A Study on Spam Detection in Twitter Based on Machine Learning

Nazia Nusrath Ul Ain[1], Meena Kumari K S[2]

[1]*Dept. of Information Science & Engineering,* [2]*Dept of Computer Science & Engineering*
Brindavan College of Engineering

***Abstract*- Spam has continued to grow at a disturbing rate despite on-going reduction efforts. This has been considerably more pervasive on micro blogging websites, given their increased popularity and ease of access. One of the most prominent micro blogging website is Twitter. Every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 500 million tweets per day. Spammers leverage on this popularity of platform to trap users in malicious activities by posting spam tweets. There are tools to stop spammers, but these tools can only block malicious links, however they cannot protect the user in real-time as early as possible. Researchers have applied different approaches to detect spam. In this paper, we study the different approaches, some of them are only based on user-based features or tweet-based features or tweet-text feature. Using tweet text feature helps us to identify spam tweets even if the spammer creates a new account which was not possible only with the user and tweet based features. The existing system which used tweet text feature evaluated four different machine learning algorithms namely – Support Vector Machine, Neural Network, Random Forest and Gradient Boosting [1]. In our proposed system, using cross validation techniques, the best performance was obtained using Naive Bayes Model. With Naïve Bayes Model, we are able to achieve accuracy surpassing the existing solution.**

***Keywords*-Naïve Bayes , Random Forest, Spam, ham**

I. INTRODUCTION

Internet and social media have become increasingly popular in the recent years. Often internet users spend lot of time on social media to follow the events of their interest, post their messages, share their ideas and make friends around the world. These platforms have become integral part of people's daily lives. One such platform is twitter which rated as the most popular social network [2].

But with great possibilities come great challenges. Exponential growth of twitter also invites unwanted activities on this platform. Every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 500 million tweets per day. Spammers leverage on this popularity of platform to trap users in malicious activities by posting spam tweets. These spam tweets not only interfere with the user experience but also can possibly cause temporary shutdown of internet services all over the world .Consequently, the research community, as well as Twitter itself, has proposed some spam detection schemes to make, Twitter as a spam-free platform. Despite, many of these existing methods, there are very few comprehensive solutions that be used for blocking spam tweets in real time.

In this paper, we study different machine learning approaches and propose a model to detect spam which takes the client and tweet based features along with the tweet text features. We have assess our solution with different machine learning algorithms and the best performance was obtained using Naive Bayes Model

II. RELATED WORK

Twitter's Growing Spam Problem has already attention from industry and researchers and was approached with many other techniques. Most of these works are utilizing machine learning algorithm to separate spam and non spam.

In 2010, Grier et al. analyzed 25 million URLs from 200 million public tweets, and found that 2 million URLs were spam, which accounts for 8% of all crawled unique URLs [3]. They further found that Twitter spam was much more harmful than email spam with a click-through rate of 0.13%, compared to a much lower rate (0.0003%–0.0006%) for email spam. Grier et al. also examined the performance of blacklists, and the results indicated that blacklists' delay failed to stop the spread of spam on Twitter.

Some preliminary works, including [4], [5], [6], [7], made use of account and content features, such as account age, number of followers or followings, URL ratio, and the length of tweet to distinguish spammers and non spammers. These features can be extracted efficiently but also fabricated easily.

Consequently, some works [8], [9] proposed robust features which rely on the social graph to avoid feature fabrication. Song et al. extracted the distance and connectivity between a tweet sender and a receiver to determine whether the tweet is spam or not [9]. However, collecting these features are very time-consuming and resource-consuming, as the Twitter social graph is extremely huge. In addition, it is unrealistic to collect those features as tweets are incoming in the form of stream.

Instead, [10] solely relied on the embedded URLs in tweets to detect spam. A number of URL-based features were used by [10], such as the domain tokens, path tokens,

and query parameters of the URL, along with some features from the landing page, domain name system (DNS) information, and domain information. These two works can only block malicious links, however they cannot protect the user in real-time as early as possible.

Several different approaches were applied to make spam free social network problem. Some of them are only based on user-based features while others are based on tweet based features only. These approaches fail to detect spam if the spammer created a new account.

Although a recent work used a tweet text feature for spam detection, there lacks of a performance evaluation of existing machine learning-based twitter spam detection methods.

In this paper, we aim to assess the Machine Learning algorithms used in the existing work and come to a conclusion with which algorithm gives better performance comparatively.

III. MACHINE LEARNING AND ITS TYPES

Machine learning [1] is a type of Artificial Intelligence where the machine learns from its code, we write the program once and when the machine encounters another problem, it should not be programmed again. It changes the code according to the new scenario's it discovers. Machine learning can be categorized into major groups as supervised, unsupervised machine learning and reinforcement learning[13]. These groups represent how the learning method works.

Supervised learning: It is a machine learning algorithm that uses a known dataset to make predictions. The dataset includes input data and response values. From it, this algorithm seeks to build a model that can make predictions of the response values for new dataset. Unsupervised learning: It is a machine learning algorithm used to draw interfaces from datasets consisting of input data without labelled responses. It finds a pattern or structure behind those inputs. Reinforcement learning: It is an area of machine learning concerned with how software agents take action in environments so as to maximize the reward.

Classification: Classification algorithm is a part of supervised learning, used to classify records. It is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y) .The output variables are often called labels or categories. The mapping function predicts the class or category for a given observation. For example, a file or a document can be classified as belonging to one of two classes: "spam" and "not spam" or "malicious" and "benign".

Random Forest: Random forest is a type of supervised machine learning algorithm based on esemble learning. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest".The random forest algorithm can be used for both regression and classification tasks.

Naïve Bayes: Naïve Bayes is a probabilistic machine learning algorithm that can be used in a wide variety of classification tasks. Typical applications include filtering spam, classifying documents, sentiment prediction etc. It is based on the works of Thomas Bayes[1702-61] and hence the name.

IV. METHODOLOGY

We prepare our dataset via accumulating tweets similar to 12,000 tweet ids from UCI dataset. So one can get facts from tweets' text, we need to extract the ones words that may be sturdy indicators to categorize the tweets in one of the lessons: spam or non-spam. We will use two algorithms –Naïve Bayes and Random Forest to compare them and then find their accuracy.

A. Information Gain from Bag-of-Word Model

After characterizing the spam and non-spam tweets' text into two separate documents, we construct the following sets:

U_S= Collection of unique words in the spam tweets' text.

NS= Collection of unique words in the non-spam tweets' text.

For each word in $_S$ and $_{NS}$,we calculate the following probability values:

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

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

We calculate the information gain $\gamma_T$for each word as follows:

$$\gamma_{\mathcal{T}} = \left| \frac{P(\mathcal{T}|\mathcal{U}_S)}{P(\mathcal{T}|\mathcal{U}_{NS})} \times \log_{10}[\frac{P(\mathcal{T}|\mathcal{U}_S)}{P(\mathcal{T}|\mathcal{U}_{NS})}] \right| \quad (3)$$

B. Extracting Light-Weight Features

After collecting 12,000 labelled tweets, we extracted around 10,000 English tweets. Since we are receiving an arbitrary independent tweet from Twitter API, so we could not obtain the complete social graph of Twitter's users.

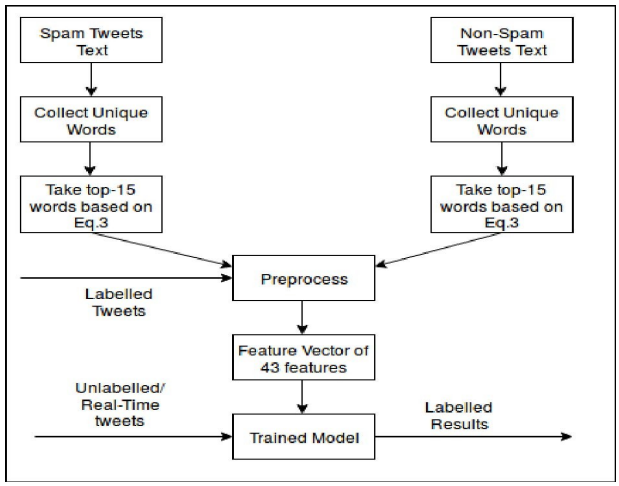

Fig. 1: Flow Diagram to preprocess the dataset for Information gain

We combine these user and tweet based 13 features along with our top-20 words as extracted in Section III-A based on tweet's text. For each top-20 word, the value in feature set corresponds to a frequency of that word in a particular tweet.

Table1. An example of a table

Top 10 Words from Spam Tweets	Top 10 Words from Non-Spam Tweets
Words	Cine
Free	Go
Entry	Jurong
Wkly	Point
Comp	Crazy
Win	Available
FA	Bugis
Cup	already
Final	Great
Tkts	World

V. EXPERIMENTAL SETUP ANDRESULTS

In this section, we will measure the Twitter spam detection performance on our dataset by using two machine learning algorithms, *Naïve Bayes* and *Random Forest(Existing algorithm used in the recent work)*.We even patterned three different feature sets for our experiment. To evaluate the performance of our created classification and make it comparable to current approaches, we consider the spam class as a positive class and non-spam class as a negative class.

We determine the Recall, Precision, F-measure and Accuracy as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

TABLE 2: Performance Evaluation on Feature-set-1 and Feature-set-2

Unit %	Feature-set 1		Feature-set 2		Feature-set 3	
Classifier	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy
Naïve Bayes	86.18	85.95	84.28	83.88	96.9	95.1
Random Forest	75.39	86.25	-	-	93.6	92.9

Table 2 shows that Naïve Bayes outperforms Random Forest in almost all the Feature sets generally and Feature set-3 specifically.

VI. FUTURE WORK

The future of Spam will grow exponentially. Some of Spammers achieve their aims by psychologically manipulating the victim into clicking unsafe links just like Malware and thus Malware detection techniques [12] could also be used to detect spam. In the real world, spam tweet's feature keeps on changing in an unanticipated way. This problem is referred as "Spam Drift". To cope up with it, we should keep updating our Bag-of-Words model based on new spam tweets by implementing self-learning algorithm.

VII. CONCLUSION

This work showed, that even quite simple Machine Learning algorithm such as Naïve Bayes Classifier may show a good result on such an important problem as spam classification. Therefore the results of this work suggest even more, that Machine learning and Artificial Intelligence techniques may be successfully used to tackle this important problem.

REFERENCES

[1]. H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*, Bengaluru, 2018, pp. 380-383.

[2]. A. Greig, "Twitter Overtakes Facebook as the Most Popular Social Network for Teens, According to Study, Daily Mail, 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]. C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@SPAM: The underground on 140 characters or less," in Proc. 17th ACM Conf. Comput. Commun. Sec., 2010, pp. 27–37.

[4]. F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammer on Twitter," presented at the 7th Annu. Collab.Electron. Messaging Anti-Abuse Spam Conf., Redmond, WA, USA, Jul. 2010.

[5]. A. H. Wang, "Don't follow me: Spam detection in Twitter," in Proc. Int. Conf. Sec. Cryptogr. (SECRYPT), 2010, pp. 1–10.

[6]. G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in Proc. 26th Annu.Comput. Sec. Appl. Conf., 2010, pp. 1–9.

[7]. 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.

[8]. C.Yang,R.Harkreader,andG.Gu,"Empirical evaluation and new design for fighting evolving Twitter spammers," IEEE Trans. Inf. Forensics Sec., vol. 8, no. 8, pp. 1280–1293, Aug. 2013.

[9]. J. Song, S. Lee, and J. Kim, "Spam filtering in Twitter using

sender receiver relationship," in Proc. 14th Int. Conf. Recent Adv. Intrusion Detect., 2011, pp. 301–317.

[10]. K. Thomas, C.Grier, J. Ma, V. Paxson, and D. Song, "Design and evaluation of a real-time URL spam filtering service," in Proc. IEEE Symp. Sec. Privacy, 2011, pp. 447–462