www.ioirp.com
International Journal of Innovative Research in Computer Science and Engineering (IJIRCSE)
ISSN: 2394-6364, Volume – 3, Issue – 3. May 2018

A Method for Detecting and Filtering Drifted Twitter Spam Using Statistical Features

S. ROJA
Department of CSE,
Idhaya Engineering College For Women,
ChinnaSalem, Villupuram(District)-606201.
rojakalai143@gmail.com

A. SHOBA
Department of CSE,
Idhaya Engineering College For Women,
ChinnaSalem, Villupuram(District)-606201.

Mr. S. JAYAPRAKASH M.E.,(Ph.D),
Associative Professor, Department of CSE,
Idhaya Engineering College For Women, ChinnaSalem,
Villupuram(District)-606201. 9791643642

Abstract-_ Twitter spam has long been a critical but difficult problem to be addressed. So far, researchers have developed a series of machine learning-based methods and blacklisting techniques to detect spamming activities on Twitter. According to our investigation, current methods and techniques have achieved the accuracy. However, due to the problems of spam drift and information fabrication, these machine-learning based methods cannot efficiently detect spam activities in real-life scenarios. Moreover, the blacklisting method cannot catch up with the variations of spamming activities as manually inspecting suspicious URLs is extremely time-consuming. In this paper, we proposed a novel technique based on deep learning techniques to address the above challenges. The syntax of each tweet will be learned through Word Vector Training Mode. We then constructed a binary classifier based on the preceding representation dataset. In experiments, we first studied the performance of different classifiers, and then compared our method to other existing text-based methods. We found that our method largely performed existing methods. We further compared our method to non-text-based detection techniques. According to the experiment results, our proposed method was more accurate.

***Index Terms*— Twitter spam, blacklist, spamming activities, classifiers, detection techniques.**

I.INTRODUCTION

Twitter is an online social networking service that enables users to send and read short 140-character messages called "tweets". Registered users can read and post tweets but those who are unregistered can only read them.

So many people's are posting photos as well as short messages to Twitter every minutes from everywhere on the earth. People send photos with short messages to twitter soon after taking photos on the spot. Therefore, by monitoring the Twitter stream and picking up Tweet photos, I get to know the current state of the world visually. This is the biggest difference of Twitter to other social media. By taking the account of these unique characteristics of Twitter, I am working on mining photos from the Twitter stream.

Social networks have become popular in recent years with millions of daily users sharing their everyday activities with friends and family. Users link themselves by defining others to follow, and consequently have their own followers based not only on social relations but also related with topics of interest. Twitter is one of the most well-known social media platforms, being characterized by providing a micro blogging service where users are able to post text-based messages of up to140 characters, mimicking the SMS (Short Message Service)messages, and known as tweets. According to the Twitter website (http://www.twitter.com) the broad cover age of this social network is confirmed by having 255 million monthly active users that post 500 million tweets per day. Another interesting characteristic of Twitter is the presence of hash tags, single words started with the symbol "#", used to classify each message content. Recently, the use of hash tags became popular, being adopted by other social networks like Face book or Instagram, as more roles were identified to the use of hash tags, like bringing a wider audience into discussion, spreading an idea, get affiliated with a community, or bringing together other Internet resources. Although mostly considered as an entertainment tool, tweets may contain information of broad interest and are being widely studied as they have a wide range of applications and uses, like event detection, academic tool, news media, or mining political opinion.

A. DEEP LEARNING TECHNIQUES

Twitter spam has long been a critical but difficult problem to be addressed. So far, researchers have developed a series of machine learning-based methods and blacklisting techniques to detect spamming activities on Twitter. According to my investigation, current methods and techniques have achieved the accuracy of around 80%. Moreover, the blacklisting method cannot catch up with the variations of spamming activities as manually inspecting suspicious URLs is extremely time-consuming. In this paper, proposed a novel technique based on deep learning techniques to address the above challenges. The syntax of each tweet will be learned through Word Vector Training Mode. In experiments, I collected and implemented a 10-day real Tweet datasets in order to evaluate my proposed method. I first studied the performance of different classifiers, and then compared in this method to other existing text-based methods. I found that method largely out performed existing methods and further compared this method to non-text-based detection techniques. According to the experiment results, my proposed method was more accurate.

B. ONLINE SOCIAL NETWORKS (OSNs)

Online social networks (OSNs) have become an important source of information for a tremendous range of applications and researches such as search engines, and summarization systems. However, the high usability and accessibility of OSNs have exposed many information quality (IQ) problems which consequently decrease the performance of the OSNs dependent applications. Social spammers are a particular kind of ill-intentioned users who degrade the quality of OSNs information through misusing all possible services provided by OSNs.

Social spammers spread many intensive posts/tweets to lure legitimate users to malicious or commercial sites containing malware downloads, phishing, and drug sales. Given the fact that Twitter is not immune towards the social spam problem, different researchers have designed various detection methods which inspect individual tweets or accounts for the existence of spam contents. However, although of the high detection rates of the account-based spam detection methods, these methods are not suitable for filtering tweets in the real-time detection because of the need for information from Twitter's servers. At tweet spam detection level, many light features have been proposed for real-time filtering; however, the existing classification models separately classify a tweet without considering the state of previous handled tweets associated with a topic. Also, these models periodically require retraining using a ground-truth data to make them up-to-date. Compared to the classical time-independent classification methods such as Random Forest, the experimental evaluation demonstrates the efficiency of increasing the quality of topics in terms of precision, recall, and F-measure performance metrics.

C. REPORTING SPAM ON TWITTER

"Spam" refers to a variety of prohibited behaviors that violate the Twitter Rules. Spam can be generally described as unsolicited, repeated actions that negatively impact other people. This includes many forms of automated account interactions and behaviors as well as attempts to mislead or deceive people. Behaviors that constitute "spamming" on Twitter will continue to evolve.

Twitter takes fighting spam seriously, and the users to enjoy the service without being concerned about spam. My anti-spam team continues to evolve and respond to new forms of spam to enable a spam-free environment on Twitter. While I have systems and tools to detect spam on Twitter, I also rely on you to help by reporting spam.

II. LITERATURE SURVEY

Spammers Are Becoming 'Smarter' On Twitter C.Chen, J. Zhang, Y. Xiang, W. Zhou, and J. Oliver Apr. 2016 stated a while researchers develop various approaches to detect Twitter spam, spammers thwart their efforts with more complex spamming strategies. The authors' in-depth analysis of more than 570 million tweets revealed three new spamming strategies: coordinated posting behavior, finite-state machine-based spam template, and passive spam.

Sifting Robotic From Organic Text: A Natural Language Approach For Detecting Automation On Twitter E. M. Clark, J. R. Williams, C. A. Jones, R. A. Galbraith, C. M. Dan forth, and P. S. Dodd's Twitter, Sep. 2016 stated a popular social media outlet has evolved into a vast source of linguistic data, rich with opinion, sentiment, and discussion. Due to the increasing popularity of Twitter, its perceived potential for exerting social influence has led to the rise of a diverse community of automatons, commonly referred to as bots. These inorganic and semi-organic Twitter entities can range from the benevolent (e.g., weather-update bots, help-wanted-alert bots) to the malevolent (e.g., spamming messages, advertisements, or radical opinions). Existing detection algorithms typically leverage metadata (time between tweets, number of followers, etc.) to identify robotic accounts.

Million Spam Tweets: A Large Ground Truth For Timely Twitter Spam Detection C. Chen, J. Zhang, X. Chen, Y. Xiang, and W. Zhou Jun. 2015 stated a Twitter has changed the way of communication and getting news for people's daily life in recent years. Meanwhile, due to the popularity of Twitter, it also becomes a main target for spamming activities. In order to stop spammers, Twitter is using Google Safe Browsing to detect and block spam links. Despite that

blacklists can block malicious URLs embedded in tweets, their lagging time hinders the ability to protect users in real-time. Thus, researchers begin to apply different machine learning algorithms to detect Twitter spam. However, there is no comprehensive evaluation on each algorithm's performance for real-time Twitter spam detection due to the lack of large ground truth. To carry out a thorough evaluation, They collected a large dataset of over 600 million public tweets. They further labeled around 6.5 million spam tweets and extracted 12 light-weight features, which can be used for online detection.

Asymmetric Self-Learning For Tackling Twitter Spam Drift C. Chen, J. Zhang, Y. Xiang, and W. Zhou Apr. 2015. Stated a Spam has become a critical problem on Twitter. In order to stop spammers, security companies apply blacklisting services to filter spam links. However, over 90% victims will visit a new malicious link before it is blocked by blacklists. To eliminate the limitation of blacklists, researchers have proposed a number of statistical features based mechanisms, and applied machine learning techniques to detect Twitter spam. In the labeled large dataset, the statistical properties of spam tweets vary over time, and thus the performance of existing ML based classifiers are poor. This phenomenon is referred as "Twitter Spam Drift". In order to tackle this problem, They carry out deep analysis of 1 million spam tweets and 1 million non-spam tweets, and propose an asymmetric self-learning (ASL) approach.

Whole Product Dynamic Real-World Protection Test, Av Comparatives Csiszar and J. Korner, Aug. 1, 2015 stated a AV-Comparatives is an independent organization offering systematic testing that checks whether security software, such as PC/Mac-based antivirus products and mobile security solutions, lives up to its promises. Using one of the largest sample collections worldwide, it creates a real-world environment for truly accurate testing. AV-Comparatives offers freely accessible results to individuals, news organizations and scientific institutions. Certification by AV-Comparatives provides an official seal of approval for software performance which is globally recognized. Currently, AV-Comparatives' Real-World Protection Test is the most comprehensive and complex test available when it comes to evaluating the real-life protection capabilities of antivirus software. Put simply, the test framework replicates the scenario of an everyday user in an everyday online environment – the typical situation that most of the experience when using a computer with an Internet connection. AV-Comparatives works closely with several academic institutions, especially the University of Innsbruck's Department of Computer Science, to provide innovative scientific testing methods. If you plan to buy an Anti-Virus, please visit the vendor's site and evaluate their software by downloading a trial version, as there are also many other features and important things for an Anti-Virus that you should evaluate by yourself. Even if quite important, the data provided in the test reports on this site are just some aspects that you should consider when buying Anti-Virus software.

Fighting Spam With Botmaker, Twitter R. Jeyaraman Aug. 1, 2015. Stated a Spam on Twitter is different from traditional spam primarily because of two aspects of their platform: Twitter exposes developer APIs to make it easy to interact with the platform and real-time content is fundamental to that user's experience. These constraints mean that spammers know (almost) everything Twitter's anti-spam systems know through the APIs, and anti-spam systems must avoid adding latency to user-visible operations. These operating conditions are a stark contrast to the constraints placed upon more traditional systems, like email, where data is private and adding latency of tens of seconds goes unnoticed.

A Survey On Concept Drift Adaptation J. Gama, I. Žliobait˙e, A. Bifet, M. Pechenizkiy, and A. Bouchachia, Apr. 2014. Stated Concept drift primarily refers to an online supervised learning scenario when the relation between the input and the target variable changes over time. Assuming a general knowledge of supervised learning in this paper they characterize adaptive learning process, categorize existing strategies for handling concept drift, overview the most representative, distinct and popular techniques and algorithms, discuss evaluation methodology of adaptive algorithms, and present a set of illustrative applications. The survey covers the different facets of concept drift in an integrated way to reflect on the existing scattered state-of-the-art. Thus, it aims at providing a comprehensive introduction to the concept drift adaptation for researchers, industry analysts and practitioners.

Spam Ain't As Diverse As It Seems Hongyu Gao, Yi Yang, Kai Bu, 2014 stated In online social networks (OSNs),spam originating from friends and acquaintances not only reduces the joy of Internet surfing but also causes damage to less security-savvy users. Prior countermeasures combat OSN spam from different angles. Due to the diversity of spam, there is hardly any existing method that can independently detect the majority or most of OSN spam. An inspiring finding is that the majority (63.0%) of the collected spam is generated with underlying templates. Therefore propose extracting templates of spam detected by existing methods and then matching messages against the templates toward accurate and fast spam detection. The implementation for this insight through an OSN spam's filtering system that performs online inspection on the

stream of user-generated messages. This is automatically divides OSN spam into segments and uses the segments to construct templates to filter future spam. Experimental results show that highly accurate and can rapidly generate templates to throttle newly emerged campaigns.

Real-Time Credibility Assessment Of Content On Twitter New York City, NY, USA: Springer, 2014. Stated that Lately, Twitter has grown to be one of the most favored ways of disseminating information to people around the globe. However, the main challenge faced by the users is how to assess the credibility of information posted through this social network in real time. In this paper, that presented a real-time content credibility assessment system named Cred Finder, which is capable of measuring the trustworthiness of information through user analysis and content analysis. The proposed system is capable of providing a credibility score for each user's tweets.

III. CONCEPTS

A. PROPOSED SYSTEM:

Consequently, the research community, as well as Twitter itself, has proposed some spam detection schemes to make Twitter as a spam-free platform. For instance, Twitter has applied some "Twitter rules" to suspend accounts if they behave abnormally. Those accounts, which are frequently requesting to be friends with others, sending duplicate content, mentioning others users, or posting URL-only content, will be suspended by Twitter. Twitter users can also report a spammer to the official @spam account. To automatically detect spam, machine learning algorithms have been applied by researchers to make spam detection as a classification problem. Most of these works classify a user is spammer or not by relying on the features which need historical information of the user or the exiting social graph. For example, the feature, "the fraction of tweets of the user containing URL" used in must be retrieved from the users' tweets list; features such as, "average neighbors' tweets" in and "distance" in cannot be extracted without the built social graph. However, Twitter data are in the form of stream, and tweets arrive at very high speed. Despite that these methods are effective in detecting Twitter spam, they are not applicable in detecting streaming spam tweets as each streaming tweet does not contain the historical information or social graph that are needed in detection.

The benefit of "old" labeled spam is to eliminate the impact of "spam drift" to classify more accurate spam tweets in future days. The effectiveness of "old" spam has been proved by my experiments during a short period. However, the effectiveness will decrease as the correlation of "very old" spam becomes less with the new spam in the long term run. In the future, I will incorporate incremental adjustment to adjust the training data, such as dropping the "too old" samples after a certain time. It can not only eliminate un useful information in the training data but also make it faster to train the model as the number of training samples decrease.

B. SYSTEM ARCHITECTURE:

The solutions used for the project have been determined and to combine them, an appropriate overview of the design need to constructed.

It is very simple, new e-mails first go through the black & white lists to do an initial filtering process. If the filter cannot identify the e-mail as spam or ham, it then goes through the Bayesian's filter. Of course at the end of the whole filtering process, all databases

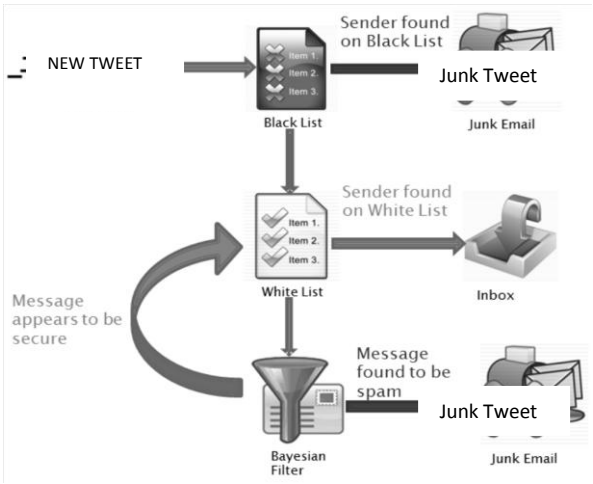

Figure 1: Architecture overview of the filter

including the black list, white list as well as the Bayesian's words database need to be updated according to the result.

C. MACHINE-LEARNING ALGORITHMS:

Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions expressed as outputs. Machine learning is closely related to and often overlaps with computational statistics; a discipline which also focuses in prediction-making through the use of computers.

Process of ML-Based Twitter Spam Detection This section describes the process of Twitter spam detection by using machine learning algorithms. Illustrates the steps involved in building a supervised classifier and detecting Twitter spam. Before classification, a classifier that contains the knowledge structure should be trained with the pre labeled tweets.

After the classification model gains the knowledge structure of the training data, it can be used to predict a new incoming tweet. The whole process consists of two steps: 1) learning and 2) classifying. Machine Learning (ML) based detection schemes involve several steps. First, statistical features, which can differentiate spam from non-spam, are extracted from tweets or Twitter users (such as account age, number of followers or friends and number of characters in a tweet). Then a small set of samples are labeled with class, *i.e.* spam or non-spam, as training data. After that, machine learning based classifiers are trained by the labeled samples, and finally the trained classifiers can be used to detect spam.

D. SCOPE:

Remember, everything on Twitter is public by default. However, I can easily make your own private experience. If I did like to make twitter a place where you privately interact with friends, just set your account to private. Turning this setting on means I will have to manually give permission to anyone who wants to follow you if you'd like them to be able to see your tweets and communicate with you. With a private account, only the people who you've given permission to follow you will see your tweets.

Most importantly, if all accounts are public and someone is acting a fool—posting mean tweets or just bugging you constantly don't hesitate to block them, mute them, or report their behavior. All of these options are available to you, so don't be afraid to use them.

E. Well-Known Spamming Strategies

At the most basic level, spammer's use various Twitter functions such as @ and hashtags (#) to engage victims (see the "Twitter Features" sidebar for a breakdown of Twitter terms). Spammers can use @ to make spam tweets appear on the victim's feed without being a follower of the victim—for example, a spam tweet will appear on Barack Obama's timeline if it is written with @obama. By embedding popular hashtag keywords, one spam tweet can become part of a trending topic that can then be viewed by a victim who is interested in that topic. For example, a spam tweet with #007 will be disseminated to victims who are browsing the popular James Bond book and film series. Spammers also use other Twitter functions, such as "reply," "favorite," and "following" to spread spam. Fortunately, researchers can also use these features (such as the number of followers or the number of hash tags) to detect Twitter spam.

IV. DISCUSSION

A. RELATED WORK

In research community, there are also some machine learning approaches related to my proposed method. For example, online learning and incremental learning. They are both common machine learning algorithms to continuously update the prediction model with new training data for better future classification. They can generate a prediction model and put it into operation without much training data at first, but they require new training data to update the model. When it comes to online Twitter spam classification, it is very difficult to label enough training samples to update the model. The reasons are two-folds. Firstly, it is significantly time-consuming to label a large amount of tweets by human. Secondly, it is difficult to gain enough spam tweets even that got a large number of human-labeled tweets, as the spam rate of Twitter is about 5%. If there are not enough spam samples (Lfun does not need non-spam samples as non-spam tweets are not drifting) to retrain the model, it is not able to solve the "spam drift" issue. In the Lfun approach has the same advantage of online learning and incremental learning, *i.e.*, it can be deployed without much training data at the beginning, but to be updated when new training data comes. Different to online and incremental learning, that will be incorporating both automated labeling and human labeling. The LDT component learns from the detected tweets. This competent is automatically updated with detected spam tweets with no human effort. To better adjust the prediction model, I also import LHL component, which learns from human labeling. To minimize human effort, LHL only samples a very small number of tweets for labeling, for example, 100 tweets in my experiments. In addition, it does not randomly pick up tweets to label, but to be in line with selection criteria called "Probability Threshold Filter Model" which can choose the most useful tweets. Benefiting from these two components, the Lfun approach can successfully deal with "spam drift", but with the least human effort.

a) SOCIAL NETWORKS

Social networks have gained significant importance and are being widely studied in many fields in the last years. Modern challenges in social networks involve not only computer science matters but also social, political, business and economical sciences. In computer science, and considering the focus on Twitter, recent works comprise event detection, information spreading, community mining, Crowd sourcing and sentiment analysis. I have proposed the use of meta-classes to boost the performance of Twitter messages classification. This preliminary study shed light on the possibility of evaluating message content in order to predict hash tags ("#"). Regarding Twitter Hash tags, and particularly hash tag recommendation, that have also identified the recent study presented in, where an approach for hash tag recommendation is introduced. This approach computes a similarity measure between tweets and uses a ranking system to recommend hash tags to new tweets. In the use of hash tags to classify Twitter messages is done by clustering similar tweets in a graph based collective classification strategy. Although

the presented results seem promising, it identified the lack of adaptiveness in this strategy. A different approach is proposed in, where an event detection method is described to cluster Twitter hash tags based on semantic similarities between the has hash tags. This work is in line with the previous work except for the fact that the semantic similarities are computed based on the message content similarities rather than being based on semantic hash tag similarities.

b) CONCEPT DRIFT

In the presence of concept drift, the learning task is not easy and requires a special approach, different from those commonly used, as the arriving instances cannot be treated as equally important contributors to the final concept. In non-stationary environments like the Twitter stream, effective requires a learning algorithm with the ability to detect context changes without being explicitly informed about them, quickly recover from the context change and adjust its hypothesis to the new context. It should also make use of previous experienced situations when old contexts and corresponding concepts reappear. According to, there are 3 approaches to handle concept drift: 1.instance selection, 2.instance weighting and 3.ensemble learning. The related work presented so far sheds light on the importance of dealing with concept drift especially in dynamic scenarios like social networks, and particularly in Twitter, where important information can be mined. Multiple applications like spam email filtering, intrusion detection, recommendation systems, event detection or improve search capabilities are just pointed examples.

In research community, there are also some machine learning approaches related to my proposed method. For example, online learning and incremental learning. There are both common machine learning algorithms to continuously update the prediction model with new training data for better future classification. That can generate a prediction model and put it into operation without much training data at first, but that require new training data to update the model. When it comes to online Twitter spam classification, it is very difficult to label enough training samples to update the model. The reasons are two-folds. Firstly, it is significantly time-consuming to label a large amount of tweets by human. Secondly, it is difficult to gain enough spam tweets even that have got a large number of human-labeled tweets, as the spam rate of Twitter is about 5%. If there are not enough spam samples (Lfun does not need non-spam samples as non-spam tweets are not drifting) to retrain the model, it is not able to solve the "spam drift" issue. The Lfun approach has the same advantage of online learning and incremental learning, *i.e.*, it can be deployed without much training data at the beginning, but to be updated when new training data comes. Different to online and incremental learning, that incorporate both automated labeling and human labeling. The LDT component learns from the detected tweets. This competent is automatically updated with detected spam tweets with no human effort. To better adjust the prediction model, it also import LHL component, which learns from human labeling. To minimize human effort, LHL only samples a very small number of tweets for labeling, for example, 100 tweets in my experiments. In addition, it does not randomly pick up tweets to label, but to be in line with selection criteria called "Probability Threshold Filter Model" which can choose the most useful tweets. Benefiting from these two components, the Lfun approach can successfully deal with "spam drift", but with the least human effort.

c) TWITTER API BEST PRACTICES: AUTOMATED SPAMMER DETECTION

Spam is a huge problem on Twitter. In certain areas it can account for the majority of tweets. This can get in the way of delivering quality results when you try collecting tweets for aggregation sites or data mining.

Because of this network phenomenon with spamming, I have found that the key to blocking spam is automating the discovery of spam accounts, and then creating a blacklist to exclude those accounts from any data collection I do for clients. Once an account is identified as a spammer, I ignore all of their tweets. I have seen reductions of spam looking tweets by as much as 50% just by blocking a hundred or so accounts. Of course, new spam accounts are being created as fast as Twitter can suspend the old ones, so it is important to add a level of automation to this process. When clients see how much cleaner their tweet data is, they understand that this blacklist is a valuable resource that can be applied to all of their future tweet collection.

d) Twitter Spam-Fighting Techniques

Here is a basic list of techniques I typically use for building a spam account blacklist. It is important to realize that this approach will reduce your flow of tweets, that is the goal after all, and you will block some non-spam tweets. If you make the following techniques tunable, you can adjust the level of spam my activity you use to blacklist someone. Then you can test these settings until you get the highest yield of good tweets while blocking as much spam as possible. In my opinion getting a tweet stream that has most of the spam blocked, while losing about 10% of the good tweets, is much better than a stream with 50% spam.

The goal of this process is to identify spam accounts, not just to block individual tweets, so you need to build a database table with all users who sent the tweets you are collecting. Then you can collect scores for each user on several blacklist criteria. I typically use a spam score field that counts the number of times they use spam words, and a duplicate count

field that records the number of times they send duplicate tweets.

I do blacklisting in two levels. First I block all tweets from accounts that look too new or have a few signatures of a spammer. For example, if the creation date is within the first month, or the default avatar, often called the "egg" on Twitter is being used. I also block tweets from users who have only a dozen or so tweets. A common spam technique is to create an account, tweet for a week or so, and then abandon it. This doesn't give me enough time to detect their activity, so I just keep their tweets out of the results I deliver. These users are reevaluated every 24 hours. Once they get past this initial block, and if their activity is not spammy, I let their tweets into the tweet stream. The second level of blacklisting is based on their tweets over a longer period, usually about 3 or 4 days.

B. SYSTEM ANLISYS:

Although there are a few works, such as and which are suitable to detect streaming spam tweets, there lacks of a performance evaluation of existing machine learning-based streaming spam detection methods. In this paper, my aim to bridge the gap by carrying out a performance evaluation, which was from three different aspects of data, feature, and model. Others apply existing blacklisting service, such as Google Safe Browsing to label spam tweets. Nevertheless, these services API limits make it impossible to label a large amount of tweets. However, Twitter has around 5% spam tweets of all existing tweets in the real world.

In a supervised spam detection system, a learning algorithm, such as Random Forest, must be trained by sufficient labeled data to obtain more accurate detection results. However, labeled instances are very expensive and time-consuming to obtain. Fortunately, that have a huge number of unlabelled Tweets which can be easily collected. The LHL in the Lfun is best suited where there are numerous unlabelled data instances, and human annotator anticipating to label many of them to train an accurate system. LHL aims to minimize the labeling cost by using different learning criteria to select most informative samples from unlabelled data to be labeled by a human annotator.

a) Build a Twitter RSS Feed

Social networks are perfect for sharing ideas with your followers and keeping up with friends and brands around the world. They're terrible at helping you stay focused at work, however. Designed to keep you around as long as possible, it requires incredible willpower to just check *one* update without peeking at notifications and baby pictures and cat GIFs.

News can be distracting too, but RSS feeds help balance it out by letting you subscribe to only your favorite sites. You might want to follow your favorite brands' Twitter feeds via RSS—or you might gain more followers for your own brand by making your own social-powered RSS feed. And, with a Zapier-powered filter, you can make a filtered RSS Super feed that includes only the most important posts.

Twitter previously included an RSS feed on each Twitter profile but removed it several years ago. It's not hard to make a new RSS feed for any Twitter account, though.

Just log in to Zapier—or create a free account if you don't have one already. Click the *Create a Zap* button, and select the Twitter integration. You can choose to have your Zap watch for new Tweets from a user or list, monitor search queries, or track mentions of a hash tag or user. To make an RSS feed of a Twitter profile, select User Tweet (or My Tweet for your own Tweets).

V. CONCLUSION

We presented a fundamental evaluation of ML algorithms on the detection of streaming spam tweets. In the evaluation, I found that classifiers ability to detect Twitter spam reduced when in a near real world scenario since the imbalanced data brings bias. I also identified that Feature discretization was an important preprocesses to ML-based spam detection. Second, increasing training data only cannot bring more benefits to detect Twitter spam after a certain number of training samples. Firstly identify the “Spam Drift” problem in statistical features based Twitter spam detection. In order to solve this problem, I propose a Lfun approach. In the Lfun scheme, classifiers will be re-trained by the added “changed spam” tweets which are learnt from unlabelled samples, thus it can reduce the impact of “Spam Drift” significantly. There is also a limitation in the Lfun scheme. The benefit of “old” labeled spam is to eliminate the impact of “spam drift” to classify more accurate spam tweets in future days. The effectiveness of “old” spam has been proved by the experiments during a short period.

A. DRAWBACK OF EXISTING SYSTEM

Existing machine learning based spam detection methods suffer from the problem of “Spam Drift” due to the change of statistical features of spam tweets as time goes on. When “spam drifts”, the old classification model is not updated with “changed” spam samples, as a result, the classification results will gradually become inaccurate. To solve this problem, obtaining the “changed” samples to update the classification model is very important. By observing that there are such samples in the unlabelled incoming tweets which are very easy to collect, I propose a scheme called “Lfun” to address “Spam Drift” problem.

B. BAYESIAN(STATISTICAL) ANALYSIS

Bayesian analysis is a very powerful technique widely used in filtering spam messages. This can be viewed as an upgraded version of the keyword checking method. This technique exploits the fact that many words which appears in a spam message is much less likely to appear in a ham message and vice versa. It works by scanning through the message word by word and check their spam probabilities (The likelihood of this word to appear in a spam message) using an established database. These probabilities are then used to compute the probability of that email being spam (also called its spam city) using Bayes' theorem. This spam city value is then used to determine whether the email is spam or not. The most important part of this technique is the database. Because it is a statistical content analysis technique, it means that the larger the database or the larger the pool of words get, the more accurate this technique would be. One of the major advantages of this technique is that it employs an artificial intelligent approach when building the database. This meant that the database would grow when more and more data is analyzed. The database is also trained on a per-user basis which significantly increases the efficiency of the filtering process. We can use the twitter application in the Microsoft Excel. Which is the text based application for this project.

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 *Proc. IEEE Commun. Inf. Syst. Security Symp. (ICCCISS)*, Jun. 2015, pp. 8689–8694.
2. C. Chen, J. Zhang, Y. Xiang, and W. Zhou, "Asymmetric self-learning for tackling twitter spam drift," in *Proc. 3rd Int. Workshop Security Privacy Big Data (Big Security)*, Apr. 2015, pp. 237–242.
3. C. Chen, J. Zhang, Y. Xiang, W. Zhou, and J. Oliver, "Spammers are becoming 'smarter' on twitter," *IT Prof.*, vol. 18, no. 2, pp. 14–18, Apr. 2016.
4. E. M. Clark, J. R. Williams, C. A. Jones, R. A. Galbraith, C. M. Dan forth, and P. S. Dodd's, "Sifting robotic from organic text: A natural language approach for detecting automation on twitter," *J. Comput. Sci.*, vol. 16, p. 1–7, Sep. 2016.
5. *Whole Product Dynamic Real-World Protection Test*, Av Comparatives, accessed on Aug. 1, 2015
6. J. Gama, I. Žliobait˙e, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *J. ACM Comput. Surv.* vol. 46, no. 4, p. 44, Apr. 2014.
7. H. Gao *et al.*, "Spam ain't as diverse as it seems: Throttling OSN spam with templates underneath," in *Proc. 30th Annu. Comput. Security Appl. Conf.*, 2014, pp. 76–85.
8. A. Greig. (2013). *Twitter Overtakes Face book as the Most Popular Social Network for Teens, According to Study*, Daily Mail, accessed on Aug. 1, 2015. [Online]. Available: http://www.dailymail.co.uk/news/article-2475591/Twitter-overtakes-Facebook-popular-socialnetwork- teens-according-study.html.
9. A. Gupta, P. Kumar guru, C. Castillo, and P. Meier, *TweetCred: Real- Time Credibility Assessment of Content on Twitter*. New York City, NY, USA: Springer, 2014.
10. R. Jeyaraman. (2014). *Fighting Spam with Botmaker*, Twitter, accessed on Aug. 1, 2015. [Online]. Available: https://blog.twitter.com/2014/fighting-spam-with-botmaker.