International Journal of Scientific Research in Computer Science, Engineering and Information Technology

© 2017 IJSRCSEIT | Volume 2 | Issue 5 | ISSN : 2456-3307

Profiling Online Social Networks for Spam Detection

K. Srinivasan, M.Sc (IT)*1, V. Sureka, MCA, M.Phil2

*1M.Phil Research Scholar, Department of Computer Science,
2Assistant Professor, Department of Computer Science,
Sri Jayendra Saraswathy Maha Vidyalaya CAS, Coimbatore, Tamil Nadu, India

ABSTRACT

Social network has become a very general way for internet users to connect and interact online. Users spend sufficiently of time on famous social networks (e.g., Facebook, Twitter, Sina Weibo, etc.), reading news, discussing events and posting messages. Unfortunately, this popularity also attracts a significant amount of spammers who continuously expose malicious behaviour (e.g., post messages containing commercial URLs, following a larger amount of users, etc.), foremost to great misinterpretation and inconvenience on users' social activities. In this paper, a supervised machine learning based solution is proposed for an effective spam detection.

Keywords: OSN, k-NN classifier, Spam detection.

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 to build social and professional networks. Millions of users use OSNs like Facebook, Twitter and LinkedIn worldwide for fostering interpersonal relationships and the number of users using these OSNs is increasing rapidly every day [1]. 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, rumours, 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 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.

People are even able to use a feature in Facebook to automatically publish updates to their Twitter accounts simultaneously. The similar function can also be designed in other social networks, for example, Tumblr users can share the pictures or information to twitter and Facebook accounts. Most of the web pages have the functioned button at the bottom to allow viewers to share this page into various OSNs [2]. All of these make different OSN accounts for one-person exhibit high similarities. Unfortunately, high prosperity in OSNs gives rich soils for different kinds of spams. Spammers who aim to advertise their products or post victim links are more frequently spreading their malicious activities via different OSNs. Reports show that nearly 10% of tweets in Twitter are all spam, and Facebook usually blocks 200 million malicious actions every day [3]. Even if all companies developed approaches to limit the activities of spammers, spam volume is rapidly growing more than users' actions.

II. RELATED WORKS

Many researches have concentrated on this area to find efficient methods to identify spam, and are especially focused on the classification of different spam features. 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.

Sarita et. al. in [4] study structural properties of legitimate users and spammers and observe similarity between Web graph and Twitter's social graph. They hypothesize that normal users are at the center of social graph (following each other and some celebrities), celebrities are at one end (mostly being followed by normal users) while spammers lie at the other end following a lot of normal users. Zi Chu et. al. [5] analyze behavior of humans, bots and cyborgs on Twitter. According to their observations, bots post more URLs per tweet, post regularly throughout the day or week while humans tweet less on weekends and nights. They also observe that bots mostly post tweets using API-based tools while humans mostly use web interface for tweeting. They also note that bots have larger number of followings as compared to followers, while humans have similar number of followers and followings. Cyborgs, on the other hand, have larger followers than followings.

Benevenuto et. al. in [6] discuss rise of video spammers and promoters in video social networks like YouTube. They analyse users' behaviours on You tube and propose some features which could distinguish spammers from normal users and use supervised learning techniques to detect spammers and promoters on YouTube. Various features, including video-based, user-based, and social network based features are presented. Video-based features like number of views, comments, number of ratings etc. capture properties of the uploaded video. User based features like number of friends, videos watched; videos uploaded etc. give an idea about the up loader. Various social-network based features like clustering coefficient etc. try to distinguish spammers from benign users based on social relationship between friends.

Gao et. al. [7] present a technique to detect and characterize spam campaigns on Facebook. They collect an anonymized dataset of wall posts from Facebook and analyze them to identify spam campaigns on Facebook. They try to form a graph using the wall posts - adding an edge between posts with similar content or containing same destination URLs. Connected components in the graph signify similar posts contents by different users. Then they use bursty nature and geographically distributed nature of spammers in order to identify connected components which participate in spam campaigns.

III. RESULTS AND DISCUSSION

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.

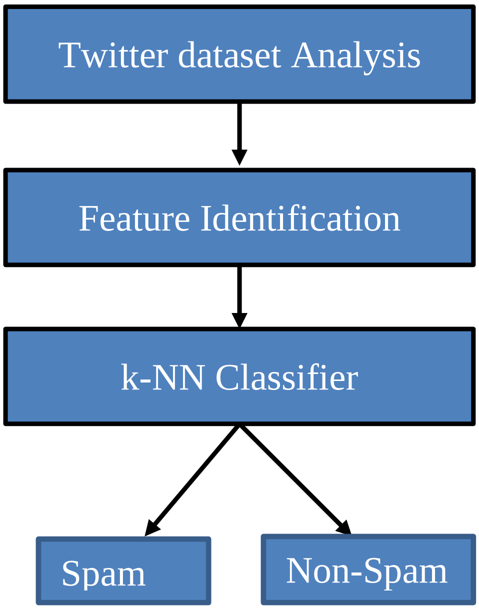

Figure 1. Proposed System

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. Twitter Data Set Description

In this work, we have used the dataset obtained from KAGGLE [8] which consists of labelled 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.

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. 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 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 be 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. K-nearest neighbor classifier

There are various different classification algorithms, which can be used to classify an account as “Spammer” or “NonSpammer”. In this work, we have used K-nearest neighbor classifier as learning algorithms.

K-nearest neighbour is a sophisticated approach for classification that finds a group of K objects in the training documents that are close to the test value [9]. To classify an unlabeled object, the distance between this object and labelled object is computed and it’s K nearest neighbours are identified. Classification accuracy mainly depends on the chosen value of K and will be better than that of using the nearest neighbour classifier. For large data sets, K can be larger to reduce the error. Choosing K can be done experimentally, where a number of patterns taken out from the training set can be classified using the remaining training patterns for different values of k. The value of K which gives the least error in classification will be chosen. If same class is shared between several of K-nearest neighbours, then per-neighbour weights of that class are added together, and the resulting weighted sum is used as the likelihood score of that class with respect to the test document.

The classification of KNN is easy to understand and implement and it can perform well in many situations. It is also scalable to new modifications as it is possible to eliminate many of the stored data objects, but still retain the classification accuracy of the KNN classifier. This is known as ‘condensing’ and can greatly speed up the classification of new objects but there comes the difficulty while deciding the value of K. If K is too small then result can be sensitive to noise points whereas if for large value of K, the neighborhood may include too many points from other classes. The choice of the distance measure is another important consideration [10]. Although various measures can be used to compute the distance between two points, but smaller distance between two objects does not always implies a greater likelihood of having the same class.

IV. EXPERIMENTAL RESULTS

The classification experiments are done using Weka [11], which had been one of the standard tools in data mining and machine learning. It contains various classification and clustering algorithms like Naïve Bayes, J-48, Random Tree, Random Forest, etc. We also use accuracy, precision, F1-Measure as criteria to evaluate the classification performance. The experimental result is shown in the table 1.

Measures	Experimental results
Accuracy	94.7
Precision	0.98
F1-Measure	0.79

Table 1. Experimental Results

Table 1 shows the performance of k-NN classifier. We can see that the k-NN classifier performs with accuracy of 94.7%. That means the classifier gained better performance in the ham tweets but poor performance in the spam tweets.

V. CONCLUSION

In this paper, we have introduced a machine learning based spam detection system for online social networks. The system starts with analyzing the twitter dataset, then identification of significant features and applied machine-learning technique to classify the data into spam and non-spam. Through k-NN classifier, we have achieved 94.7% accuracy.

VI. REFERENCES

[1]. Mccord, M., & Chuah, M. (2011, September). Spam detection on twitter using traditional classifiers. In international conference on Autonomic and trusted computing (pp. 175-186). Springer, Berlin, Heidelberg.

[2]. McMinn, A. J., Moshfeghi, Y., & Jose, J. M. (2013, October). Building a large-scale corpus for evaluating event detection on twitter. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (pp. 409-418). ACM.

[3]. Song, J., Lee, S., & Kim, J. (2011). Spam filtering in twitter using sender-receiver relationship. In Recent advances in intrusion detection (pp. 301-317). Springer Berlin/Heidelberg.

[4]. Yardi, S., Romero, D. Detecting spam in a twitter network. First Monday 15(1), pp. 7-14, 2010.

[5]. Zi Chu, Steven Gianvecchio, Haining Wang and Sushil Jajodia. Who is Tweeting on Twitter: Human, Bot, or Cyborg? In ACSAC, pp. 21-30, 2010.

[6]. F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonalves. Detecting Spammers and Content Promoters in Online Video Social Networks. In SIGIR, pp. 620-627, 2009.

[7]. H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Zhao. Detecting and Characterizing Social Spam Campaigns. In IMC, pp. 35-47, 2010.

[8]. Thomas, K., Grier, C., Song, D., & Paxson, V. (2011, November). Suspended accounts in retrospect: an analysis of twitter spam. In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference (pp. 243-258). ACM.

[9]. Agarwal, S., & Sureka, A. (2015, February). Using knn and svm based one-class classifier for detecting online radicalization on twitter. In International Conference on Distributed Computing and Internet Technology (pp. 431-442). Springer, Cham.

[10]. Galán-García, P., Puerta, J. G. D. L., Gómez, C. L., Santos, I., & Bringas, P. G. (2016). Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. Logic Journal of the IGPL, 24(1), 42-53.

[11]. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.