characters in this tweet
no_digits	Number of digits in this tweet

Table. 1 . Features Extracted [1]

The 12 characteristics may be separated into two groups, user-based features, and tweet-based features, depending on the item from which they were derived. The JSON object "user" was mined for user-specific information like account age. No retweets, no hashtags, no userments, no urls, no chars, and no digits are some of the tweet-based features. The issue of "Spam Drift" affects current machine learning-based spam detection techniques because over time, the statistical characteristics of spam tweets change. Getting the modified samples to update the classification model is crucial for solving this issue. In response to the Spam Drift problem, a solution termed as Lfun was offered when it was discovered that there are such samples in the unlabeled incoming tweets that are relatively simple to gather.[2]

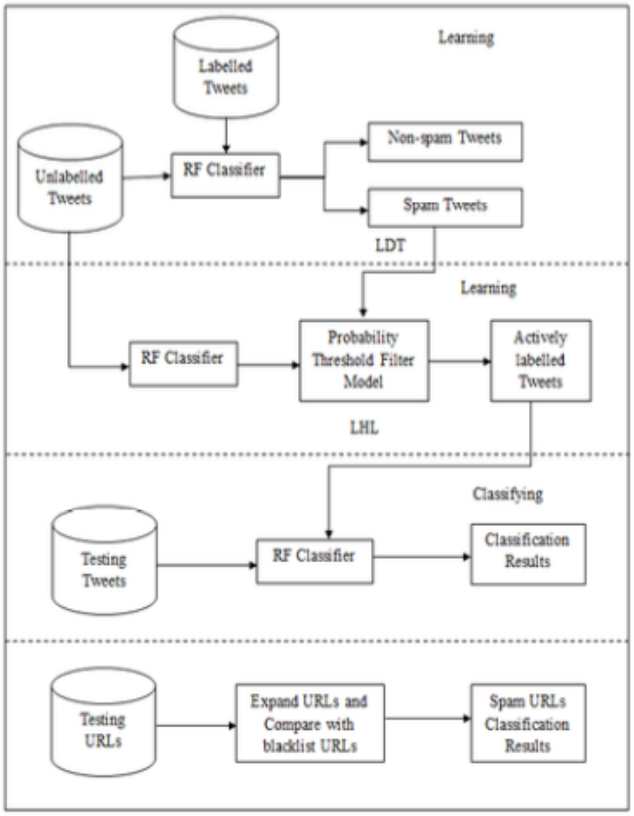
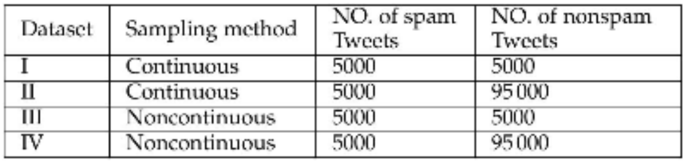


Fig. 1. Proposed Architecture [2]

This system consists of two key parts: LDT, which learns from identified spam tweets, and LHL, which learns from human labelling. To make Twitter a spam-free environment, the research community as well as Twitter itself have suggested several spam detection strategies. Twitter has implemented several Twitter rules to suspend accounts if they act inappropriately. Twitter will suspend any accounts that repeatedly add new friends, send identical messages, reference other users, or provide material that is merely links. The official @spam account is another place where Twitter users may report spammers. Researchers have used machine learning algorithms to classify spam detection as an issue to automatically identify spam. Most of these efforts determine if a person is a spammer or not by depending on characteristics that demand past user data or an existing social network. For instance, to extract data like average neighbour tweets in and distance in, a social graph must be established. Instead, features like the percentage of a user's tweets having URLs in them must be collected from the user's tweets list. Tweets, on the other hand, appear relatively quickly on Twitter since the data is in the form of a stream.[2]

D. *Feature Statistics*

Spammers participate in more lists than regular users do to be seen by more people. Naturally, spammers send more tweets than non-spammers do to distribute their messages further. Compared to spammers, non-spammers use less hashtags. Around 80% of non-spam tweets do not use hashtags in their transmitted messages, compared to just 60% of spam tweets.



Dataset	Sampling method	NO. of spam Tweets	NO. of nonspam Tweets
I	Continuous	5000	5000
II	Continuous	5000	95 000
III	Noncontinuous	5000	5000
IV	Noncontinuous	5000	95 000

Table. 2. Dataset Sampled [1]

IV. ML-BASED SPAM DETECTION PROCESS

Researchers frequently incorporate measures from information retrieval to assess the effectiveness of spam detection methods.

1) Positives and Negatives: Spam class is S, and tweet is t. Whether or not t belongs to S is the classifier's output. Using true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) is a typical method of assessing the performance of the classifier (FN). [1]

The metrics are defined as

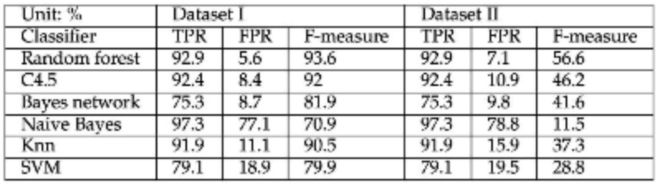
a) TP tweets accurately identified as being in class S.
b) FP tweets that should not have been classed as S-class tweets.
c) TN tweets accurately identified as not belonging to class S.
d) FN tweets of class S that were mistakenly classified as not being of class S.

We also import true positive rate (TPR) and false positive rate to assess spam detection performance (FPR).

a) TPR is defined as the proportion of spam tweets properly identified as being in the class of spam to all tweets in the class.
b) FPR is calculated as the proportion of non-spam tweets that were mistakenly assigned to the spam class S to all non-spam tweets.

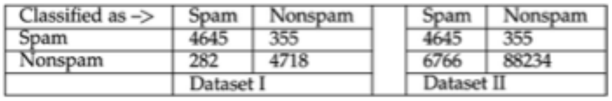
2) F-measure, Precision, and Recall : F-measure, Precision, and Recall are additional metrics used to assess performance within each class.

a) The ratio of tweets that belong to class S to those that are classified as class S is known as precision.
b) The ratio of tweets accurately identified as belonging to class S to all users in class S is known as recall.
c) F-measure, a frequently used statistic to assess per-class performance, combines accuracy and recall.



Unit: %	Dataset I			Dataset II		
Classifier	TPR	FPR	F-measure	TPR	FPR	F-measure
Random forest	92.9	5.6	93.6	92.9	7.1	56.6
C4.5	92.4	8.4	92	92.4	10.9	46.2
Bayes network	75.3	8.7	81.9	75.3	9.8	41.6
Naive Bayes	97.3	77.1	70.9	97.3	78.8	11.5
Knn	91.9	11.1	90.5	91.9	15.9	37.3
SVM	79.1	18.9	79.9	79.1	19.5	28.8

Table. 3. Performance Evaluation on Datasets I and II [1]



Classified as ->	Spam	Nonspam		Spam	Nonspam
Spam	4645	355		4645	355
Nonspam	282	4718		6766	88234
	Dataset I			Dataset II	

Table. 4. Confusion Matrix of Random Forest on Both Datasets [1]

We assess the effects of the above-mentioned machine learning techniques spam to nonspam ratio on Datasets I and II. In this series of tests, each classifier was trained using a dataset that included 1000 spam tweets and 1000 non-spam tweets. These trained classifiers were then applied to the four sampled datasets to identify spam. Apart from TPR, FPR, and F-measure, these classifiers' performance is assessed using these metrics.[1]

Apart from Bayes network and SVM, most classifiers can obtain more than 90% TPR on both datasets. For Dataset I, these classifiers can likewise achieve good F-measure. However, when analysing on Dataset II, that is, when the ratio of spam to nonspam is 1:19, the F-measures drop down considerably.

When random forest is evaluated on both datasets, F-measure lowers on Dataset II, and the confusion matrix is shown in the above table. Since the classifiers were trained using the same dataset, the change in the spam to nonspam ratio had no effect on the TP and FN of the spam class. As a result, Recall, which is defined as the ratio of tweets correctly classified as spam to all tweets that are spam, remained constant. The number of FP, however, grew significantly as more nonspam tweets were included in the test. As a result, the precision defined as the number of tweets accurately identified as spam to the total number of anticipated spam tweets declined. As a result of the sharp decline in precision, the F-measure, which combines precision and recall, substantially declined. The F-measure of machine learning-based classifiers is typically relatively low since there are far more nonspam tweets than spam tweets, according to our research.[1]

V. Conclusion

In the present world, Social Network Sites are gaining popularity at a rapid rate. Due to the increase in the users, spammers who try to hack them are also increasing. Various ML algorithms to detect these spammers are present but there lacks a framework to evaluate the performance of all the algorithms. In this report, performance metrics were calculated by using 600 million public tweets from Trend Micro and 6.5 million tweets as spam. Using this dataset, 12 light-weight features were identified which are used for machine learning-based spam classification to distinguish between spam and non-spam tweets. Four distinct datasets to replicate different circumstances were collected to examine the spam detection capabilities of various classifiers. We found that feature discretization was a critical preprocess for ML-based spam identification, after a certain number of training samples, adding more training data alone is not sufficient to improve spam detection on Twitter.

In my opinion, further modifications can be done to improve the analysis of the performance metrics on ML algorithms to detect the Spam Tweets. Firstly, when the tweets were sampled continually as opposed to randomly, classifiers could detect more spam messages. By continuous selection of tweets from different scenarios can provide with large range of dataset where large range of distribution of the features can be captured even though they change in later days. Secondly, there will be a problem with the ML detection models as the new tweets are coming in the forms of streams the dataset used for training the ML is not being updated. This issue needs to be addressed and perform the necessary modifications to the existing performance evaluation framework. Third, there is no cleaning process mentioned for the data. Data should be cleaned properly then results might have been even better, and more features can be extracted using the dataset.

References

[1] C. Chen et al., "A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection," in IEEE Transactions on Computational Social Systems, vol. 2, no. 3, pp. 65-76, Sept. 2015, doi: 10.1109/TCSS.2016.2516039.

[2] B. Kadam, R., Shital, & Y., G. (2019). Twitter Spam Detection using Lfun Approach based on Real -Time Statistical Features. INTERNATIONAL JOURNAL OF RESEARCH IN ELECTRONICS AND COMPUTER ENGINEERING (IJRECE), 7(3), 390–394.

[3] K. Lee, J. Caverlee, and S. Webb, “Uncovering social spammers: social honeypots + machine learning,” in Proc. 33rd Int. ACM SIGIR Conf. Res.Develop. Inf. Retrieval, 2010, pp. 435–442.

[4] N. Eshraqi, M. Jalali and M. H. Moattar, "Detecting spam tweets in Twitter using a data stream clustering algorithm," 2015 International Congress on Technology, Communication and Knowledge (ICTCK), Mashhad, Iran, 2015, pp. 347-351, doi: 10.1109/ICTCK.2015.7582694.

[5] K. Thomas, C. Grier, D. Song, and V. Paxson, “Suspended accounts in retrospect: An analysis of Twitter spam,” in Proc. ACM SIGCOMM Conf. Internet Meas., 2011, pp. 243–258.