International Journal of Engineering & Technology, 7 (2.7) (2018) 631-635

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Spam Detection on Online Social Media Networks

G.N.V.G. Sirisha[1]*, G.V.Padma Raju[2], G. Amruta[3]

[1] Asst. Professor, Dept. of CSE, S.R.K.R. Engineering College, Bhimavaram, India
[2] Professor, Dept. of CSE, S.R.K.R. Engineering College, Bhimavaram, India
[3] M.Tech. Student, Dept. of CSE, S.R.K.R. Engineering College, Bhimavaram, India
**Email: sirishagadiraju@gmail.com*

Abstract

Now-a-days people are generally using social networking sites for communicating with the other users and for sharing information across the world. The online social networking sites are becoming the significant tools and are providing a common medium for number of users to communicate with each other. The large amount of information that is accessible on the social networking sites retain the cyber attackers, who generally exploit the information available for their benefits. They generally infect the user's system, appeal the victims to click on malicious links, advertise some products only to gain money. Spam profiles are becoming major security threat used by cyber criminals and also a source of unwanted ads. Twitter is one among several social networking sites which are expanding on daily basis. Spam detection in twitter has become one of the major problems these days. A twitter spam account user nature is analyzed with a target to improve detection of social spam. An innovative technique based on deep learning technology is used for the identification of spam accounts in twitter. These techniques have an advantage that they use raw data to learn high level features on their own, unlike the traditional machine learning algorithms which require native features for the application of classification model.

Keywords: *Cyber criminals; deep learning; malicious links; spam account;*

1. Introduction

All the attractive features of OSNs have made them as everyone's first choice of using the internet. Many people are using online social networking sites to post their messages and share ideas across the world. Several online social networking sites such as Messenger, WhatsApp, and Telegram, LinkedIn are becoming the major important apps in the internet for the users. By the year 2016, all the social media sites are grouped together and it is found that there are more than two billions of users. Therefore the Communication of a user with the other users has peak with this large number of users using OSNs for various purposes. From the business perspective, these platforms are useful for both governmental and commercial agencies by analyzing the with effective ways for marketing, analysis and opinion mining.

Especially, Twitter fosters and attracts the users by providing free micro blogging service for customers, discover various other users of OSN etc. Micro blogging is a service to send message with a character limit of 140. A survey has revealed that in twitter that more than 42 millions of new profiles are generated each month. In the OSNs context, spam could take many forms such as fake profile, malicious links, advertisements, fraudulent reviews, and malicious files. Based on the popularity of the Twitter OSN, unauthorized account users are posting lot of spam, that can be distrustful URLs that generally direct the users to dangerous websites. Therefore twitter spam in twitter had become serious issue and is having a bad network experience for the users. Thus it is identified that 8 percentage of URLs are identified as spam in the data of 2-million-URLs.

Spam is described briefly as or forwarding unwanted messages to the all the remaining users of the OSN. However many of these messages will be inserted with a link a particular website which harms the security of the user account. Even though, in literature, spam messages and spam profiles are handled distinctly, they had a essential link between them. There is a greater probability that spam messages are sent by fake accounts. Usually false or fake advertisements, may mislead the users and can cause inauspicious effects in trading. Actually Spam does not have the ability to differentiate between adult and young, and therefore, there is a risk of showing children to various pornographic links. Additionally other widely existing techniques of spamming is to publish a tweet which should be of the following form "@spam @username" where "@username" is used to tag a account which is spam.

Consequently, both the virtual world and the real world are being impacted by the OSNs. Thus, it is necessary to protect these platforms. To increase users trust in these platforms it requires both regulations and technical controls. It is necessary to decrease or eliminate the existing spam accounts. Spam accounts existence reduces the trustworthiness, performance of the site used. Which leads to exhibit the account user to several issues of security. The main aim of the projects' is to increase the detection of social spam by making use of features that are existing without considering language used by the user for tweets.

2. Related work

McCord & MChuah et al. [16] firstly studied the comparisons between the messages posted by spam account users and valid users. They had identified the features that are used by the traditional algorithms that automatically differentiates the spam ac-

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

count and original accounts. Features are extracted like time at which tweets are posted, length of tweets, mentions and keywords etc. These useful features are used in evaluation of spam account identification which is applied on traditional classifiers such as Support Vector Machine, K-NN neighbor Random Forest, Naïve Bayesian schemes using the collected dataset of twitter. The disadvantage of this method is it is applied only to a small set of data.

Alex Hai-Wang et al. [10] has designed directed graph model which is used to find relationships of the distinct. "follower", "friend" of the data obtained from twitter. Depending on the security policy of Twitter, to provide the detection of spam a different features based on graph and message-based features are designed. A sequence of classification methods such as decision trees, Naive Bayes, neural networks, SVM are applied and are compared to differentiate between normal and suspicious behaviors. The dataset used is Twitter data collected using Twitter API.

Benevento et al. [2] had used tweet content and user based features to identify spammers on twitter. Tweet content features used are - number of characters in a tweet, URLs count of a tweet, numeric character count in a tweet, re-tweet number, ratio of hashtags to number of words in a tweet, word count and hash-tag count in a tweet. Large data containing of more than 45 million users of twitter OSN that had 1064 users and are manually identified and labelled as either spam or non-spam. SVM classifier which is a supervised machine learning scheme are used to identify the difference between spam accounts and. non-spam accounts. The drawback in this method is to apply classification model based on other features such as hashtags' etc.

Grace Gee, Hakson et al. [11] had implemented a learning algorithm that is supervised and it would also be able to label a Twitter account as either spam or a normal user, which is being trained with an initial training set of spammer profiles and the normal user profiles of the twitter. After that we need to collect a huge dataset of account information to correctly train the algorithm for the further classification process. Therefore Gee et al made use of this feature and identified spam accounts using several classification algorithms. They have made use of this twitter feature and collected the normal user data from API and spam accounts data is collected from "@spam" in the twitter dat. Firstly a simple Naïve Bayes algorithm is implemented, and a good initial test is being performed of how good a learning algorithm can identify spammer profiles on twitter. And finally, after the data analysis, a more complex algorithm is being applied. The SVM technique is being implemented, further improve model. The disadvantage of the method is feature engineering is more complex in model.

Amit A. Amleshwaram et al. [1] had developed a method which is a hybrid-approach that contemplates both the domain/URL centric and user properties. They have made a new fifteen features then combine them four features that are proposed. They had applied this method on data of several users with 5 recent tweets and has detected that more than 90% of them are spammers.

Ayon Chakraborty et al. [3]proposed methods for protecting authentic users from abusive users. An algorithm named SPAM (Social Profile Abuse Monitoring) was proposed which helps social media users to identify abusive users. Applied classification algorithms such as random forest, SVM, decision tree and trained over 200 recent posts of 5000 twitter users.

Yang et al. [19] had identified the the new features for the identification of spammers that is spam accounts in Twitter. Finally a total of 18 features (10 new features and 8 existing features) are used for the identification purpose and they are tested using various classification algorithms like Bayesian Network, Random Forest, Decision Tree. Among all these classifier Bayesian classifier performs well with 88.6% of accuracy. Disadvantage of this method is that it is applied on very less data, with less detection rate and only a particular type of spam accounts are being identified.

C. Chen et al.[5] had developed a spam identification method based on machine learning, performed an evaluation by comparing six algorithms such as K-NN Random Forest, Naive Bayes, and SVM. 600 million public tweets were collected out of which 6.5 million are labelled as spam and 6 million are labelled as non-spam. A web service named Trend Micro's Web Reputation Service is used for labelling. They had performed a huge number of experiments for detecting Twitter spam accounts using dataset collected from twitter API. If features are discretized before classification the classifier performance will be further more improved. Further, they had identified that classifiers are able to identify spam tweets at higher rates. The spam tweets were sent at high frequency when compared to normal tweets which are generally randomly tweeted by user.

G.Stringing et al. [18] had proposed a set of accounts and named them as honey net accounts (honey-profiles). They had created honey net accounts of major social networking sites, and recorded all the activities and then these accounts are being observed. Finaly a tool is built to identify spammers, and applied it on Facebook and Twitter dataset.

J Martines-Romeo et al. [14] has proposed two techniques for detecting spam. First is indentifying spam without statistical analysis, and the second is detecting spam by statistical analysis of language. Using the second technique they could detect spam in trending topics. The proposed methods were tested on 20 million tweets corresponding to 34K trending topics. In addition to original features, orthogonal features are used to identify trending topics. Extensive experimentation has shown that F-measure of the system is same as state-of-the-art spam detection systems.

J. Song et al. [17] developed a system which identifies spam on twitter based on filtering malicious and normal accounts. The messages are classified as either unwanted or normal messages by the identification of the sender and keeping track of the sender activities. They their experiments are being conducted on Twitter data. And two relation features, are being proposed which are distance and connectivity, to facilitate the identification of spam. Real time spam detection allows to filter spam while they are being delivered itself.

S. Lee, J. Kim et al. [13] have presented a system called WARNINGBIRD that detects spam based on correlations of URL redirect chains. It considers the direction of redirect chains for spam detection. They had identified features for malicious URLs: Where many are variations of previously identified features and few are newly identified. The disadvantage of this method is it is not applicable for dynamic data.

Arushi Gupta et al. [12] extracted the normalized feature vector from preprocessed Twitter data. These normalized features are in turn used to identify the activities of spammers. The key features extracted from the data set are Spam Words, URLs, Follower/Followee, Replies and Hash Tags. The obtained features are then used learning algorithms such as Naive Bayes, and Clustering are applied features to identify spam accounts. Therefore to improve detection of spam, a method is proposed which "combines" the advantages of the learning (Clustering, Naive Bayes and Decision Trees).

Buket Erahin et al. [8] had proposed an approach for identifying the fake accounts on Twitter social networking site. The aim of the approach proposed by them is to show the effects of the application of discretization on Naïve Bayes classification algorithm. Entropy Minimization Discretization technique is used to discretize the data with Minimal Description Length principle as stopping criteria. Several experiments have been conducted by applying this techniques and the accuracy has been increased with Naïve Bayes classification algorithm from 85.55% to 90.41% by only preprocessing our dataset using discretization technique on selected features of the data. By using discretization, the problem of normality assumption of continuous data of social media datasets used by Naïve Bayes classification algorithm is being over-comed.

Malik Mateen et al.[15] have evaluated data based and features based graphs for identification of malicious accounts. They used a data set which contains more than 12 millions of tweets. The results define that by the application of approach which hybrid precision and recall is being improved significantly. They had developed a method that used user, graph features for detection of spam.

Chao Chen et al.[4] developed a big ground-truth of data for the malicious tweet identification. Several data related factors, such as data size used for training, and data sampling, spam to non-spam ratio, performance etc., are being used. Twelve features for spam detection and identified that feature discretization is important factor for the spam detection performance. The features are mainly varied in time.

Manuel et al. [7] has presented COMPA which, is the very first spam identification system developed to identify the malicious social networking accounts. They identified that COMPA can detect compromises that affect high profile accounts very reliably. As these accounts behavior is very consistent, the false positives are minimal. Spam accounts that violate their accounts' behavioral profile are identified by clustering the messages and applying COMPA to it. By applying COMPA on to two datasets from popular social networks, such as Facebook and Twitter, it shows that their system will be able to detect hundreds of thousands of compromised accounts.

Asim Eshraqi et al. [18] had proposed an algorithm called as DenStream algorithm**.** After preprocessing the tweets, they are clustered to detect spam and real tweets. The features of spam tweets are identified by first identifying one r more spam accounts and all tweets of those accounts are taken as spam tweets.

Arpna Dhingra et al. [6] In their work, tweets are being classified into different forms as spam and fam. This method had first extracted the live tweets and then preprocessed them to improve the tweets in a similar way. Then, the extracted features and weights are stored in a particular file. After performing this process the MLP classification, accuracy, precision and recall values are calculated that explains how the spam accounts are identified accurately.

Table 1: Outline of the Related Work Including Methodologies Used, Categories, Features based and Accuracies

Title of the paper	Features based on	Methodology used	Accuracy (%)
"Spam Detection on Twitter Using Traditional Classifiers" [16]	Hybrid	Random Forest, Naïve Bayesian, Decision Tree algorithm	95.7%
"Don't follow me: Spam detection in Twitter"[10]	Hybrid	Naïve Bayesian, Decision Tree, SVM, Neural Network,	93.5%
"Detecting spammers on twitter"[2]	Hybrid	SVM	87.6%
"Twitter Spammer Profile Detection"[11]	Account	Naive Bayesian, SVM	89.6%
"CATS: Characterizing Automation of Twitter Spammers" [1]	Hybrid	Decision Tree, Random Forest ,Naive Bayes,	93.6%
"SPAM: A Framework for Social Profile Abuse Monitoring"[3]	Hybrid	Decision Tree, SVM, Random Forest, Naive Bayesian,	89%
" Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers" [19]	24 Detection Features		
"6 Million Spam Tweets: A Large Ground Truth for Timely Twitter Spam Detection" [5]	Hybrid	Bayes Network, k-NN, Decision Tree algorithm, Random Forest	TPR with 90%
"Detecting Spammers on Social Network"[18]	Hybrid	Random Forest	90.93%
"Detecting Malicious tweets in tre using statistical analysis of language" [14]	Tweet	SVM, Decision Tree, Naïve Bayesian	94.5%
"Spam Filtering Spam Filtering in Twitter using Sender - Receiver Relation ship [17]	Graph	Decision Tree, Bayes Network, Bagging, Lib SVM, FT	99.7%
"Warning Bird: A Near Real-Time Detection System for Suspicious URLs in Twitter Stream "[13]	Tweet	LIBLINEAR	0.9028
"Improving spam detection in online social networks"[12]	Tweet and user data	Decision trees Naive Bayes and Clustering	87.9%
"Twitter Fake Account Detection"[8]	Tweet	Entropy Minimization Discretization (EMD) , Naïve Bayes	85.55% to 90.41%
"A Hybrid Approach for Spam Detection for Twitter"[15]	Hybrid	J48, Decorate and Naive-Bayes	92%
"A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection"[4]	Hybrid	Bayes network, SVM, Naive Bayes	
"Towards Detecting Compromised Accounts on Social Networks"[7]	Hybrid	COMPA	
"Detecting Spam Tweets In Twitter Using a Data Stream Clustering Algorithm "[9]	Hybrid	Data stream clustering algorithm	89 %
"Content Based Spam Classification in Twitter using Multilayer Perceptron Learning "[6]	Hybrid	Multilayer Perceptron Learning	82%

3. Types of spam detection techniques

Based on several papers, spam detection techniques are broadly classified into following four categories based on their different types of features:(1) User-based features, (2) Message-based features, (3) Hybrid features (4) Tweet sender receiver relationship.

3.1. Account-based features

Generally Spammers can be identified by analyzing the information which the user enters during the registration in the social networking site. Several characteristics. As certain features such as homepage, biography, creation date and location, are controlled by account user, that are not necessary in terms of identification of spam.

The following are the facts observed when the mode of spam accounts are checked by using of user-based features:

1) The number of likes for spam account users is generally low and also spammers use links and hash tags to attract other users.
2) The frequency of tweets by spam accounts is high and spam users also follow a large number of legitimate users to attract them.
3) To make any tweet popular the spam accounts repost the same tweet or similar tweet. These spam accounts will be controlled by one or more users.

Table 2: Outline of Account-based Features

Features Used	Description of the features	Is User-controlled
Profile picture	Profile picture	✓
Homepage	Website homepage of the account	✓
Color	Accounts color theme	✓
Header picture	Header picture of the account	✓
Tweets count	Total count of tweets the account has	×
Username	Unique identifier of the account	✓
Following Count	Number of accounts an account is following	×
Biography	Biography	✓
Location	Location	✓
Followers Count	Total number of followers	×
Retweets Count	Total number of retweets	×
Moments count	Number of times a account has been tagged	✓
Likes count	Count of number of likes a tweets had received	×
Moments count	Number of times a account has been tagged	✓
Likes count	Count of number of likes a tweets had received	×

3.2. Tweet-based Features

Usually spam account users lean to post huge number of unwanted tweets to the normal users to grab the concentration. Spam accounts also be identified by evaluating the content of their messages.

The following are the facts observed when the mode of spam accounts are checked by using of message-based features:

- The likes count and the number re-tweets their tweets have received is very low when compared to legitimate users. Tweets of spam accounts generally use lots of hash tags to reach more number of users.

3.3. Hybrid based techniques

Using the user based techniques for identification of spam accounts is not effective as the user based attributes can be easily falsified by the spammer. In the same way, using content based features for spam account detection in twitter has its own limitations as the spam tweets contains only URLs or few words. Thus there are several techniques which came up that uses both based on user information and message based features to identify the fake accounts.

3.4. Features based on graph

The tweets and user relationship in twitter network can be represented using a graph. Tweets and users correspond to nodes. The relationship between nodes are shown using links. The twitter user and his tweets are connected using relationship.

The following are the facts observed when the mode of spam accounts are checked by using of graph-based features:

- Features based on graph mostly provide the robust performance for the identification of spam accounts.
- Connectivity and distance are the graph based features that best helps in distinguishing legitimate user from spammer.

Table 3: Outline of Message Based Features

Features Used	Description of the feature	Is User-controlled
Hash tag	Hash tag(s) used	✓
Links	URL in the message	✓
Replies Count	It represents the replies a particular message has gained	×
Retweet Count	How many times a tweet is retweeted	×
Sender	Sender	✓
Likes count	Count of likes a tweet has gained	×
Location	The place where the tweet is posted	✓
Mentions	Tagging of users in a tweet	✓

Table 4: Features Based on Graph

Features that are Used	Definition of the feature	Is it controlled by user or not
Connectivity	Connectivity is strength of connection	×
Distance	The shortest path between any two users is taken as distance	×

References

[1] Amleshwaram AA, Reddy N, Yadav S, Gu G & Yang C, "CATS: Characterizing Automation of Twitter Spammers", *Proceedings of 5th International Conference on Communication Systems and Networks (COMSNETS),* (2013), pp.1-10, available online: http://dx.doi.org/10.1109/COMSNETS.2013.6465541

[2] Benevenuto F, Magno G, Rodrigues T & Almeida V, "Detecting Spammers on Twitter*", Proceedings of CEAS 2010 Seventh annual Collaboration, Electronic messaging, Anti Abuse and Spam Conference*, (2010), pp.1-9

[3] Chakraborty A, Sundi J & Satapathy S, "SPAM: A Framework for Social Profile Abuse Monitoring".*Technical Report,* Dept.of Computer Science, Stony Brook University,Stony Brook, NY 11794-4400, USA, (2012)

[4] Chen C, Xie Y & Xiang Y, "A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection", *IEEE Transactions on Computational Social Systems,* Vol.2, No.3, (2015), pp. 65-76

[5] Chen C, Zhang J, Chen X, Xiang Y & Zhou W, "6 Million Spam.Tweets: A Large Ground Truth for Timely Twitter Spam Detection", *Proceedings of IEEE International Conference on Communications,* (2015), pp.7065-7070

[6] Dhingra A & Mittal S, "Content Based Spam Classification In Twitter using MultiLayer Perceptron Learning", *International Journal of Latest Trends in Engineering and Technology* ,Vol. 5,

No.4, (2016), pp.1-11, available online: https://pdfs.semanticscholar.org/2ef0/3ba493f5d4c8a5dfd9c62bcd6abd81a5c9de.pdf, last visit:10.03.2018

[7] Egele M, Stringhini G, Kruegel C & Vigna GV, "Towards Detecting Compromised Accounts on Social Networks", *IEEE transactions on Dependable Secure Computing*, Vol.: 14, No.4, (2017), pp.447-460

[8] Erahin B, Aktas O, Kilinc D & Akyol C, "Twitter Fake Account Detection", *Proceedings of IEEE International Conference on Computer Science and Engineering* , (2017), pp.388-392. doi: 10.1109/UBMK.2017.8093420

[9] Eshraqi N, Jalali M & Moattar MH, "Detecting Spam Tweets In Twitter Using a Data Stream Clustering Algorithm", *Proceedings of Second International Congress on Technology, Communication and Knowledge,* (2015), pp.347-351, doi: 10.1109/ICTCK.2015.7582694

[10] Fauci AS & Wang AH, "Don't Follow Me: Spam Detection in Twitter", *Proceedings of IEEE 2010 International Conference on Security and Cryptography (SECRYPT),* (2010), pp.1-10

[11] Gee G & The H, "Twitter Spammer Profile Detection",(2010), available online: http://cs229.stanford.edu/proj2010/GeeTeh-TwitterSpammerProfileDetection.pdf

[12] Gupta A & Kaushal R," Improving Spam Detection in Online Social Networks" *Proceedings of IEEE International Conference on Cognitive Computing and Information Processing*, (2015), pp. 1-6, doi: 10.1109/CCIP.2015.7100738

[13] Lee S & Kim J,"Warning Bird: A Near Real-Time Detection System for Suspicious URLs in Twitter Stream", *IEEE Transaction Dependable Secure Computing* 10 (2013), pp.183-195

[14] Martinez-Romo J & Araujo L, "Detecting malicious tweets in trending topics using a statistical analysis of language", *Journal of* Expert Systems Applications, Vol. 40, (2013), pp.2992–3000. doi:10.1016/j.eswa.2012.12.015

[15] Mateen M, Iqbal MA, Aleem M & Islam A, "A Hybrid Approach for Spam Detection for Twitter" *Proceedings of IEEE 14th International Bhurban conference on Applied Sciences and Technology,* (2017), pp.466-471

[16] McCord M & Chuah M, "Spam detection on twitter using traditional classifiers," *Proceedings of 8th International conference on Autonomic and Trusted Computing*, Springer, (2011), pp.175-186

[17] Song J, Lee S & Kim J, "Spam Filtering in Twitter using Sender-Receiver Relationship", *Proceedings of 14th International Conference on Recent Trends in Intrusion Detection,* (2011), pp.301-317

[18] Stringhini G, Kruegel C & Vigna G," Detecting Spammers on Social Networks", *Proceedings of 26th Annual Computer Security Application Conference*, (2010),pp.1-9

[19] Yang C, Harkreader R & Gu G,"Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers" *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 8, , (2013), pp. 1280-1293 available online:, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6553246&isnumber=6552239