Asian Journal of Information Technology 15 (7): 1253-1262, 2016
ISSN: 1682-3915

Spam Profile Detection in Online Social Network Using Statistical Approach

[1]C. Emilin Shyni, [2]Anesh D. Sundar and [3]G.S. Edwin Ebby
[1]Department of Information Technology, KCG College of Technology,
[2]Department of DOVE, Karunya University, Coimbatore,
[3]Wipro Technologies, Chennai, Tamil Nadu, India

Abstract: Online Social networks, popularly known as OSN are widely used by millions of people around the world to communicate with friends and relatives. Information sharing is done by sending links to videos, websites and files. The community structure of the online social networks help in building a network of trust which is exploited by spammers who spread spam messages that promote personal blogs, advertisements, phishing and scam. Spamming is the method of sending unsolicited bulk messages especially advertisements, indiscriminately. Two of the most popularly used OSN around the world are Facebook and Twitter. This research focuses on detecting spam profiles on the given set of Twitter profiles. Current studies identify spam profiles based on a set of 11 features. This study identifies spam profiles based on an enriched set of 17 features. The extracted features are given as input to classification algorithms and the accuracy of these different algorithms are analysed. Based on this, the best classification algorithm for Twitter has been identified.

Key words: Online social network, spam, twitter, classifier, India

INTRODUCTION

Online social networks have been widely used by people around the world to communicate with friends and relatives. The basic elements in the hierarchy of social networks are the individual users. The next in hierarchy are the communities formed by the friends, families and acquaintances. Online social networks have succeeded in building a network of trust. This trust has been exploited by spammers who spread spam messages which promote personal blogs, advertisements, phishing and scam. A user is more likely to respond to a Twitter follower's message than a message from a stranger.

Spamming is the method of sending unsolicited bulk messages especially advertisements, indiscriminately. Information sharing through URL (Uniform Resource Locator) shortening service is an important feature of OSN. Twitter plays a very important role in today's online social networking scenario. It has about a billion registered users. Over 500 million tweets are sent on an average every day. This study mainly concentrates on Twitter. In this study, a statistical approach is provided in order to identify spam profiles in Twitter. This research identifies a set of 17 features which help in the detection of malicious profiles. The features thus extracted were fed into classification algorithms which provide detection rates and false positive rates. An experiment was conducted which involves feeding the entire feature set to train the classifiers and testing the accuracy of their classification using 10 folds cross validation. The VFI (Voting Feature Intervals) classification yielded a detection rate of 99.5% for twitter.

Literature review: There are so many different approaches employed by existing systems to detect spam in online social networking sites. The aim of all the detection techniques is to increase the detection accuracy and to reduce the false positive rates. In the researc, (Thomas *et al*., 2011) a real time URL-spam detection scheme (Monarch) for Twitter was proposed by the researcher in which the browser activities while loading a page specified by a URL were logged. Yet another research by Lee and Kim (2013) presents a suspicious URL detection system in Twitter stream. Another significant work on detection of spam in OSNs is presented by Stringhini *et al*. (2010) In this research, honey-profiles were created representing different age, nationality, etc. Based on the activities in Facebook, MySpace and Twitter, six features were developed to differentiate spam profiles from normal profiles. The researcher created social honeypot to attract spam spreaders on Twitter, whose profiles are analyzed to

Corresponding Author: C. Emilin Shyni, Department of Information Technology, KCG College of Technology, Chennai, Tamil Nadu, India

identify a set of features for classification purpose. Random forest classifier was used as classification algorithm which yielded 96.75% accuracy. Similarly, the researchers Yang *et al.* (2013) did a thorough analysis of profile-based and content-based evasive tactics that are employed by spammers in Twitter. A set of 24 features were proposed which were evaluated using different machine learning techniques. The Detection Rate by using new feature set increased to 85%. The researcher, Rebecca (2012) have analyzed tweets and the elements like user name, colour scheme, background and profile picture/avatar. Jin *et al.* (2011) proposed SocialSpamGuard which was a scalable and online spam detection system in social media. Jones *et al.* (2013) proposed a welsh-speaking Twitter and expressed through the non-symmetric "follow" relationship. In the research (McCord and Chuah, 2011), a total of 6 user-based and content based features were identified to separate spam and benign accounts. The results show that their spam detection system has a 95.7% precision and 95.7% F-measure using the random forest classifier. Miller *et al.* (2014) had identified features based on the categories Content and tweet. In the research Lee and Kim (2014) proposed a scheme that detects malicious Twitter accounts at the time of their creation without waiting for the initiation of malicious behaviors.

Most of the earlier approaches have dealt with tweet level detection of spam in Twitter social network. Moreover limited statistical features were considered for spam detection. This study does a profile level detection by using an enriched set of 17 features which help in identifying the malicious nature of a Twitter profile. These features can be practically used in order to identify whether a Twitter profile is malicious and block the same if it is so and thus stop the spam propagation in Twitter network.

MATERIALS AND METHODS

Twitter data and characteristics: A set of profiles were collected from Twitter social network which included manually classified benign and spam profiles. A total of 405 profiles were taken from Twitter of which 211 profiles were benign and 194 profiles were malicious. These accounts were the main source of our data. The maximum number of tweets that were considered for analysis from a single profile is 3200 even if the user has tweeted more than this count. Only public profiles were used for collecting the data and performing the analysis. The main characteristics of Twitter from which specific features can be identified and collected include: Tweets: These are messages using which information can be disseminated by sharing a link or writing a message not >140 characters. @mentions: This feature is used to address someone. Hash-tags: popular topics that are discussed in tweets are called as hash-tags. The topic is preceded by a hash (#) and hence, it gets the name. URLs: URLs can be shared in tweets. These features are discussed in detail in the following study.

Identified features: The main characteristics of Twitter lead to some of the specific features that can help in branding a profile as malicious or benign. These features which help in characterizing a profile were identified in Twitter network and their details were crawled using the HTML parser. The Twitter features are listed and discussed and these values were extracted from all the Twitter profiles that were collected. These features thus extracted were used to identify whether an account is benign or spam. The statistical features have different kinds of values which typically indicate whether the profile is spam or not.

Feature categorization: The features are broadly classified into Interaction related features, Tweets related features, URLs related features, tags/@ mention related features and age related features. These features are listed in Table 1 where features F1, F3, F4, F5, F6, F9, F10, F11, F13, F14 and F15 were used by Ahmed and Abulaish (2013) and the remaining six features are novel features proposed in this study. Interaction related features introduced in this study are Follower of Following (FoFo) ratio (F2). FoFo ratio provides knowledge about the friendship details of the users with users as well as the popularity of the user with others in Twitter. A large number of following and a small number of followers highlights the suspicious aspect of the account. Tweets related features in this study are API count (F7) and Tweeting rate (F8). API count is the number of tweets with the tweet source of "API" to the total number of tweet count. Since increased number of API implies that the account is more suspicious (Chu *et al.*, 2012), count of API is taken into consideration. Tweeting rate is the ratio of the number of tweets to the age of the account. As high tweeting rate symbolizes malicious user, tweeting rate is captured and used.

URL related features introduced in this study are API URL ratio (F12). It is more convenient for spammers to post spam tweets using API, especially when spammers need to manage a large amount of accounts. Thus, a higher API URL ratio of an account implies that this account's tweets sent from API are most likely to contain

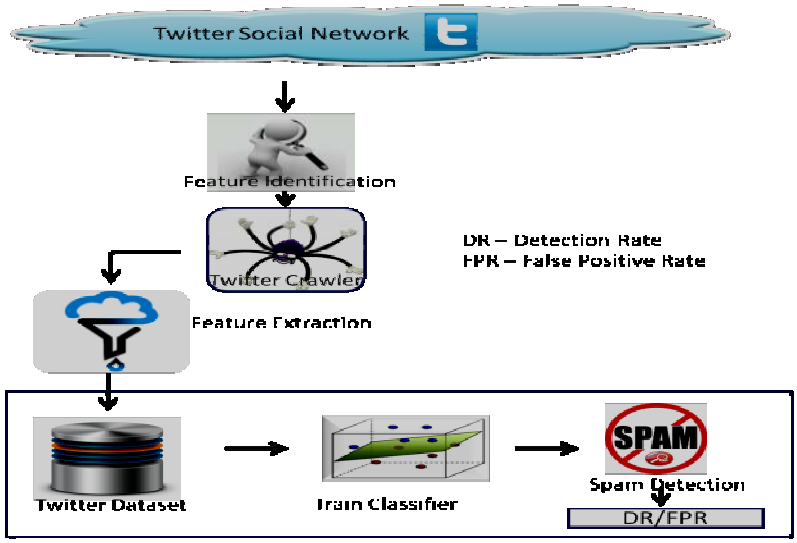

Fig. 1: Architecture for malicious profile detection

```
while(character position < length of the complete profile)
        Character parsed = character at the current position
        if  (Character parsed  is '#')
                while(character position < length of the complete profile)
            Concatenate character to String1
                        Extract character at next position
                        if (Character is Whitespace)
                                if(length of String1 > 1 )
                                        if ((Character in the first position of String1!='!' and '>'))
                                                if (Substring within String1 contains("http"))
                                                        Remove "http" from String1
                                                Increment the count of hash tags present in the
                                                Twitter profile
                                                        Write the hash tag found in the output file
                                Initialize String1
Check the next character position and Store the count of hash tags in a separate file.
```

Fig. 2: Sample pseudo code for twitter crawler

URLs, making this account more suspicious. Age related features in this study are age of the account (F16) and following rate (F17). Age of the account is the number of months that the user has been attached to Twitter. Most of the spammers create the twitter for short span of time. Following rate reflects the speed at which an account follows other accounts. These newly introduced features and the previous 11 features are increasing the accuracy of the spam detection.

Architecture for malicious profile detection in Twitter: Classification of a profile as spam or not involves several steps that must be executed before one can confirm of its malicious nature. The architectural design of the work done in the twitter social network is given in Fig 1. gives an insight into the actual sequence of steps executed for the identification of malicious profiles. It starts with the identification of the features from the twitter network, which would help in the detection of malicious profiles. From the twitter public directory, various public profiles were randomly considered. This constituted the majority of benign profiles that were used for the purpose of producing the dataset that would train the classifier. A carefully examined and manually collected set of spam profiles were also used to produce the dataset. Subsequent to the identification of the profiles, a Twitter crawler was written which would crawl the data required from Twitter profiles. The Twitter crawler is further explained in the study

Twitter crawler: A Twitter crawler is a Java code which is used to parse a Twitter profile and to extract the data that is required from the profile. The crawler parses every character in the profile and checks for specific character or word or sequence of characters. Based on the presence or absence of the characters or word checked, certain actions are taken and decisions made. A sample pseudo code of the crawler is provided in Fig. 2. Using this

Table 1: List of features

Feature number	Feature name	Explanation
Interaction related features		
F1	No. of followers/No. of following	This indicates the number of followers for the account as well as the count of people who are following this account
F2	FoFo ratio	Ratio of the number of following to the number of followers. A high FoFo ratio is an indication that the profile is malicious
F3	Total Number of Hashtags	This feature provides the total number of hash tags or popular topics that are discussed in the profile
Tweets related features		
F4	Unique number of hash tag	This is the number of unique hash tags present in a profile.
F5	Maximum frequency of hash tags	This is the maximum of the frequency of hash tags that have been used in the profile
F6	Average frequency of hash tags	This is the average value of the frequency of hash tags that are used in a profile. The frequency of every hash tag is found out and the average value of all these frequencies taken
F7	API Count	Spammers usually use API to post tweets automatically as it is very difficult for them to manually handle multiple accounts. This count is the number of tweets with the tweet source of API or the total number of tweets posted by the profile using API
F8	Tweeting rate	This is the ratio of total number of tweets in a profile to the age of the profile. The age of the profile has to be computed and the total number of tweets that are posted by the profile. Based on this, the ratio is calculated
URL related features		
F9	Total number of URLs shared by a user	This is the total count of URLs that are present in a profile.
F10	Total number of unique URLs	This is the count of the total number of unique URLs that are present in a profile
F11	Average frequency of the URLs	The frequency of every URL that is posted by an account is considered and their average is taken
F12	API URL ratio	This is the ratio of the number of tweets that are posted by an API which contains a URL to the total count of tweets posted by API
@mention related features		
F13	@mention count	This is the total number of @mentions that appears in a user's account
F14	Unique @mentions	This is the number of unique user names that are @mentioned in a user's profile
F15	@mention rate	This is the average number of @mentions that is used by a user in his account. To find this, the frequency of every name that is @mentioned is taken and the average of these frequencies is found out
Age related featuresa		
F16	Age of the account	This is the number of months that the user has been attached to Twitter
F17	Following rate	This is the ratio of the account's following number to the age of the account

crawler, the features that are required can be extracted from the Twitter profile. This constitutes the next step. In the subsequent step, the extracted features are formatted to create twitter dataset which is used to train the classification algorithms for the detection of malicious profiles. These algorithms classify the profiles based on the feature set fed to them as malicious or benign.

Feature set analysis: Feature set forms the basis of classification and hence, this study deals with the analysis of various features that were mentioned in the previous sections. The specific values of these features were examined and for every feature that was analyzed, certain conclusions were reached based on its property which would aid in the classification of the profiles. A detailed analysis of every feature and the contribution of each for the identification of spam content are presented in this study.

Number of followers/following (F1): more number of followers indicates that more people trust this account. Hence this is an indication that the profile is a trustworthy one. Following number indicates how many people this account follows. Usually spammers follow a large number of profiles to gain popularity (Table 1). FoFo ratio (F2): this ratio provides knowledge about the friendship details of the users with others as well as the popularity of the user with the others in Twitter. A large number of following and a small number of followers highlights the suspicious aspect of the account. A high Fofo ratio is an indication that the profile is malicious.

Total Number of Hashtags (F3): The number of hash tags is indicative of the interaction of the profile with a large community. The total number of hash tags will be high for a malicious profile. A spammer will be able to spread spam more easily by spreading the malicious content using a hash tag so that it is visible to a larger community.

Numbers of unique hash tags (F4): Normal or benign users typically use a variety of hash tags whereas spammers use the most popular hash tags as they are seen by most people and it is easier to spread malicious content over there. The number of unique hash tags will be less in the case of malicious profiles as spammers tend to use popular hash tags frequently.

Maximum frequency of hash tags (F5): This value does not provide any information of its own but is useful along with the hash tagging rate or the average of the frequency of hash tags. A large maximum value along with a high average hash tagging rate indicates the malicious nature of a profile.

Average frequency of hash tags (F6): a high average value along with a small number of hash tags highlights the malicious nature of the profile.

API Count (F7): When more number of tweets is generated automatically and a large number of these automated tweets have a URL embedded in them, the profile is likely to be a malicious one. So, this feature along with the API URL ratio (F15) indicates the malicious nature of a profile.

Tweeting rate (F8): A high tweeting rate indicates that the profile is malicious in nature. That is, in a short span of time, a malicious profile posts a large number of tweets in order to spread spam at a great speed.

Total number of URLs shared by a user (F9): malicious profiles tend to share a large number of URLs. Most of these URLs are repetitive in nature as a small number of unique URLs are shared, repetitively by these malicious profiles. A very high value for this feature indicates that the profile is malicious.

Total number of unique URLs (F10): the number of unique URLs will be very small for a malicious profile. This feature along with the total number of URLs shared by a profile models the malicious behavior of the profile. A spammer will have a small number of URLs which spread spam content. These spam URLs will be circulated at large by the malicious profiles.

Average frequency of the URLs (F11): A high value for this feature indicates that a small number of distinct URLs are shared a large number of times which is indicative of the malicious behavior of a profile.

API URL ratio (F12): This ratio indicates as to how many of the automated tweets contain a URL. A high ratio indicates that the profile is malicious. A spammer would spread URLs using automation as he would have multiple accounts to manage and would resort to automation to make his job easier.

The @mention count (F13): A large number of @mentions indicates that the user is interacting with a large number of people many times. Benign users never do this as they communicate only with a small number of people and address only a small number of people. Thus, a large value illustrates that the profile is malicious.

Unique @mentions (F14): This feature along with the @mention rate can indicate the malicious nature of a profile.

The @mention rate (F15): In twitter, a high @mention rate along with a large number of following count or followers count indicates that the profile is malicious. Also, another indicator is an extremely small value of the mention rate.

Age of the account (F16): The age of the account together with the number of tweets of the profile gives the tweeting rate. A high rate indicates that the profile is malicious. Similarly, the age of the account along with the number of following indicates the following rate of a profile. A high following rate also indicates that the profile is malicious in nature.

Following rate (F17): A spammer would follow more accounts in a short span of time to gain popularity. This is indicated by the following rate. This A high following rate indicates that the profile is malicious.

Based on the analysis done, an algorithm was developed which provides the set of steps involved in the detection of malicious profiles. The algorithm has been summarized in

Algorithm for malicious profile detection: This study deals with the set of steps and computations that were followed in order to identify whether a profile is malicious. Compute Fofo ratio. Fofo ratio computation is given by Eq. 1:

$$\text{Fofo ratio} = \text{No. of following}/\text{No. of followers} \quad (1)$$

If Fofo ratio <0.09, then set spam flag. Count the total number of hash tags in the profile. If this count exceeds 785, then set the spam flag. Let n = number of unique hash tags in the profile. Let freq (xi) be the frequency of occurrence of hash tag xi. Now compute maximum number of hash tags in the profile by the formula specified in Eq. 2:

$$\text{Maximum number of hash tags} = (\text{freq}(xi)),\ I \text{ varying from 1 to n} \quad (2)$$

Average value of frequency histogram used in the profile (F6) is computed using Eq. 3:

$$F6 = (\sum \text{hash tags in the profile})/(\sum \text{unique hash tags in the profile}) \quad (3)$$

If ((maximum number of hash tags generated by the profile>63) and (F6>19)), then set the spam flag.

Calculate the total number of @mentions that are used by the account (F13) If F13 exceeds 2000 set the spam flag. Extract the number of unique @mentions in the profile (F14). Compute the number of @mentions per friend (F15). The formula for this is given in Eq. 4:

$$F15 = F13/F14 \quad (4)$$

If ((followers>1000) || (following>1000)), then If ((F15 >100) || (F15<2.5)) ,Set spam flag Extract the total number

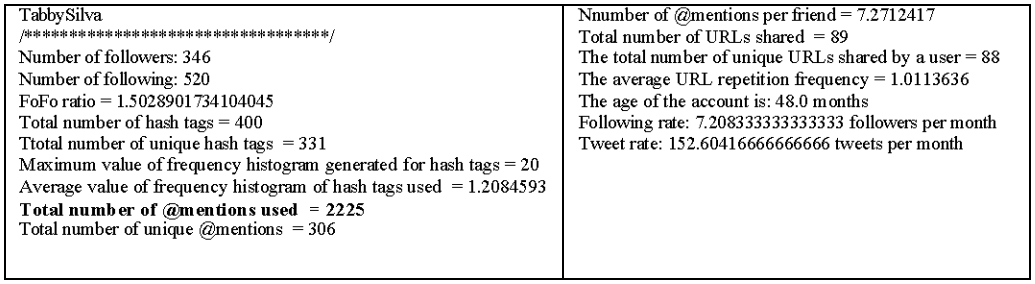

Fig. 3: Real example for the algorithem

of URLs shared by the profile (F9). Extract the total number of unique URLs in the profile (F10). If ((F9>2500) and (F10<30)) then set the spam flag. Compute average URL frequency (F11). This can be computed by the formula specified in Eq. 5:

$$F11 = F9/F10 \quad (5)$$

Find the age of the account, F16. Compute the following rate, F17. This is given by the formula mentioned in Eq. 6:

$$F17 = (\text{Following number})/ F16 \quad (6)$$

If F17>100, set the spam flag Extract the total number of tweets posted by the account. Compute the tweet rate. The formula for its computation is specified in Eq. 7:

$$\text{Tweet rate} = (\text{Tweet count})/F16 \quad (7)$$

If tweet rate>195, set the spam flag. Find the total number of tweets with the tweet source of API (F7). Compute API URL ratio (F12). This can be calculated by the formula given in Eq. 8:

$$F12 = (\text{Number of tweets posted by API containing a URL})/F7 \quad (8)$$

If ((F7>50) and (F12>0.8)), then set the spam flag. If spam flag has been set, then the account is a spam profile. This algorithm was run for every profile. Based on the setting of spam flag, it can be decided whether the profile is spam or benign. The detection is based on statistical features of a profile as can be seen from this algorithm. A real example of Twitter profile classification is shown in Fig. 3. For the sample profile, the classification result is spam profile because the total number of @mentions> 2000.

RESULTS AND DISCUSSION

The complete set of features which were extracted from Twitter were collected and fed into classification algorithms. The classification algorithms that were trained to detect malicious profiles included VFI (Voting Feature Intervals) classifier, ADTree (Alternating Decision Tree) classifier and random committee classifier. Of these, VFI provided the best performance (as far as the detection rates are concerned) with a detection rate of 99.5% and a false positive rate of 5.2%. The ADTree classifier provided a detection rate of 97.9% with a false positive rate of 2.8%. The Random committee classifier yielded a detection rate of 96.4% and a false positive rate of 6.6%. The VFI classifier performs much better than the other classifiers as every feature is independently considered. The precision for VFI classifier was 0.946 while that for ADTree was 0.969 and the same for random committee was 0.93. The precision is greater for ADTree as the false positive rate is the least for the algorithm. Precision is lowest for Random Committee classifier as false positive rate is the highest for this classifier when compared to the other two classifiers. It can also been seen that VFI has a high recall value, 0.995 as it returns most of the relevant results. It is closely followed by ADTree with 0.979 and finally random committee which has a recall value of 0.964. The F-Measure which is the weighted average of precision and recall is the highest for ADTree which is 0.974. F-measure is next highest for VFI classifier for which the value is 0.97. The F-measure is the least for Random Committee as both the precision and recall values are lesser for this classifier when compared to the other two. These results are tabulated in Table 2. Figure 4 shows the graph which compares the detection rates and the false positive rates for all the three classifiers. Figure 5 illustrates differences between the precision, recall and f-measure for all the three classifiers.

Table 2: Analysis of the performance classify

Classifier	Detection rate	False positive rate	Prcision	Recall	F-measure
VFI	0.995	0.052	0.946	0.995	0.970
AD tree	0.979	0.028	0.969	0.979	0.974
Random committee	0.964	0.066	0.093	0.964	0.974

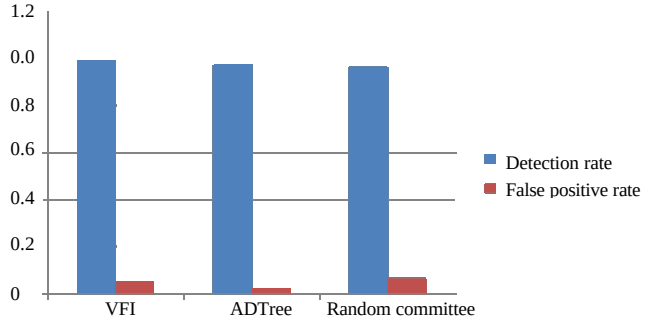

Fig. 4: Comparison of the performance of classifiers

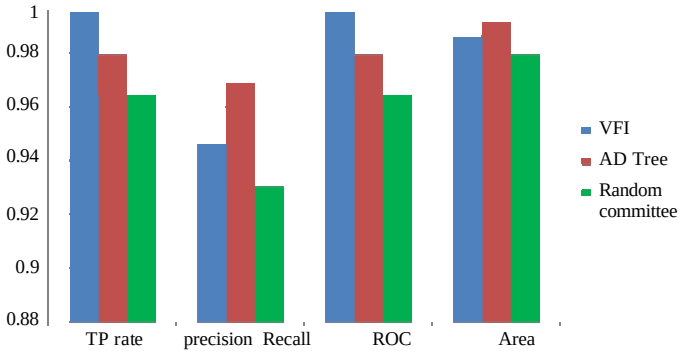

Fig. 5: Graphical comparison of the performance of classifiers

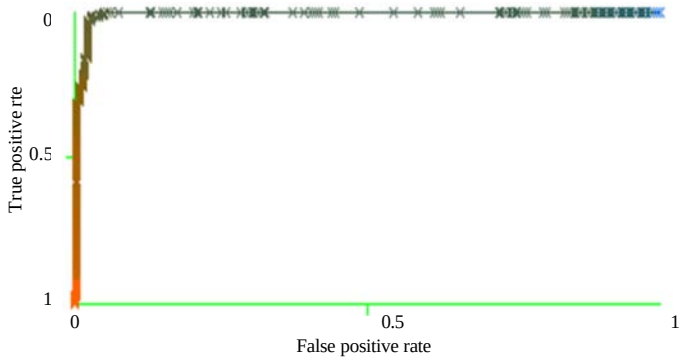

Fig. 6: ROC curve for VFI classifier

The next analysis is based on Receiver Operating Characteristic (ROC) curve. This curve is a representation of the true positive rate against the false positive rate. This indicates the tradeoff between sensitivity and specificity, that is an increase in sensitivity is accompanied by a decrease in specificity. The ROC curve for VFI classifier has been plotted in Fig. 6 and that for ADTree classifier is plotted in Fig. 7 The ROC curve for Random Committee classifier is plotted in Fig. 8. It can be observed that Fig. 7 has more accuracy with the area under the ROC curve being 0.991. The curve for the next classifier which yields better accuracy is Fig. 6 with the area under the ROC curve being 0.986. The least accuracy among the three is portrayed in Fig. 8 with the area under the ROC curve being 0.979. This indicates that as far as the ROC curve is concerned, the ADTree classifier outshines the other two in terms of better accuracy. The ROC curve for ADTree classifier is closer to the y-axis, while it rises above when compared to the other curves, before it turns right. The significance of the results obtained for various classification algorithms have been discussed in detail in the following subsection.

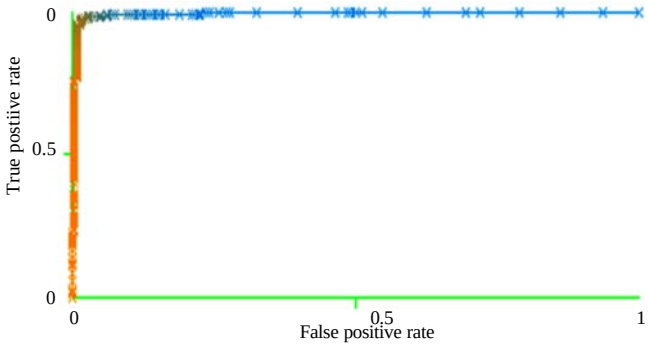

Fig. 7: ROC curve for ADT classifier

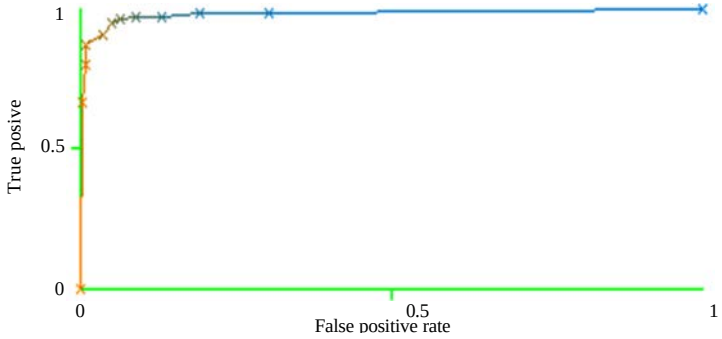

Fig. 8: ROC curve for random committee classifier

Classification algorithms and their impact on the results: Three different classification algorithms were used namely Voting Feature Intervals classifier, Alternating Decision Tree Classifier and Random Committee classifier. A brief description of these algorithms along with their impact on the result will be discussed in the following subsections.

Voting feature intervals classifier: As per the authors, (Demiro'z and Guvenir ,1997) in this classifier, the set of feature intervals is represented on each feature dimension separately. It can be compared with the Naïve Bayesian classifier, in which each of the features are considered separately. VFI performs much better than Naïve Bayesian classifier in terms of Detection Rates. This classifier performs better than the other ones for the proposed malicious profile identification method because, it considers every feature separately and every class is given real valued votes. Based on the votes obtained, the profile is classified as spam or benign. But the False Positive Rate is a bit high when compared to ADTree classifier and hence the area under the ROC curve is slightly lesser than the ADTree classifier.

Adtree classifier: This classifier is otherwise called as alternating tree classifier. There are two nodes in an alternating decision tree. One is the decision nodes and

```
: 0.042
| (1)tweet_rate < 35.879: 1.165
| (1)tweet_rate >= 35.879: -0.726
| | (3)mention_count < 1997.5: 0.47
| | | (6)tweet_rate < 198.043: 0.688
| | | (6)tweet_rate >= 198.043: -1.636
| | (3)mention_count >= 1997.5: -1.817
| (2)fwers < 241.5: 0.684
| (2)fwers >= 241.5: -0.823
| (4)fwing < 1052: 0.122
| | (5)count < 668.5: 0.281
| | | (7)foforatio < 0.115: -1.563
| | | (7)foforatio >= 0.115: 0.488
| | | | (8)tweet_rate < 195.239: 0.378
| | | | | (10)mention_count < 2004.5: 0.882
| | | | | (10)mention_count >= 2004.5: -0.959
| | | | (8)tweet_rate >= 195.239: -1.262
```

Fig 9: Alternating decision tree generated for the given dataset

is specified by the decision nodes while the prediction nodes have a single number. The prediction nodes are present in the ADTree both as leaves and root always. A portion of the alternating decision tree generated for the dataset is shown in Fig. 9. The detection rates, though lesser than the VFI classifier, the false positive rates are also less, thereby leading to more area under the ROC curve. Hence, the accuracy is more for this classifier.

```
Follrate < 10.35
|  tweet_rate < 0.02 : spam (1/0)
|  tweet_rate >= 0.02
|  |  Avg1 < 0.5
|  |  |  hashtag_freq_max < 0.5 : benign (9/0)
|  |  |  hashtag_freq_max >= 0.5
|  |  |  |  fwing < 1013.5 : benign (3/0)
|  |  |  |  fwing >= 1013.5 : spam (1/0)
|  |  Avg1 >= 0.5 : benign (138/0)
Follrate >= 10.35
|  fwers < 523.5
|  |  count3 < 99.5 : benign (7/0)
|  |  count3 >= 99.5
|  |  |  fwers < 274 : benign (1/0)
|  |  |  fwers >= 274 : spam (1/0)
|  fwers >= 523.5 : spam (6/0)
```

Fig. 10: Random tree generated during classification

Random committee classifier: The random committee classifier creates an ensemble of various base classifiers, finds out their predictions and averages them. A portion of the random tree that was generated during the classification is given in Fig. 10. When compared to the other two classifiers, the Random Committee classifier has comparatively lesser detection rate and a higher false positive rate. Thus, the other two classifiers outperform when compared to this classifier for the collected dataset.

In addition to those classifiers, experiments were conducted with three more classifiers, namely, Naïve Bayes, Jrip and J48. Decision rates of 97.3, 96.2 and 96.3 are obtained respectively. Any of these classifiers can be selected for classification the profile as benign or malicious.

The decision rate ranges from 96.3-99.5. The rich feature set used for classification in this research, which includes the newly proposed six features discussed in detail in study 3 results in better classification achieving atleast 96.3 decision rates.

CONCLUSION

In this study, we have proposed a set of 17 features in order to identify malicious profiles in Twitter. A large dataset of 405 profiles were used, of which 211 were benign and 194 were malicious ones. We have also analysed the accuracy of our detection algorithm by feeding the feature set extracted from the profiles into various classification algorithms. We have compared the efficiency of these algorithms in terms of their detection rates and accuracy based on the false positive rates. The algorithms that were considered for the classification were VFI, ADTree and Random Committee. Out of these, best results were obtained for VFI classification in terms of detection rates.

From accuracy point of view, best result was achieved by ADTree classifier. The detected spam profiles can be blacklisted. Thus spam profiles can be blacklisted and removed hereby preventing harm to other accounts. From these results, we intend to perform cluster analysis for detecting various spam campaigns in Twitter network. Also, we propose to extend this research for Facebook as well.

REFERENCES

Ahmed, F. and M. Abulaish, 2013. A generic statistical approach for spam detection in Online Social Networks. Comput. Commun., 36: 1120-1129.

Chu, Z., S. Gianvecchio, H. Wang and S. Jajodia, 2012. Detecting automation of twitter accounts: Are you a human, bot or cyborg?. Dependable Secure Comput. IEEE. Trans., 9: 811-824.

Demiroz, G. and H.A. Guvenir, 1997. Classification by voting feature intervals. Proceedings of the 9th European Conference on Machine Learning: ECML-97, April 23-25, 1997, Springer Berlin Heidelberg, Ankara, Turkey, ISBN: 978-3-540-68708-5, pp: 85-92.

Jin, X., C. Lin, J. Luo and J. Han, 2011. A data mining-based spam detection system for social media networks. Proc. VLDB. Endowment, 4: 1458-1461.

Jones, R.J., D. Cunliffe and Z.R. Honeycutt, 2013. Twitter and the Welsh language. J. Multilingual Multicultural Dev., 34: 653-671.

Lee, S. and J. Kim, 2013. Warningbird: A near real-time detection system for suspicious urls in twitter stream. Dependable Secure Comput. IEEE. Trans., 10: 183-195.

Lee, S. and J. Kim, 2014. Early filtering of ephemeral malicious accounts on Twitter. Comput. Commun., 54: 48-57.

McCord, M. and M. Chuah, 2011. Spam Detection on Twitter Using Traditional Classifiers. In: Autonomic and Trusted Computing, Calero, J.M.A., L.T. Yang, F.G. Marmol, L.J.G. Villalba, A.X. Li and Y. Wang (Eds.). Springer-Verlag GmbH, Berlin, pp: 175-186.

Miller, Z., B. Dickinson, W. Deitrick, W. Hu and A.H. Wang, 2014. Twitter spammer detection using data stream clustering. Inf. Sci., 260: 64-73.

Rebecca, W., 2012. Newspaper Twitter: Applied drama and microblogging. Res. Drama Educ. J. Appl. Theatre Perform., 17: 569-581.

Stringhini, G., C. Kruegel and G. Vigna, 2010. Detecting spammers on social networks. Proceedings of the 26th Annual Conference on Computer Security Applications, December 06-10, 2010, ACM, New York, USA., ISBN: 978-1-4503-0133-6, pp: 1-9.

Thomas, K., C. Grier and J. Ma, 2011. Design and evaluation of a real-time url spam filtering service. Proceedings of the IEEE Symposium on Security and Privacy, May 22-25, 2011, Berleley, California, USA., pp: 16-31.

Yang, C., R. Harkreader and G. Gu, 2013. Empirical evaluation and new design for fighting evolving Twitter spammers. Inf. Forensics Secur. IEEE. Trans., 8: 1280-1293.