Improving Spam Detection in Online Social Networks

Arushi Gupta
Department of Information Technology
Indira Gandhi Delhi Technical University for Women
Kashmere Gate, Delhi
arushigupta12@gmail.com

Rishabh Kaushal
Department of Information Technology
Indira Gandhi Delhi Technical University for Women
Kashmere Gate, Delhi
rishabh.kaushal@gmail.com

***Abstract* — Online Social Networks (OSNs) are deemed to be the most sought-after societal tool used by the masses world over to communicate and transmit information. Our dependence on these platforms for seeking opinions, news, updates, etc. is increasing. While it is true that OSNs have become a new medium for dissemination of information, at the same time, they are also fast becoming a playground for the spread of misinformation, propaganda, fake news, rumors, unsolicited messages, etc. Consequently, we can say that an OSN platform comprises of two kinds of users namely, Spammers and Non-Spammers. Spammers, out of malicious intent, post either unwanted (or irrelevant) information or spread misinformation on OSN platforms. As part of our work, we propose mechanisms to detect such users (Spammers) in Twitter social network (a popular OSN). Our work is based on a number of features at tweet-level and user-level like Followers/Followees, URLs, Spam Words, Replies and HashTags. In our work, we have applied three learning algorithms namely Naive Bayes, Clustering and Decision trees. Furthermore, to improve detection of Spammers, a novel integrated approach is proposed which "combines" the advantages of the three learning algorithms mentioned above. Improvement of spam detection is measured on the basis of Total Accuracy, Spammers Detection Accuracy and Non-Spammers Detection Accuracy. Results, thus obtained, show that our novel integrated approach that combines all algorithms outperforms other classical approaches in terms of overall accuracy and detect Non-Spammers with 99% accuracy with an overall accuracy of 87.9%.**

Keywords—Spam detection, classification algorithm, twitter, online social networking.

I. INTRODUCTION

Online Social Networks (OSNs) are a platform where people with common interests and beliefs, interacts and connect. People visit OSN platforms to collect information relevant to them and also to build social and professional networks. OSNs like Facebook, Twitter and LinkedIn are used by millions of users worldwide for fostering interpersonal relationships and the number of users using these OSNs is increasing rapidly every day. These OSNs are becoming a new platform for dissemination of information, opinions and news. However, at the same time, some of the users, called Spammers, are misusing these OSN platforms, thereby spreading misinformation, propaganda, rumors, fake news, unsolicited messages, etc. Sometimes, this spamming is done with the intent of advertising and other commercial purposes, where spammers subscribe to various mailing lists and then send spam messages indiscriminately to promulgate their interests. Such activities disturb the genuine users, called Non-Spammers and also decrease the reputation of OSN platforms. Therefore, there is a need to devise mechanisms to detect Spammers so that corrective actions can be taken thereafter.

Our work focuses on detection of Spammers over one of the most popular OSN platforms, Twitter [1]. Twitter is viewed as one of the most prevalent and sought after online website utilized for micro blogging. Founded in March 2006, it was an instant hit in the Internet space, with more than 100 million users registering by 2012 which increased to 500 million users by July 2014. Its users can send out 140 character short messages which can be accessed by its other users across different interfaces such as the Twitter website, SMS and mobile device app. Twitter is chosen as an OSN platform for our work because it offers a large number of user bases and also because information on Twitter is publicly available by default which can be accessed through APIs provided by Twitter.

Being one of the most prominent OSNs, Twitter is continuously under attack by Spammers. One of the modus operandi of Spammers is that they gain credibility of users by following accounts of famous celebrities, and when those accounts follows them back, they get legitimized in some sense, thereby leading other Twitter users to follow them. The Spammers, thereafter, successfully proliferates spam messages among their highly connected communities. Another way in which Spammers work is by sending the victim large number of direct messages called Direct Messaging (DM) spamming. In addition to this, there exists a huge black market, which allows a person who intends to spam to buy a million followers that mainly consist of fake accounts in order to disguise as a genuine user (Non-Spammer). This not only

978-1-4799-7171-8/15/$31.00 ©2015 IEEE

gives the identity of a spammer a legitimate appearance but also enables DM spamming, where the spammer can send direct messages that contain malicious content. There are many other techniques through which spammers can possibly gain popularity and spread malicious tweets.

In our work, we obtained the Twitter dataset from the authors of [2] and thereafter have performed preprocessing over it to obtain normalized set of features based on which the activities of spammers were studied. The key features extracted were Followers/Followees, URLs, Spam Words, Replies and HashTags. After obtaining these features, three classical learning algorithms namely Naive Bayes, Clustering and Decision trees were applied on these features to detect Spammers in the dataset. Moving on, in order to improve detection of Spammers, a novel proposed approach was devised which "combines" the advantages of the three classical algorithms (Naive Bayes, Clustering and Decision Trees).

Finally, the improvement in spam detection is measured on the basis of accuracy parameters and the results, thus obtained show that our novel integrated approach that combines all algorithms outperforms other classical approaches in terms of overall accuracy and non-spammer detection accuracy.

The key contribution of our work is the proposed integrated approach which combines the three learning algorithms namely Naive Bayes, Clustering and Decision Trees with an aim of improving spam detection accuracy.

II. Related Work

The issue of spamming over emails and in many other forms is a well studied problem. Spam Detection has been the area of interest of many researchers. Many solutions have been propounded in regard to spam detection. However, spam detection in the social networks, which is a recent phenomenon, has not been studied so widely. Also, the fact that Tweet messages are small in size, restricted to 140 characters only (as opposed to email or web content), the problem of spam detection becomes more difficult. This section summarizes the main contributions of other researchers on spam detection in social networks. Fabricio Benevenuto *et al.* [2] detected spammers by identifying various user social behaviors and the characteristics of tweet content. These characteristics were used in a machine learning approach to classify the users as spammers and non-spammers. De Wang *et al.* [3] in their study proposed a general framework to detect spam account across all the OSNs. The main contribution of their work was a new spam detected in any one social networking could be quickly identified across all other OSNs. Alex Hai Wang *et al.* [5] proposed a model which uses a directed graph that depicts the relationship between "friends" and "follower" relationship in twitter. Bayesian Classifier was also used in his work, to detect spam accounts. Xin Jin *et al.* [6] propounded a method for detecting spam accounts in social media network. They employed a GAD Clustering algorithm integrated with designed active learning algorithm to deal with spam accounts. M. McCord and M. Chuah [7] discussed various features related to user and tweet content which can be utilized in the detection of accounts intended for spamming. In their work, he evaluated four classifiers and compared there accuracies. Carlos Castillo *et al.* [9] constructed a spam detection system that exploits the linked dependencies of web pages. The algorithm assumed that the linked web pages belonged to same class, i.e. if one web page is spam than its linked web pages must also be spam. Hongyu Gao *et al.* [11] studied spam accounts in one of the popular OSN, Facebook. A dataset of "wall" messages between various Facebook users was used to identify spam accounts. In their study, they found out that spamming was most common during early hours, when regular users were asleep. Kenichi Yoshida *et al.* [15] worked on email spam. They used an unsupervised approach to detect spam accounts. Benjamin Markines *et al.* [16] devised a mechanism to detect social spam. In their work, six features were recognized which were used to distinguish between spammers and non-spammer users. The features were passed to various machine learning algorithms which further classified the spam and non-spam accounts.

To the best of our knowledge, some of these works applied each of the machine learning algorithms separately but not in a combined manner which has resulted in improving spam detection, one of the key contributions of our work.

III. Proposed Approach

An overview of the complete process of spam detection is shown in the diagram in Figure 1, each of whose steps are explained in this section.

The preliminary step for the detection of spammers in any OSN is data collection and necessary preprocessing to convert it into a form, which can be used by the learning algorithms.

A. Data Set Description

In our work, we have used the dataset obtained from Fabricio Benevenuto *et al.* [2] which consists of labeled record of 1064

Twitter users. Dataset comprises of 62 features containing user specific and tweet specific information. The spammer accounts comprised of around 36% of the dataset. Also, as per [2], the users were chosen randomly and not based on any of their characteristics. They [2] have used SVM based machine learning approach as opposed to our work in which we have used other learning approaches namely Naive Bayes, Clustering, Decision Trees and finally combined all of them together to achieve a higher spam detection accuracy.

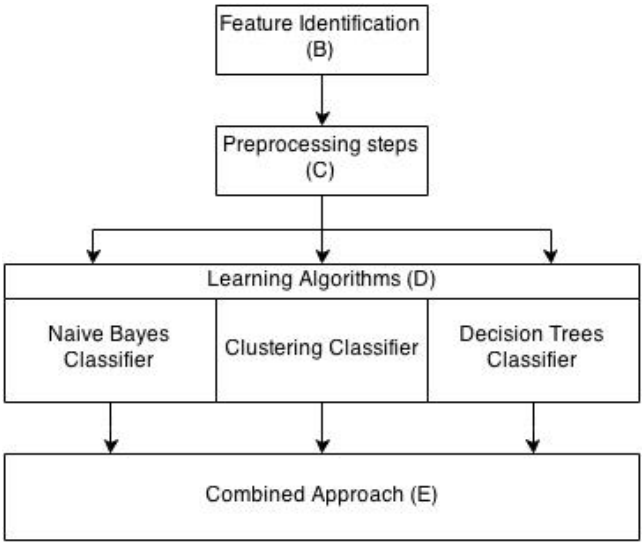

Figure 1: Proposed Spam Detection Approach

B. Feature Identification

Since, spammers behave differently from non-spammers; therefore we can identify some features or characteristics in which both these categories differ. Various features which we have used to detect spam accounts include:-

Number of followers and followees: Followers are the users who follow a particular user, while followees are the users whom the user follows. Generally speaking, spammers have small number of followers but follow large population with the motive to get noticed by many. Therefore, account with large followees and small number of followers can potentially be considered as a spam account.

URLs: URLs are the links which direct to some other page on the browser. With the development of URL shorteners, it has now become easy to post malicious links on any OSN. This is because URL shorteners hides the source of the link, thereby making it difficult for the detection algorithms (used to detect malicious links) to detect such links. Too many URLs in tweets of a user are a potential indicator of the user being a spammer.

Spam Words: An account with spam words in almost every tweet can be considered to be a spam account. Therefore, "Fraction of tweet with spam words" can be considered as an important factor for detecting spammers.

Replies: Since, information or message sent by a spammer is useless, therefore people rarely replies to its post. On the other hand, a spammer replies to a large number of posts in order to get noticed by many people. This pattern can be used in the detection of spammers.

Hashtags: Hashtags are the unique identifier ("#" followed by the identifier name) which is used to group similar tweets together under the same name. Spammers use large number of hashtags in their posts, so that their post is posted under all the hashtag categories and thereby gets wide viewership and is read by many.

C. Preprocesor

Twitter user accounts in the dataset [2], labeled as Spammer and Non-Spammers, and were used for training the learning algorithms and also in accuracy calculations.

In preprocessing step, all the continuous features were converted into discrete. The procedure adopted to select the intervals for a particular feature was obtained from [4] according to which all user accounts are arranged in increasing order of their feature values. Processing begins from the first account, if we encounter an account whose category is different from the category of the next account, and then an interval is created as a mean of both the feature values.

D. Learning Algorithms

There are various different classification algorithms, which can be used to classify an account as "Spammer" or "Non-Spammer". In our work, we have used Naive Bayes, Clustering and Decision trees as learning algorithms. Although, each of these approaches can be solely used to classify user accounts, but in order to increase the accuracy, we have combined these approaches into an integrated algorithm in our work.

In our proposed approach as outlined in Figure 1, the preprocessed data is first classified using different learning algorithms to predict the class {"Spammers" or "Non-Spammers"} of all Twitter user accounts.

In *Naive Bayes approach,* accounts were classified by calculating the probability of the given account to be

Spammer/Non-Spammer, given the feature values of that account. Bayes theorem was used to calculate this probability. Mathematically, Bayes theorem can be expressed into a simple form:

$$P(C \mid F) = \frac{P(F \mid C) * P(C)}{P(F)}$$

Here,
F is a vector of the features of a user
C is the class (Spammer/Non-Spammer) of the user.

If the calculated probability of an account to be spam is more than 0.5, than that account is classified as Spammer, otherwise it is considered as a Non-Spammer (genuine account).

Clustering is basically an unsupervised learning technique. Unlike Naive Bayes, a separate training data need not be prepared for this algorithm. On the basis of similar feature values (like similar kind of reply trend, similar usage of spam words in tweet), this algorithm could classify the entire set of accounts unto two classes. One of this class was labeled as Spammer and another as Non- Spammer.

Decision trees learning method was also used to classify Twitter user accounts. In this technique, a tree structure was prepared and the decision was made at every level of the tree. These decisions were made on the basis of feature values. The Twitter account to be classified was designated as root node and then after passing through a series of decisions, this node eventually reaches the leaf node, which then decides its class. Thus, in this way, the user accounts were classified.

E. Combined Approach

As part of combined approach, we compare the classification results of any two learning algorithms, if both the learning algorithms predict the same result, then we finalize the class of the Twitter account under investigation. Otherwise, if the predicted class of both the classification techniques differ, then we use the prediction of third algorithm as the final class.

Results obtained following this approach show an improvement in spam detection.

IV. Experiments & Evaluation

For the corroboration of the accuracy of the algorithm the results obtained from the algorithm were compared with the labels (Spammers/Non-Spammers) in the dataset [2]. Output of various algorithms implemented is depicted in Tables I – IV below.

Table I: Naive Bayes Results

	Predicted Values	
Actual Values	Non-Spammers	Spammers
Non-Spammers	225/232	7/232
Spammers	42/133	91/133

Table II: Clustering Results

	Predicted Values	
Actual Values	Non-Spammers	Spammers
Non-Spammers	230/232	2/232
Spammers	68/133	65/133

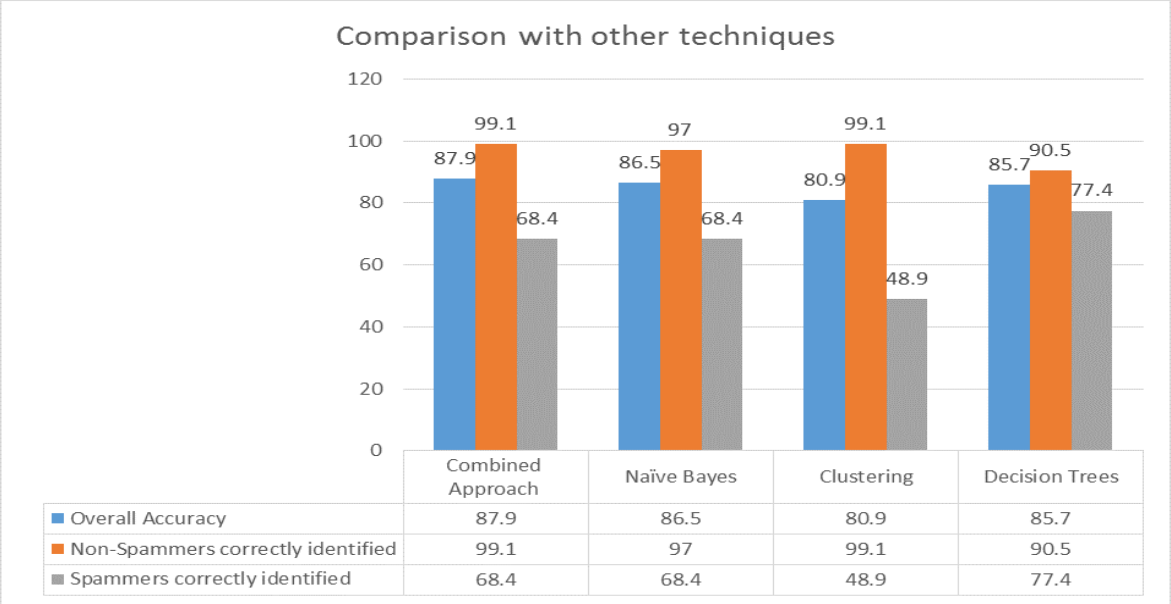

Figure 2: Comparison of Improvement in Spam Detection using four learning approaches

Table III: Decision Trees Results

	Predicted Values	
Actual Values	Non-Spammers	Spammers
Non-Spammers	210/232	22/232
Spammers	30/133	103/133

Table IV: Integrated Approach Results

	Predicted Values	
Actual Values	Non-Spammers	Spammers
Non-Spammers	230/232	2/232
Spammers	42/133	91/133

It is evident that the proposed algorithm was able to successfully identify an account as Spammer or Non-Spammer with 87.9% accuracy. The algorithm's accuracy of detection of non-spammers was higher (99.1%) as compared to the accuracy of detection of spammers (68.4%).

This integrated algorithm was then compared with each of the learning algorithm, Naive Bayes, Clustering and Decision Trees. The results showed that Clustering algorithm performs better in detection of non-spam accounts but was very poor in detecting spam accounts. Our algorithm was able to maintain the high accuracy of Clustering algorithm in detecting non-spam and at the same time, retain the accuracy of Naive Bayes in detecting Spammers accounts thereby, increasing the overall accuracy. The graphical representation of the above data is shown in Figure 2.

V. CONCLUSION & FUTURE WORK

In our work, an algorithm, combining three different learning algorithms (namely Naive Bayes, Clustering and Decision trees) was implemented. This integrated algorithm categorizes an account as Spammer/Non-Spammer with an overall accuracy of 87.9%.

Finally, this algorithm was compared with all the three learning algorithms taken alone. It was observed that the combined approach could give best results in terms of overall accuracy and in detection of non-spammers. Though, Decision Trees alone perform better in detection of spammers but it is poor in detecting non-spam accounts, thus it can't be used solely.

As a future work, the integrated approach can be further improved by maintaining high accuracy of Decision Trees approach with respect to detection of Spammer accounts.

VI. REFERENCES

[1] "Twitter" https://twitter.com

[2] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida, "Detecting Spammers on Twitter", Proceedings of Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS), 2010.

[3] De Wang, DaneshIrani, and Calton Pu, "A Social-Spam Detection Framework", Proceedings of Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS), 2011.

[4] Dewan Md. Farid, Nouria Harbi, and Mohammad Zahidur Rahman, "Combining Naive Bayes And Decision Tree For Adaptive Intrusion Detection", International Journal of Network Security & Its Applications (IJNSA), Volume 2, Number 2, April 2010.

[5] Alex Hai Wang, "Don't Follow Me: Spam Detection In Twitter", Proceedings of Security and Cryptography International Conference (SECRYPT), 2010.

[6] Xin Jin, Cindy Xide Lin, Jiebo Luo and Jiawei Han, "A Data Mining-based Spam Detection System for Social Media Networks", Proceedings of the VLDB Endowment, Volume 4, Number 12, August 2011.

[7] M. McCord and M. Chuah, "Spam Detection on Twitter Using Traditional Classifiers", Proceedings of Autonomic and Trusted Computing International Conference (ATC), 2011.

[8] Kurt Thomas, Chris Grier, Vern Paxson and Dawn Song, "Suspended Accounts in Retrospect: An Analysis of Twitter Spam", Internet measurement conference (IMC), 2011.

[9] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "Know your neighbors: Web spam detection using the web topology", in Int'l ACM SIGIR, 2007.

[10] G. Stringhini, C. Kruegel and G. Vigna, "Detecting Spammers on Social Networks", Proceedings of ACM ACSAS, 2010.

[11] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Zhao, "Detecting and characterizing social spam campaigns", Proceedings of the Internet Measurement Conference (IMC), 2010.

[12] F. Benevenuto, T. Rodrigues, V. Almeida, J. M. Almeida, C. Zhang, and K. W. Ross, "Identifying video Spammers in online social networks", in AIRWeb, pages 45–52, 2008.

[13] C. Pu and S. Webb, "Observed trends in spam construction techniques: a case study of spam evolution", Proceedings of Conference on Email and Anti-Spam (CEAS), 2006.

[14] Leyla Bilge, Thorsten Strufe, Davide Balzarotti and Engin Kirda, "All your contacts are belong to us: automated identity theft attacks on social networks", Proceedings of ACM World Wide Web Conference, 2009.

[15] K. Yoshida, F. Adachi, T. Washio, H. Motoda, T. Homma, A. Nakashima, H. Fujikawa, and K. Yamazaki, "Density-based spam detector", Proceedings of the Tenth ACM SIGKDD International Conference, 2004.
[16] B. Markines, C. Cattuto, and F. Menczer, "Social spam detection", in AIRWeb, pages 41–48, 2009.
[17] H. Drucker, D. Wu, and V. Vapnik, "Support vector machines for spam categorization", IEEE Transactions on Neural Networks, 10(5): pp. 1048–1054, 1999.
[18] S. Webb, J. Caverlee, and C. Pu, "Introducing the webb spam corpus: Using email spam to identify web spam automatically", Proceedings of the Conference on Email and Anti-Spam (CEAS), 2006.
[19] I. Drost and T. Scheffer, "Thwarting the nigritude ultramarine: Learning to identify link spam", Proceedings of the European Conference on Machine Learning (ECML), 2005.