Machine Learning For Prediction Of Malicious Or SPAM Users On Social Networks

Gaurav Kumar, Dr. Vinay Rishiwal

Abstract: The use of social networking is very much prevalent now a days as a part of an individual for interaction with friends and family. The reason of such development is the limited time and greater geographical distances between the net users. Further it provide the easy mean for interacting. But at the same time the social interaction platforms are very much at the target of spammers and content polluters. In this regards it is very high time for researchers to explore and provide the necessary framework for monitoring and identifying such individuals and bots which are based on machine learning approach. A manual way of identifications using human intervention is definitely a very time consuming and rigorous process. In this paper a provisioning of machine learning based approach based on artificial neural networks usage in determined for identifications of such bots or individuals. Various types of neural networks and tier relative outcomes are compared and evaluated for spam identification in this paper.

Index Terms: SPAM, Machine Learning, ANN, Social Networks, BOTS, .

1. INTRODUCTION

The use of social media is one of most common way for millions of user to share the information and get the updates associated with their friends and known persons. C. P.-Y. Chin, N. Evans, and K.-K. R. Choo have explored the various factors influencing the social network firms that reflected online social networks (OSNs), the most popular one are Twitter, Facebook, considered as enterprise systems [1], with popularity growth and number of users grown exponentially. Due to lack to personal interaction with reason being different geographical locations and timings individuals spends social interaction time in OSNs for information sharing on birthday and other social political issues through text, images, emozies and videos. H. Tsukayama given the over view of twitter fro the beginning. Twitter, started in 2006, can be considered as one of the most popular micro blogging site which has vast base in terms of users and most of the time users share their views on the political and global issues as favor or against. An estimated figure of 200 million Twitter users provide 400 million new tweets every day [2]. The spam associated with is referred as unsolicited or non-required tweets those may contain malicious links that directs legitimate users to vulnerable sites containing phishing, drugs, sex rackets, honeypots, etc. given by F. Benevenuto, G. Magno et. al [3], the result is overall reputation of platform got tarnished and at the same time effects legitimate users. On 11 June, 2019 Mr. Amitabh Bacchant tweeter's account got hacked by Turkish hackers as they put face of Pakistan prime minister. Many of its followers received direct spam messages which contained malicious links. The similar example is referred with respected to electoral commission of Australia twitter account hacked cited by author [4]. Therefore it is the need of the hour to find out various approaches with ability to sort out useful information is critical for both academia and industry to discover hidden insights and predict trends on Twitter. Because spam are a sort of pollution on tweeter [5]. Identification of such tweeter accounts is very much needed and suspension mechanism should be automatically incorporated in such scenario. If a friend request so very frequently raised by an account and duplicate contents are transferred then such type of content may be spam. And may be suspended by tweeter [6]. Another option is official@spamaccount can be reported by twitter users for spammer. To automatically detect spam, machine learning algorithms have been applied by researchers to make spam detection as a classification problem [3], [7],[23]. 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 [3], must be retrieved from the users’ tweets list; features such as, “average neighbors’ tweets” in [13] and “distance” in [17] cannot be extracted without the built social graph. According to A. Bifet and E. Frank However, Twitter data are in the form of stream, and tweets arrive at very high speed [24]. 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 contents provided on the OSN sites by legitimate users is very much there as daily routine but at the same time the misuse of these sites by the spam senders and chat bots is very much trending as a vulnerable side by spammers or malicious users too. The spammers tries to send the content as legitimate users and may consider their spoofed presence with very much relevant content. The paper gives an overview of various basic algorithms of machine learning to classify the contents on tweeter or Facebook as spam or non span. One of the key highlights of these frameworks is their dependence on clients as essential donors of substance and as annotators and raters of other’s substance. This dependence on clients can lead to numerous positive impacts, counting large-scale development within the size and substance within the community, bottom-up revelation of “citizen-experts”, fortunate disclosure of modern assets past the scope of the framework originators, and modern social-based data look and recovery calculations. Spammers. In specific, social spammers are progressively focusing on these frameworks as portion of phishing assaults [14], to spread malware [5] and commercial spam messages [7], [26], and to advance affiliate websites [17]. Within the past year alone, more than 80% of social organizing clients have “received undesirable companion demands, messages, or postings on the online posts.

- *Gaurav Kumar is currently pursuing Ph.D. program in computer science and information technology Department, in Mahatma Jyotiba Phule Rohilkhand University UP India,. E-mail: midhagaurav1@rediffmail.com*
- *Dr. Vinay Rishiwal is currently working as Associate Professor in computer science and information technology Department, in Mahatma Jyotiba Phule Rohilkhand University UP India. E-mail: rishi4u100@gmail.com*

2 PROBLEM STATEMENT AND FRAMEWORK

As per the classification algorithms available following are the major algorithms available for binary classification which were considered in this paper.

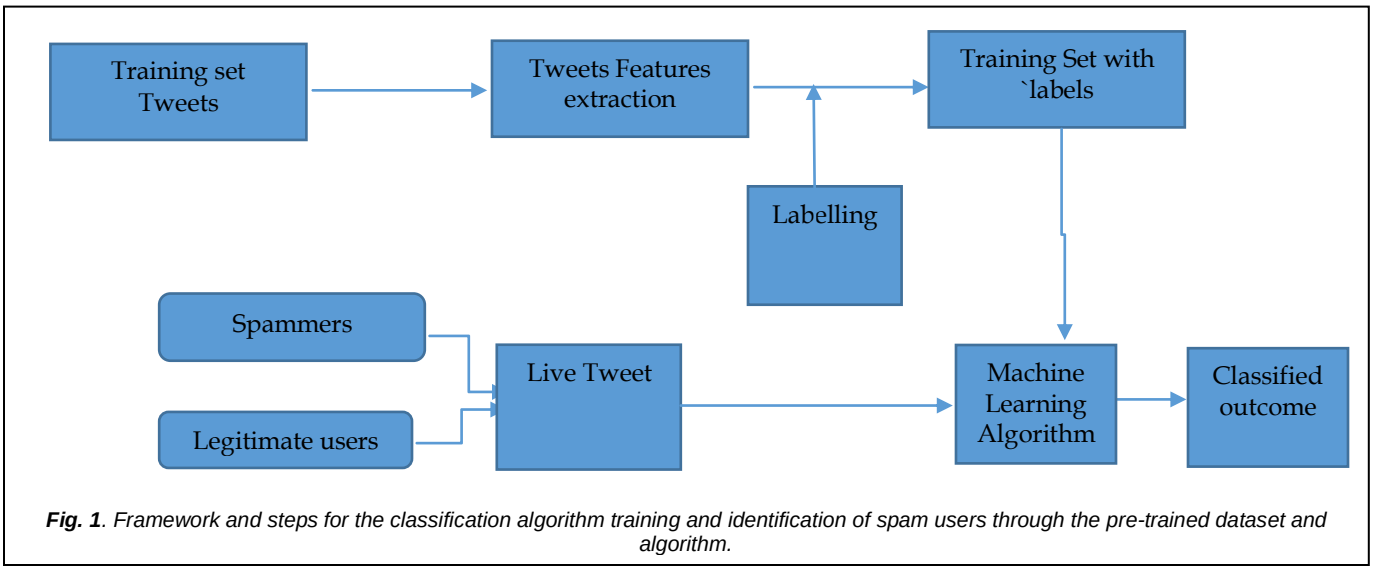

***Fig. 1**. Framework and steps for the classification algorithm training and identification of spam users through the pre-trained dataset and algorithm.*

2.1 Problem Statement

Fig1. Provide an overview and steps pictorial representation where we have already available data set which is required to be On the social interaction platform such as tweeter or facebook there we can consider the set of m users N={n_1,n_2, n_3, n_4,…..n_k}Each user tries to send some message that comprise some words. Those words can be considered as bag of words for identification of spam and non spam users as per the labeled data available on the plat form available by the legitimate users. Let the profile of user n_k is v_kThe objective of the problem is to classify the user n_k is spammer or malicious or not. Mathematically the set can be defined as

S:n_k→ {malicious user(spammer), legitimate user(non spammer} (1)

To identify the user as spammer first step is to identify the features of the tweeted text which contains certain words which can help in the identification of spam content. These contents can be further segregated in the words those can be termed as features for the data segregation. Thereby to build the given set the features can be given for identification of malicious user and non-malicious users as W={w_1,w_2,w_3,…… w_{m}} from S for profile v_k. The given accounts if identified as malicious on the basis of contents, and those can be blocked to stop their malefic intentions. A classifier can give the real time analysis of live streaming of messages and give a relative outcome on the basis of method chosen for classification.

2.2 Approach for Solving the Problem

The problem can be solved by considering the textual information of the tweet as main source for the feature extraction. The features can be part of the tweet text which can be disintegrate in the different words associated. On the basis of pre-labeled approach we can have groups of data which is labeled either SPAM or malicious and other type can be considered as legitimate text message from the legitimate user or non-spam user. The given labeled data can be used for the machine learning part where some data can be used for training and part of the data can be used for testing the particular algorithm of machine learning.

a. Multinomial Naïve Bayes considers a vector of features and classify the given data on the bases of linear approach when expressed with the log space. The problem starts with the events set (e1, e2, e3, e4…en) with existence probabilities associated with them (p1, p2, p3….pn). The likelihood prediction of class c_k for the given by the following expression

$$p(C_k|e_1e_2e_3 \dots e_k) = P(c_k)\prod_{i=1}^{n} P(e_i|c_k) \quad (2)$$

b. The multinomial Naïve bays classification can be applied on bag of words and works quite efficiently in binary classification where text based bag of words and vectors containing tf-idf can take entire tweet text with the applied label

c. Bernoulli Naïve Bayes Classifier. It is good in case if large dataset is there for the purpose of classification

$$p(C_k|e_1e_2e_3 \dots e_k) = P(c_k)/\| \; \|P(e_i|c_k) \quad (3)$$

d. Support Vector Machine is a discriminative classifier based upon the concept of hyper plane that use to separate the two classes. The equation of hyperplane in as given

Wx+b=0 (4)

3 EXPERIMENT AND OUTCOMES

On the basis of algorithms discussed in the last section. A data set of Tweets containing the text messages and their categorization labeling was taken with total 6000 tweets text which are already labelled as spam or malicious and ham or non-malicious. The following steps carried out for the experimenting with given data set with python 3.7 environment.

3.1The Dataset segregation into Spam(Malicious) and Non Spam

The given Dataset is taken into individual words and the converted all the words into lowercase. A removal process of stop words is considered further which does not effect the overall evaluation process of the message which is tokenized in words. Non English words left with no any specific emphasis. The data is further vectorized to count occurrence with TF-IDF.

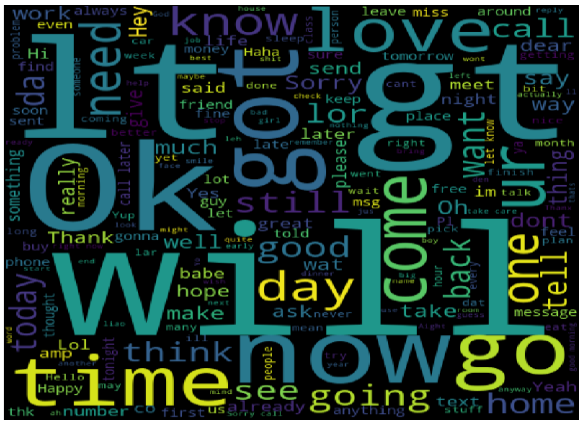

Fig. *2. Represent the non spam text word cloud after processing the data.*

The occurrence weight of the words then can be diagrammatically represented as word clouds for malicious and non-malicious words.

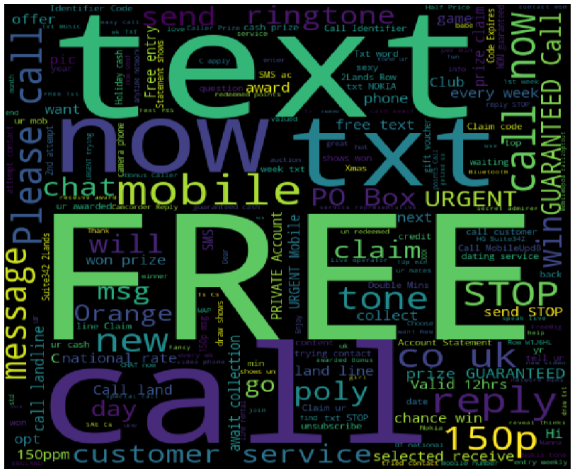

Fig. *3. Represent the spam or malicious text word cloud after processing the data.*

The spam or malicious word cloud for the same data is represented as in fig3. The data which is already labelled and split in to two sets with 80% for training and 20% for testing

3.2 Implementation of data mining algorithms discussed

Various algorithm trained for the purpose of analysis of data. The training and testing outcome for various algorithms after evaluation on the data for the which was segregated in two parts is given in the Fig. 4

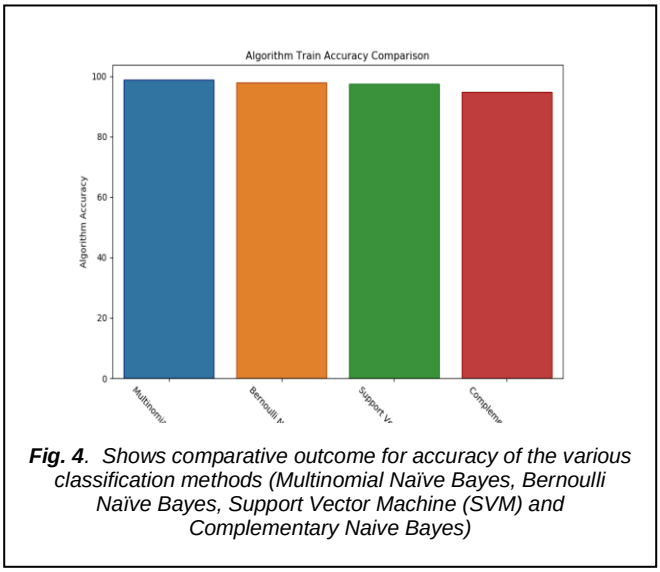

Fig. 4*. Shows comparative outcome for accuracy of the various classification methods (Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Support Vector Machine (SVM) and Complementary Naive Bayes)*

3.3 Performance Evaluation Metrics

The main performance evaluation metrics used by most of the researchers to evaluate the working of the classification algorithm considers the following.

a) Positives and Negatives: True Positive (TP) is a type of estimation where the spam which is considered to belonged to spam class also evaluated as spam by the classification algorithm. True Negative (TN) is a type of estimation that checks whether the non spam tweet which was to be estimated by classifier is evaluated as non-spam. The same thing is applicable to the non spam labelled tweets which are further categorized as spam (FP) by the algorithm and FN alternatively for spam which are classified as non spam. Table 1 represents the positive and negative performance Evaluation Matrix

Table 1
Performance Evaluation Metrics

		Prediction by Algorithm	
	Actual Labelled	Spam or Malicious	Non Spam or Non Malicious
True	Spam or Malicious	TP	FN
False	Non spam	FP	TN

The results of the given data after processing for the above specified table 1 are as given.

Fig 5 is for the Multinomial Naïve Bayes confusion Matrix

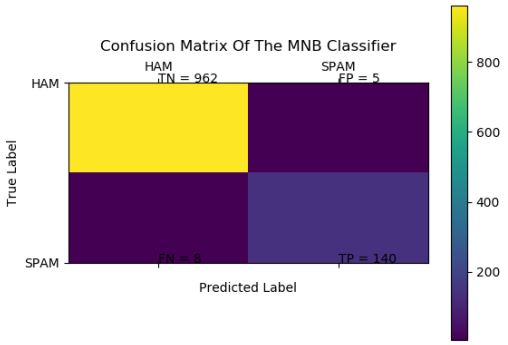

Fig. 5. *Multinomial Naïve Bayes confusion matrix with 20 % data test*

Fig 6 shows the confusion matrix of Bernoulli Naïve Bayes comparison outcome

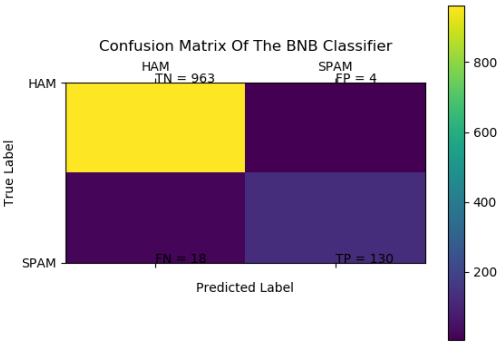

Fig. 6. *Bernoulli Naïve Bayes confusion matrix with 20 % data test*

Fig 7 shows the confusion matrix of Complement Naïve Bayes comparison outcome

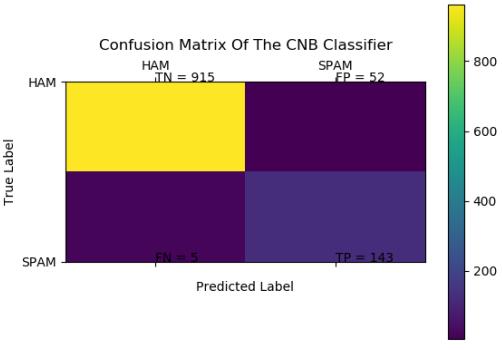

Fig. 7. *Complement Naïve Bayes confusion matrix with 20 % data test*

Fig 8 shows the confusion matrix of Support Vector Machine comparison outcome

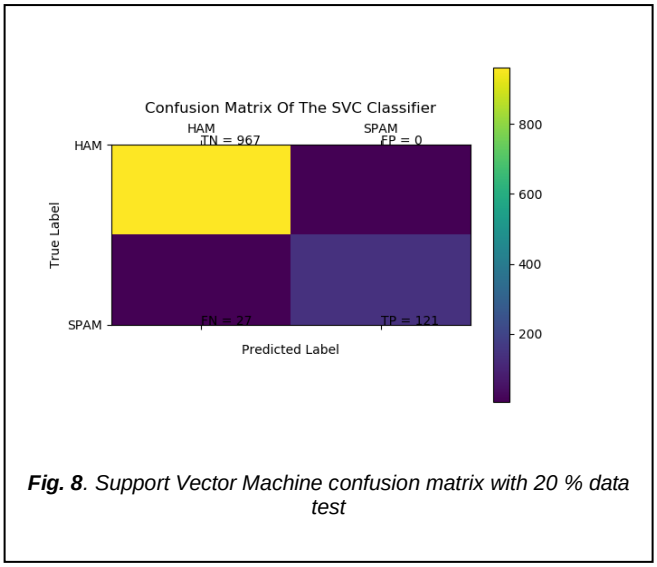

Fig. 8. *Support Vector Machine confusion matrix with 20 % data test*

b) TPR Rate : The true positive rate is defined as ration of TP to total numbers of predicted as spam(TP+FN)

$$TPR = \frac{TP}{TP + FN}$$

c) FPR Rate: It is ratio of non-malicious or non-spam tweets wrongly classified put in the category of spam class S to the sum of all spam tweets.

$$FPR = \frac{FP}{FP + FN}$$

d) Precision and Recall

The Precision in the factor which required to calculate the which calculates the ratio of TP to sum of spam tweets(TP+FP)

$$\mathrm{Precision} = \frac{TP}{TP + FP}$$

Recall is given as ratio to the true positive spam classification to the total number of users in the given class or category

$$\mathrm{Recall} = \frac{TP}{TP + FN}$$

e) F-Measure: A relation between recall and precision and mostly used for the pre-class evaluation

$$F - Measure = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

The experimental outcome for the above specified parameters are given with 5572 tweets with labeled spam and non-spam

classes are already available for the experiment in the given data set the initial data processing provide the following features calculation, length count, mean, standard deviation with the percentage variations as shown in Table 2. The bifurcation of all tweets is considered with two sets. One for the training with 80% input of existing labeled data set and 20% with training and comparison with the real outcome and as per the classifier expected outcome

Table 2

DATASET GROUPED BY LABEL

	Length Count	mean	std	min	25%	50%	75%	Max
NS	4825	70.71	57.7	2	33.0	52.0	91	910
S	747	138.32	29.0	13	132	149	157	223

The accuracy score of the given data which was already given in the Fig. 4. is given in table 5 .

Table 3

DATASET GROUPED BY LABEL

Classifier	Accuracy Level
Multinomial Naive Bayes score:	99%
Bernoulli Naive Bayes score:	98%
Support vector machine score:	98%
Complement Naive Bayes machine score:	95%

The most effective and best classification of spam account detection is considered with Multiple Naïve Bayes and Bernoulli Naïve Bayes classifier and then other outcomes lying in the list thereafter.

Table 4

True Positive and False Positive Outcome of classifiers

	TN	FP	TP	FN
Multinomial Naive Bayes	962	5	140	8
Bernoulli Naive	963	4	130	18
Complement Naive Bayes	915	52	143	5
Support vector Machine	967	0	121	27

The other measures TPR, FPR, Precision and Recall on the basis of above specified measures are given in table 5.

Table 5

TPR, FPR and Precision and Recall Values

	TPR	FPR	Precision	Recall
Multinomial Naive Bayes	0.946	0.385	0.995	0.946
Bernoulli Naive	0.878	0.182	0.996	0.878
Complement Naive Bayes	0.966	0.912	0.946	0.966
Support vector Machine	0.818	0.000	1.000	0.818

As per the data given the most optimistic result for the calculation of the multinomial naïve bays which predict the spam users very correctly but at the same time if SVM result is calculated for the false positive metrics is quite encouraging. The ROC curve for the above specified algorithms are calculated and the following outcome is appearing in the given graph. AUC - ROC bend is an exhibition estimation for grouping issue at different edges settings. ROC is a likelihood bend and AUC speaks to degree or proportion of distinguishableness. It tells how much model is fit for recognizing classes. Higher the AUC, better the model is at foreseeing 0s as 0s and 1s as 1s. By similarity, Higher the AUC, better the model is at recognizing spam and non spam with more appropriate outcome. An incredible model has AUC close to the 1 which implies it has great proportion of detachability. A poor model has AUC close to the 0 which implies it has most exceedingly awful proportion of distinctness. Truth be told it implies it is responding the outcome. It is anticipating 0s as 1s and 1s as 0s. Also, when AUC is 0.5, it implies model has no class partition limit at all. The following figure shows the ROC-AUC outcome of the various classification models discussed in the paper for the purpose of comparison but at the same time it can be seen most of the discussed algorithms have better value of ROC for the small data set as per the binary classification as given in Fig. 9. where as other metrics shows variation in the accuracy.

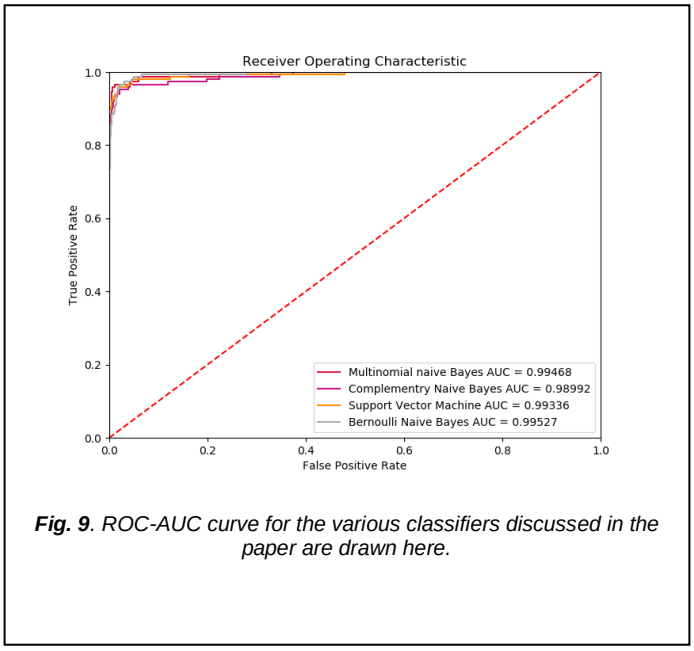

Fig. 9*. ROC-AUC curve for the various classifiers discussed in the paper are drawn here.*

4 CONCLUSION

In this paper, we give an essential assessment of ML calculations on the location of spilling spam tweets. So as to play out this assessment, we first gathered the data for the pre-labeled tweets and that was applied snow bowling framework which further calculated the tf-idf feature associated with labelled tweets and provided the frequencies of occurrence of specific terms in the tweet text and . Besides, we utilized cdf figures to delineate the attributes of removed highlights. We utilized these highlights to AI based spam classification later in our examinations. To examine the

capacity of spam identification of various classifiers, we inspected four diverse datasets to mimic different situations. In our assessment, we found that classifiers' capacity to distinguish Twitter spam diminished when in a close to genuine situation since the imbalanced information brings predisposition. We likewise identified that Feature discretization was a significant pre-procedure to ML-based spam identification. Second, expanding preparing information just can't bring more benefits to recognize Twitter spam after a specific number of preparing tests. We should attempt to bring progressively discriminative highlights or better model to additionally improve spam identification rate. Third, classifiers can identify more spam tweets when the tweets were examined ceaselessly instead of haphazardly chose tweets. From the third point, we completely investigated the motivation behind why classifiers' exhibitions diminished when preparing and testing information were in various days from three point of perspectives. We reason that the presentation diminishes because of the way that the appropriation of highlights changes of later days' dataset, though the dispersion of preparing dataset remains the equivalent. This issue will existing spilling spam tweets identification, as the new tweets are coming in the types of streams, however the preparation dataset isn't refreshed. We will take a shot at this issue later on.

REFERENCES

[1] C. P.-Y. Chin, N. Evans, and K.-K. R. Choo, “Exploring factors influencing the use of enterprise social networks in multinational professional service firms,” J. Organizat. Comput. Electron. Commerce, vol. 25, no. 3, pp. 289–315, 2015.

[2] H. Tsukayama, “Twitter turns 7: Users send over 400 million tweets per day,” Washingon Post, Mar. 2013 [Online]. Available: http://articles. washingtonpost.com/2013-0321/business/37889387_1_tweets-jack-dorsey-twitter

[3] 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.

[4] L. Timson, “Electoral commission Twitter account hacked, voters asked not to click,” Sydney Morning Herald, Aug. 2013 [Online]. Available: http://www.smh.com.au/it-pro/security-it/electoral-commission-twitteraccount-hacked-voters-asked-not-to-click-20130807-hv1b5.html

[5] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, “Twitter spammer detection using data stream clustering,” Inf. Sci., vol. 260, pp. 64–73, Mar. 2014.

[6] 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.

[7] 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.

[8] X. Jin, C. X. Lin, J. Luo, and J. Han, “Social spam guard: A data miningbased spam detection system for social media networks,” PVLDB, vol. 4, no. 12, pp. 1458–1461, 2011.

[9] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, “Aiding the detection of fake accounts in large scale social online services,” in Proc. Symp. Netw. Syst. Des. Implement. (NSDI), 2012, pp. 197–210.

[10] S. Ghosh et al., “Understanding and combating link farming in the Twitter social network,” in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 61– 70.

[11] H. Costa, F. Benevenuto, and L. H. C. Merschmann, “Detecting tip spam in location-based social networks,” in Proc. 28th Annu. ACM Symp. Appl. Comput., 2013, pp. 724–729.

[12] E. Tan, L. Guo, X. Zhang, and Y. Zhao, “Unik: Unsupervised social network spam detection,” in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage., San Fransisco, CA, USA, Oct. 2013, pp. 479–488.

[13] C. Yang, R. Harkreader, and G. 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.

[14] S. Lee and J. Kim, “Warning bird: A near real-time detection system for suspicious URLs in Twitter stream,” IEEE Trans. Dependable Secure Comput., vol. 10, no. 3, pp. 183–195, May/Jun. 2013.

[15] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, “Analyzing spammers’ social networks for fun and profit: A case study of cyber criminal ecosystem on Twitter,” in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 71–80.

[16] X. Zhang, S. Zhu, and W. Liang, “Detecting spam and promoting campaigns in the Twitter social network,” in Proc. IEEE 12th Int. Conf. Data Mining (ICDM), 2012, pp. 1194–1199.

[17] J. Song, S. Lee, and J. Kim, “Spam filtering in Twitter using senderreceiver relationship,” in Proc. 14th Int. Conf. Recent Adv. Intrusion Detect., 2011, pp. 301–317.

[18] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, “COMPA: Detecting compromised accounts on social networks,” presented at the 20th. Annu. Netw. Distrib. Syst. Sec. Symp., San Diego, CA, USA, Feb.. 24–27, 2013.

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

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

[21] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd, “Detecting spam in a Twitter network,” First Monday, vol. 15, nos. 1–4, Jan. 2010.

[22] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a social network or a news media?” in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 591–600

[23] C. Chen, J. Zhang, Y. Xiang, and W. Zhou, “Asymmetric self-learning for tackling Twitter spam drift,” in Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM), Apr. 2015, pp. 208–213.

[24] A. Bifet and E. Frank, “Sentiment knowledge discovery in Twitter streaming data,” in Proc. 13th Int. Conf. Discov. Sci., 2010, pp. 1–15.

[25] B. Wang, A. Zubiaga, M. Liakata, and R. Procter, “Making the most of tweet-inherent features for social spam detection on Twitter,” arXiv preprint arXiv:1503.07405, 2015.

[26] J. Oliver, P. Pajares, C. Ke, C. Chen, and Y. Xiang, “An in-depth analysis of abuse on Twitter,” Trend Micro, Irving, TX, USA, Tech. Rep., Sep. 2014.

[27] 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.
[28] 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.
[29] K. Lee, J. Caverlee, Z. Cheng, and D. Z. Sui, “Content-driven detection of campaigns in social media,” in Proc. 20th ACM Int. Conf. Inf. Knowl. Manage., 2011, pp. 551–556.