Malicious Accounts Detection based on Short URLs in Twitter

Rasula Venkatesh

Roll. 213CS2174

under the guidance of

Prof. Sanjay Kumar Jena

Department of Computer Science and Engineering

National Institute of Technology Rourkela

Rourkela – 769 008, India

Malicious Accounts Detection based on Short URLs in Twitter

Dissertation submitted in

June 2015

to the department of

Computer Science and Engineering

of

National Institute of Technology Rourkela

in partial fulfillment of the requirements

for the degree of

Master of Technology

by

Rasula Venkatesh

(Roll. 213CS2174)

under the supervision of

Prof. Sanjay Kumar Jena

Department of Computer Science and Engineering

National Institute of Technology Rourkela

Rourkela – 769 008, India

Department of Computer Science & Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India. www.nitrkl.ac.in

Declaration by Student

I certify that

- I have complied with all the benchmark and criteria set by NIT Rourkela Ethical code of conduct.
- The work done in this project is carried out by me under the supervision of my mentor.
- This project has not been submitted to any other institute other than NIT Rourkela.
- I have given due credit and references for any figure, data, table which was being used to carry out this project.

Place: NIT,Rourkela-769008 **Rasula Venkatesh**

Date:01/06/2015

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India.

Certificate

This is to certify that the work in the thesis entitled " ***Malicious accounts detection based on short URLs in Twitter***" submitted by ***Rasula Venkatesh*** is a record of an original research work carried out by him under our supervision and guidance in partial fulfillment of the requirements for the award of the degree of Master of Technology in Computer Science and Engineering, National Institute of Technology, Rourkela. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Prof. Sanjay Kumar Jena
Professor
Department of CSE
National Institute of Technology
Rourkela-769008

Place: NIT,Rourkela-769008
Date: 01 - 06 - 2015

Acknowledgment

First of all, I would like to express my deep sense of respect and gratitude towards my supervisor Prof. Sanjay Kumar Jena, who has been the guiding force behind this work. I want to thank him for introducing me to the field of social Network and giving me the opportunity to work under him. His undivided faith in this topic and ability to bring out the best of analytical and practical skills in people has been invaluable in tough periods. Without his invaluable advice and assistance it would not have been possible for me to complete this thesis. I am greatly indebted to him for his constant encouragement and invaluable advice in every aspect of my academic life. I consider it my good fortune to have got an opportunity to work with such a wonderful person.

I thank our H.O.D. Prof. S K Rath and Prof. S K Jena for their constant support in my thesis work. They have been great sources of inspiration to me and I thank them from the bottom of my heart.

I would also like to thank all faculty members, PhD scholars, my seniors and juniors and all colleagues to provide me their regular suggestions and encouragements during the whole work.

At last but not the least I am in debt to my family to support me regularly during my hard times.

I wish to thank all faculty members and secretarial staff of the CSE Department for their sympathetic cooperation.

Rasula Venkatesh

Abstract

The popularity of Social Networks during the last several years has attracted attention of cybercriminals for the spreading of spam and malicious contents. In order to send spam messages to lured users, spammers creating fake profiles, leading to fraud or malware campaigns. Sometimes to send malicious messages, cybercriminals use stolen accounts of legitimate users. Nowadays they are creating short URLs by the short URL service provider and posted on to friends board. Lured users unknowingly clicking on these links, then they are redirected to malicious websites. To control such type of activities over Twitter we have calculated a trust score for each user. Based on the trust score, one can decide whether a user is trustable or not. With usage of trust score we have got an accuracy of 92.6% and F-measure is 81% with our proposed approach.

***Keywords*:** Short URLs, Cybercrime, Twitter, Spam Messages, Trust Score

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Social networking is a platform provides to build a social relationship among people using the Internet. Over recent years, social networks are largest and fastest growing networks. There are hundreds of online social networks are present like Facebook, Twitter, LinkedIn etc. are the most popular based on the number of active users. In this networks the users are sharing their personal information. These sites can be used by the government to get opinion of public quickly. On Twitter, users are communicating through tweets. Twitter playing a crucial role for connecting peoples and peoples can discuss on a particular topics like earthquake in Nepal. In Twitter the user can send a massage maximum upto 140 characters only. Twitter allows only unidirectional relationship among the users. User can add tags to the tweets (i.e. # tags) which provides easily combines all the related information.

Twitter has a concept of following. Suppose if a user A follows B signifies that all tweet posted by B would be posted on timeline of A. But user B cannot see the

tweets posted by the user A. By this we can specify that whose tweets the user having an interest to see. These user could be friends, co-workers, celebrities, researchers etc. Twitter acting as news social media for spreading the breaking information over the globe. Twitter has trending topics on the left side of the user timeline. Trending topics contain top 10 hot topics to discuss. In order post a tweet related to trending topic user must include # followed by topic name.

There are millions of tweets are generating per day, the increasing concerns about the trustworthiness of information disseminated throughout the social networks and the privacy breaching threats of participant's private information. In the few years ago the users are limited to viewing of information on the websites. Now online social networks are providing a platform for the users to actively participate over the websites. At the same time there is a cybercriminals attacks like stealing credentials, fake messages etc. Cybercrimes are serious threat for Internet users. Twitter is the one of social network attracted by the most of the malicious users. They are providing malicious links and fake information for advertising purpose or get the money from the lure users.

Twitter having limitation that we can on send 140 characters, the user can not send whole URL in a tweet. There are some of the URL shortening service provider (goo.gl, bit.ly, t.co) present for shortening the long URL to short URL. Spammers are masquerading the actual URLs, i.e. user doesn't know the actual link behind the short URL.

In this project, we are mostly concentration on *"trust score"* of a user. In social

network (like Twitter) user can participate in several social activities. How much trustable a person in social networks. Based on the trust score the user can decide tweet posted by the particular user is trustable or not. If the user is having higher the trust score the information posted by user is legitimate content. Lesser the trust score the information posted by him is more vulnerable, i.e. containing malicious information. The trust score is numerical score with in the range of 0 to 1. For calculating trust, we are considering many parameters are user activities, social connection, user profiles etc.

In the past years, several machine learning algorithms are analyzed features of social network user, still not accurately classifying the malicious users.

1.1 Motivation

Most interaction between two users in online social network is based on trustworthiness between them. In a Twitter network users are posted their tweets and the other cant able to decide how much trustable [1].

See in Figure 1.1 Bob is providing services to the Alice, he dont know Bob is trusted service provider or not. By assigning trust score each user we are classifying the user is malicious are not. Based on this score the online user can decide the respected user tweets are trustable or not.

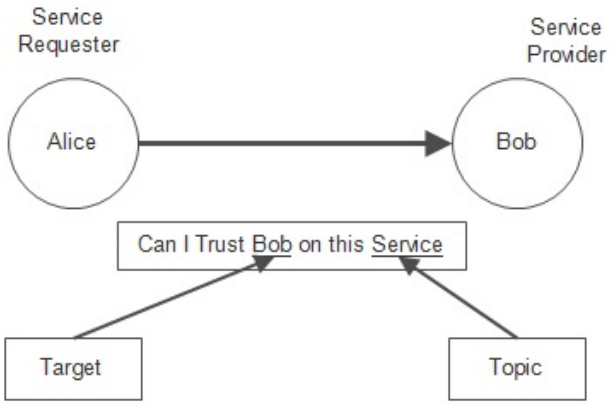

Figure 1.1: Trust Relationship

1.2 Problem Statement

As more and more people are spending increasing amounts of time on social networking sites there is a growing concern for the privacy and legal rights surrounding them.

Spammers and rumors are increased in the social networks for gaining profit.

Lack of inefficiency and incapability of detecting malicious activities in timely fashion and as soon as malicious user detected, then they were creating new profiles.

1.3 Objective

Protecting users from clicking malicious short URLs. This can be done by avoiding user from posting malicious link tweets and detecting such users in social network. In this thesis, we are going to classifying users into malicious or legitimate by using

trust score feature along with user profile features. Here the spammer users are classified in offline.

1.4 Issues

There are a lot of issues while using the social networking sites, like discloser of confidential information, cyberbullying, privacy, defamation, identity theft, spam, malware etc. All these are done mostly by using the fake profiles.

Spam is defined as an electronic messaging system sends unsolicited bulk messages. Spammers on Twitter are user, they try to send unsolicited messages to large number of users for advertising purpose or infecting the user system.

Initially spammers create a legitimate looking profile. For making a friendship with users over the Twitter first he sends legitimate URLs links to build trust. Later the attacker start sending malicious links. So the victim already trusted the attacker, the clicking the URL then malware downloaded into the system, it may not be limited to the malware. Depending on the vulnerability the attacker may steal the session information to impersonate victims on social network.

1.4.1 Neighborhood Attack

Online social network can be represented by the social graph. Each node in the graph is a social network user and the relationship among the users is represented by the edges. There is a neighborhood attack when the malicious user know the

friends (neighbors) of the victim user i.e. the malicious user knows the relationship among the friends also. Then he can find the identity of the user [2]. In social network every user have unique neighborhood graph.

1.4.2 Drive by Download Attack

In this attack the victim visited through the vulnerable browser. It landed on to the actual page after many redirection. This type of attack mostly by the advertisement. It acts as medium to spread malware over the network. The attacker post ads on the users wall. As shown in Figure 1.2, when the user clicking on the ads it is redirecting to malware website. A malware downloaded into the user system, then user computers gets infected [3].

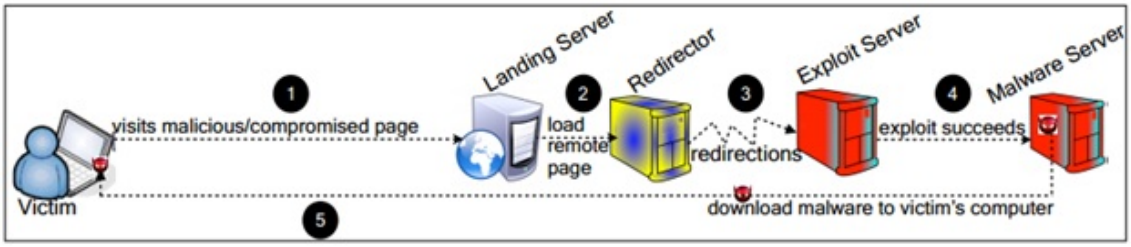

Figure 1.2: Drive by Download

As shown in Figure 1.3, when we are clicking on the malware links, then it downloaded on our system and it sends the keystrokes and screen shots to malicious user or attacker [4]. Then the attacker know the our credential information.

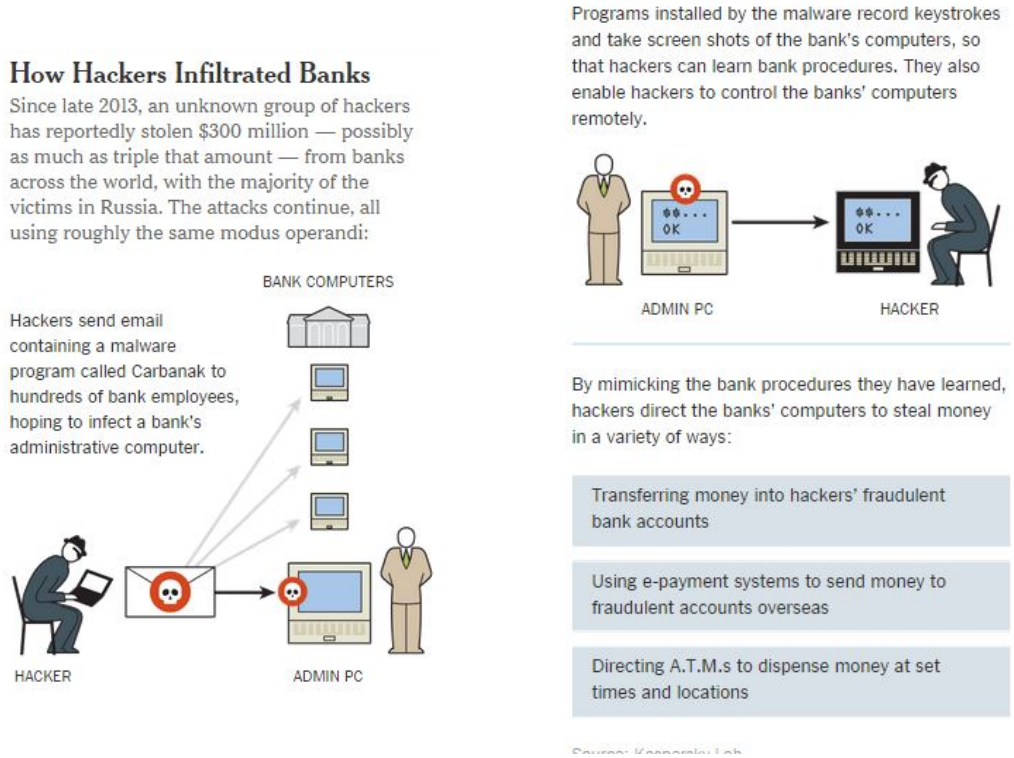

Figure 1.3: Malware Installation

1.4.3 Phishing

Phishing is a social engineering, in which the attacker gets the confidential information from unsuspected victims.

In phishing attack, the attacker provides a fake websites it looks same like original websites. So the lure victims are providing their sensitive information such as passwords, financial information. The attacker gather information from the social network users. Extract the useful information to trick users to phishing websites like as shown in Figure 1.4. For example, attackers can send a phishing website to victims by using the victims friends names.

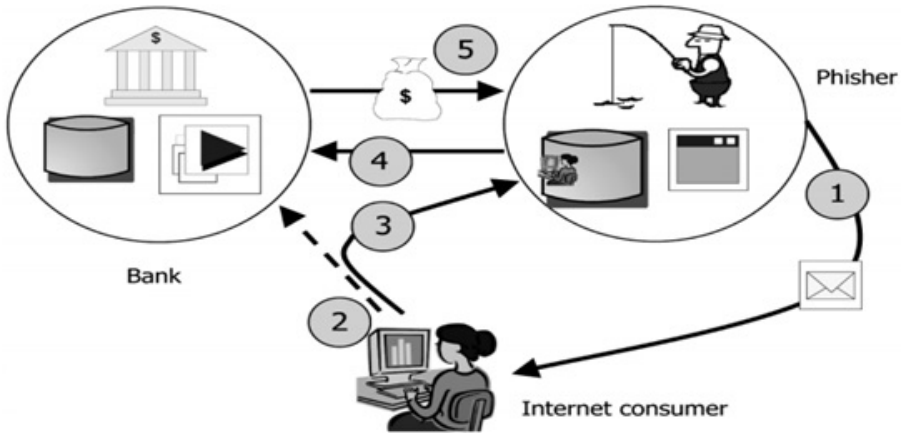

Figure 1.4: Phishing

1.4.4 Shortened and Hidden Links

URL shortening is popular method for reduce the size of URL because the most of the URLs are too long. Users can easily access the shortening service. The user submitting the original URL and the service providing the shortened URL that will redirecting to the original webpage. The social network users can not know to which website it is pointing to. Attacker creating a malicious websites, then instead of posting original links they were using the short URLs. Initially they are making a good relationship with the users by sending legitimate URLs [2]. After making a trust among them, they start sending malicious links usually the user trust the link that is posted. So this increase the click rate of the malicious link.

1.5 Heterogeneous Social Graph Representation of Twitter

In heterogeneous graph representation, three types of vertices in the graph which correspond to three major entities in online social networks (e.g., users, tweets, and hashtag topics).

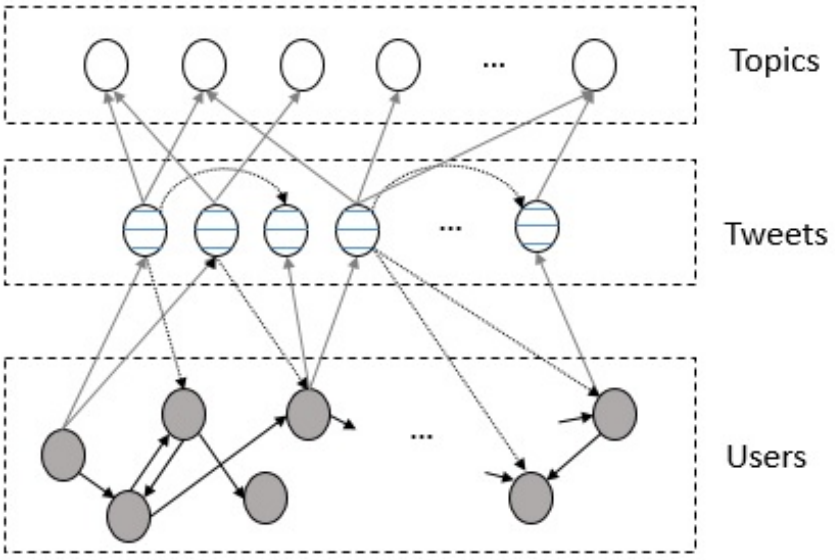

Figure 1.5: Heterogeneous Social Graph

Directed edges connecting vertices in the graph as shown in Figure 1.5 represent different types of social activities. First, an edge from user u_i to user u_j means that u_i relates to u_j in the network (e.g., u_i is following u_j in Twitter). Second, an edge from user u_i to tweet t_j indicates that u_i is the author of t_j (e.g., u_i posts a tweet t_j in Twitter). Third, an edge from tweet t_i to topic h_j represents that h_j is one of the

topics covered in t_i (e.g., h_j is a hashtag topic in a tweet t_i). In addition, there are two more types of directed edges in the graph. One edge starts from tweet t_i and points to another tweet t_j. This represents that t_j is a retweet of t_i. Another type of edges connects a tweet t_i and a user u_j. This specifically captures the mention function in Twitter.

Chapter 2

Literature Review

This chapter gives the overview of existing works on detection of malicious accounts in social networks. Due to raising of social networks, numerous studies have been done related to the detection of malevolent users. Malicious account detection is rely on the behavior of the user. Detection of spammers in online social networks is difficult not only by the nature of spammer. Malicious user easily adopting existing techniques. Different Online Social Networks(OSNs) like Facebook, YouTube and so on has been focused by spammers to connect with clients. OSNs gives a perfect stage to spammers to mask as a benevolent client and attempt to get malevolent posts clicked by ordinary clients.

Some malicious accounts participating in social bots. Social bot automated computer programs. Malevolent post URLs attached with bots. When user is clicking on that it downloading on to the machine. Then it stealing all information from the victims machine. Social bots are controlled by the boot master. Bots may

or may not require input from the user. Bots are looks like an original profile but it randomly selects the profile name, randomly chosen profile image. Social bots are randomly select a user from the list to send request. If the user is accepting the request then it send to all the friends of victim user. Which increases the acceptance rate so that attacker gets more benefit. Bots are monitor the tweets among the two users also [5].

Spam are generally refers to the unsolicited message deliver to the large number of peoples directly or indirectly [6]. There are many different techniques to detect spam message and these techniques depending on the many features which are extracted from behavior of the user and social interaction [7–9]. Lee *et.al.* [10] classified users in to polluter and legitimate users based on the 18-profile based features.

In online social network rumor identification taken much attention. Rumor are malicious users whose true value definitely unverifiable i.e. the value is always false [11]. Sarita *et.al.* [12] study on structural properties of a graph based on the web graph and social graph. Users are present at the center of the graph. The users who having a followers count high they are at borders of the graph. For example celebrities have more number of followers so we are ignoring the celebrities. The normal users who having the maximum followers count, they are taken more attention.

Sangho *et.al.* [13], have given the techniques used by the attacker to void URLs form blacklist of URL service providers. They suggested many URL based features like length of the URL and redirection etc. Pasquale *et.al* [14], have proposed

the classification of malicious and fraudulent behavior of user by using the global and local reputation. A user in the online social network predict and assign the trustworthiness of another user. In past, global reputation is based on the feedbacks of previous activities of the user. Here the malicious user can send as many as feedback about him.

Gupta *et.al* [15], have studied on the bit.ly short URLs. They were classifying the bit.ly short URLs in malicious and benign. Bit.ly facing a problem of work from home, phishing, pornographic information propagation over the network. They were identified some short URL based features and are coupled with the domain related feature for improving the accuracy of classification. De wang *et.al* [16] have analyzed the misuse of short URLs and the characteristics of non-spam and spam users based on the click traffic of URLs. Many supervised learning algorithms like markov model [17] and SVM model [18] are used for detection of rumors over the social networks by the selected features. They are network-based features, content-based features and social network specific features [19]. Michael *et.al* [20] have proposed a Software Privacy Protector (SPP) for Facebook. It improves privacy of a user by implementing methods for detecting malicious users.

Online Impersonation

As shown in Figure 2.1, the attacker or hacker creating a fake accounts and pretending it is created by the original user. They are acting like a correct

NEW DELHI: Union Minister of State for Home Affairs Kiren Rijiju is now a man in search of an identity - not in the real world, but on Twitter. While attempting to set up a personal Twitter handle, the junior minister realised that the handle with his name had already been taken, as was another suitable name.

Home Ministry officials said Mr Rijiju, 43, on Thursday sought to register the handle most obviously close to his name - @Kiren_Rijiju. But he found it already existed, as a fake account.

The handle carried a photo of Mr Rijiju, had been set up in 2009 and has not seen a new tweet since then.

Another possible handle for the minister, @krijiju, too turned out to be fake. This one was followed by over 600 people.

Figure 2.1: Online Impersonation

person [21]. Initially for making a friends they are posting genuine tweets. Then after made a trust relationship, they will start posting malicious links. The friends might think that it is also genuine message. The lure user will get attacked.

2.1 Page Rank Algorithm

The internet can be seen as a large graph. In this graph, each node is considered as a web page, links among the web pages is known as edges of the graph. The connections among the web pages is in single direction or multi direction. Page Rank algorithm is the heart of search engine. It will decides how much important a specific page is and how high to show in search results.

The underlying idea of Page Rank algorithm is *a page is important if other pages are pointing to it.* It means every page connection taking it as vote and it is recommending that page important. It seems like a Page Rank algorithm is counter

of online ballots. Votes given by the pages important to other pages. Based on this results the page is reflected in search results.

Page Rank algorithm is best for calculating trust propagation over a network. It does not require the explicit collection of votes for rating. This approach is related to approaches used in this work.

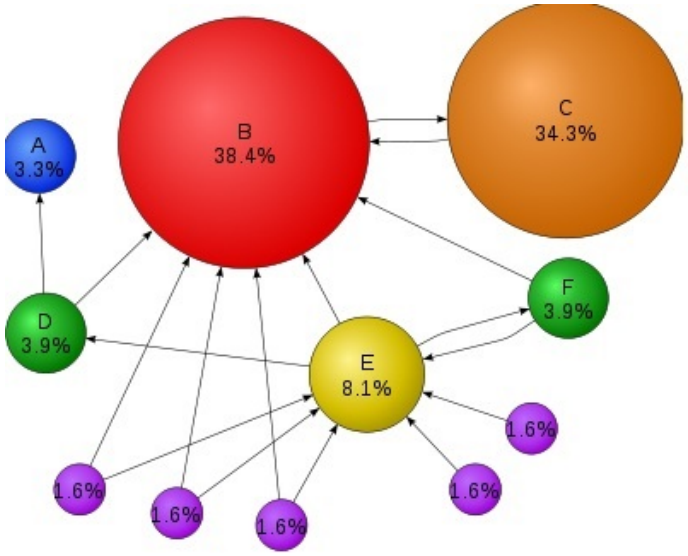

Figure 2.2: Page Rank for Simple Network

Page Rank algorithm is basic technique for citation counting, the term implies that citation counting calculates the references pointing to the object. Rank all the objects accordingly. It has weakness a single link from most important page has more significant than many links from unimportant page [22].

$$R(v) = c \sum_{u \in B_v} R(U)/N_u \tag{2.1}$$

Let v be a web page, then let F be the set of pages v points to and B be the set of pages that point to v. Let $N_v = \|F_v\|$ be the number of links from$v$ and let c be a factor used for normalization. Thus the value assigned to a web page v will be propagated in equal parts to all pages it links to, as shown in Figure 2.2 .

Chapter 3

Proposed work

In this chapter, we are presented an approach for data collection, analysis of data, feature selection, proposed algorithm which is used for calculating special feature trust score and classification algorithms used for classifying malicious users.

3.1 Methodology for Data Collection

Now, we will describe the procedure for data collection. The first step for our analysis is to gather data from Twitter. We collected a data and information of 4230 users. All these information is verified by the Twitter. Twitter and used machine learning algorithms to classify as malicious or not. We used a Twitter API to collect the data and we can collect only the information there in the public domain. If the user is keeping his data secret (i.e. does not allowing other to access his personal information). We have also collected the information of 380 suspicious users. All

these suspicious users are blocked by the Twitter network. As shown in Figure 3.1, later we have collected the data (tweets) of each user. The stream of tweets are accessible by Twitter stream API. Which gives information of tweets are posted in Twitter. There is limit that we can access only 40 latest posts of a user. Some of the tweets contains the short URLs and related hashtags. Here hashtags indicates, the tweets are related to the specific topics.

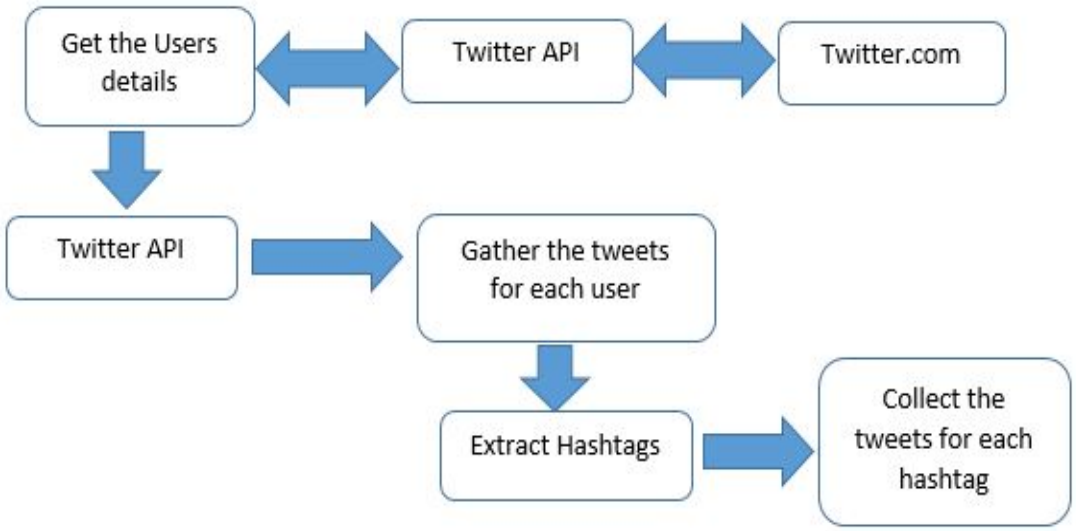

Figure 3.1: Data Collection

Later, we have extracted the all the tweets related to the hashtags. We have collected the tweets of all the hashtags. For example profile data shown in Figure 3.2. It contains the ID of a user, profile name, followers count, friends count etc.

```
1127 1127 108 551 112 Fullerton, CA None
1128 ericcogan 1330 365 1363 New York City http://t.co/m2EzXuNtgC
1129 1129 4 10 232 02116 USA http://t.co/d1C1pVjXKs
1130 Ewe 269 3 12  None
1131 ammonkc 942 594 7473 La'ie, Hi http://t.co/KjlHMFXvVC
1132 1132 142 3 3  None
[{u'message': u'Sorry, that page does not exist.', u'code': 34}]
1134 1134 148 0 0  None
1135 brandon 3440 757 23216 New Jersey, NYC, Delray Beach http://t.co/pp5sMbEkj2
1136 1136 148 15 2  None
1137 1137 27 0 0  None
1138 Faisal 6519 570 12828 Saudi Arabia http://t.co/buyqEae9zZ
1139 1139 1 0 4  None
1140 paul 4476 1772 2996 Elgin, Speyside, Scotland http://t.co/0MYyCNXVj9
1141 1141 139 1 1  None
1142 1142 133 7 2  None
1143 seanluce 655 812 3782 Michigan http://t.co/e0xyf3Lx9T
1144 1144 68 27 59 Flint, MI http://t.co/m2i7rbpRjM
1145 1145 163 1 0  None
1146 1146 153 21 0  None
1147 stevent 307 233 211 Connecticut, USA None
1148 indiefeed 12112 11849 3301 San Diego http://t.co/yO37i9SF6K
1149 sbb 714 908 3788 Boston, MA None
1150 1150 43 1 0  None
1151 Wendi 1 18 10  None
1152 1152 169 0 0  None
1153 AndrewCrow 8345 948 23981 San Francisco http://t.co/liQW9R3dNH
1154 peterme 10482 500 13082 Oakland http://t.co/aU5nTNRKOp
1155 1155 515 281 15492  http://t.co/I7EzIWzg0F
1156 1156 584 1121 85  None
1157 1157 127 0 2  None
```

Figure 3.2: User Profile Data

```
1131 ammonkc 942 594 7473 La'ie, Hi http://t.co/KjlHMFXvVC
1132 1132 142 3 3  None
[{u'message': u'Sorry, that page does not exist.', u'code': 34}]
1134 1134 148 0 0  None
```

Figure 3.3: Suspended User

By the Figure 3.3. we can see that the user ID 1133 details are not available i.e. the user is suspended from the Twitter. We can treat that user as malicious user. Definitely we are assigning trust score 0 to the user. If users are connecting to these malicious users then the trust score of users is decreased.

```
Uh oh dont make your dad mad
Block it like it's hot!! #pacwins
RT @jimmykimmel: My fight outfit #PacquiaoMayweather http://t.co/s1gwoK
Alc4
Yeah, this is the thing - we the people can't support fast tracking thi
s secret thing which is secret.  https://t.co/G6VC6CM14u
RT @jakegagne: Ready to get down with the get down back at VIR this wee
kend 📷 @geocrashphoto https://t.co/NDb9bscCx1
RT @RikerGoogling: why don't combadges use ssl
We can't fast track TPP, it's not comprehensible yet.
RT @tommypjr: SF tech CEO: If @NancyPelosi supports open Net, she shoul
d oppose #FastTrack &amp;amp; #TPP http://t.co/H0Fj6haULG @idltweets http:/
/…
RT @RoadRaceFactory: Catch @CamPetersen72 on @nextmotochamp! Also #cove
rmodel @jakegagne! It's a double dose of #RRF  https://t.co/d7iz0Lux…
RT @TheOnion: Snowden Documents: NSA Can Search For Words Spoken In Pho
ne Calls http://t.co/dQNLtUzdLG #WhatDoYouThink? http://t.co/HlPBjjp…
Mauna Kea telescope protests: Scientists need to reflect on history and
culture. http://t.co/T8efmIMDM1
RT @davepell: I'd love to watch my wife beat the shit out of Mayweather
```

Figure 3.4: Hashtags

As shown in the Figure 3.4 all those hashtags or trending topics are extracted from tweets.

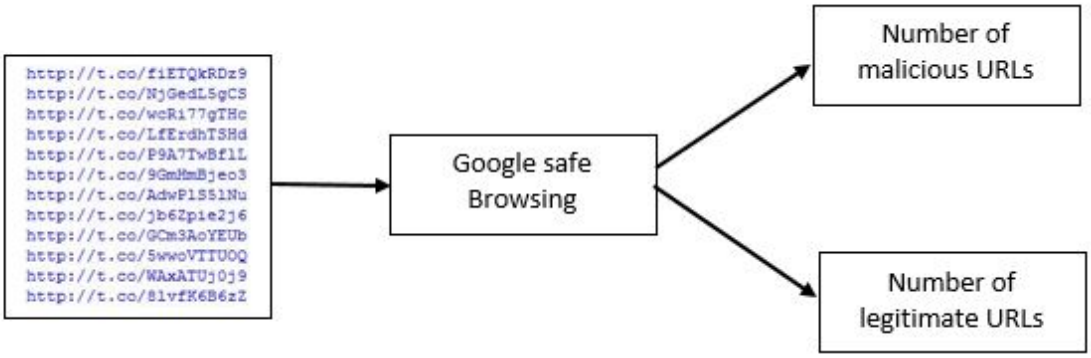

Figure 3.5: Short URLs Labeling

Twitter quickly reacts to detected malicious profile, as well as deletes any malicious tweet found in order to get the Social Network clean from fraud. So if we want to get this malicious data for our analysis we should be quicker than Twitter and gather as much data as possible before it is deleted.

As shown in Figure 3.5 each extracted short URLs from hashtag tweets, is queried to google safe browsing API to find whether the short URLs are malicious or not. Google safe browsing maintain a black listed URLs. When the request is sent, it searches against blacklisted URLs. If query returns *false* then the requested URL is malicious. If it returns *true* then the URL is legitimate. We are assigning trust score to *hashtags* based on number of legitimate URLs i.e.

$$\#\ Trust\ score\ of\ hashtag\ =\ \frac{\#\ Number\ of\ legitimate\ URLs}{\#\ Total\ no\ of\ URLs} \tag{3.1}$$

If the hashtag having the high trust score, then the information related to that is more trustable. If the trust score value is low all the information related to that is malicious. If the trust score is 0.5 then it is not decided (i.e. it may be either malicious or legitimate).

Many Twitter spam detection schemes have been proposed. These schemes use different strategies for classifying suspicious users or suspicious tweets.

Analyzing user features: such as the account creation date or the number of followers. The advantage of this approach is that the information is easily available; the problem is that attackers to bypass detection mechanisms could forge some of

these features.

Analyzing relationships between users: The advantage is that it is more complicated for an attacker to create a complete user network to bypass detection; the downside is that it is difficult and slow to recreate this network for an analysis.

Analyzing tweets: This is a different approach that usually does not take the user features into account, just the tweet itself. Usually, there is not much to analyze but the links, this tweet information may be correlated with other features for a more complete approach. The usual approach here is to compare tweets with other ones gathered from known malicious campaigns.

3.2 Proposed Algorithm

Data: a heterogeneous graph representation G (V, E), a trust threshold Θ;

Result: a set of malicious activities Mal;

1 Initialize a trustworthiness score of 0.5 to each node in G;

2 Initialize a trust score to each T in G based on the formula 3.1

3 **repeat**

4 $\quad \forall$ v $\forall$ u Trust score(u)=$\sum_{x \in B_u} Trustscore(x)/N_u$

5 **until** *all nodes are visited in U* ;

6 **repeat**

7 $\quad \forall$ v Trust score(v)=$\sum_{x \in B_v} Trustscore(x)/N_v$

8 **until** *all nodes are visited in V* ;

9 Repeat step 6 to 8 until reaching a stable status; each vertex v is calculated a trust score T(v);

10 initialize Mal to be $\emptyset$;

11 **for** *every* $v \in V$ **do**

12 $\quad$ **if** *(T(v)*$\leq \Theta$ *)* **then**

13 $\quad\quad$ let Mal = Mal $\cup$ v;

14 return Mal;

Where

N_u, N_v is the out degree of the node U , V

B_u, B_v is the set of nodes pointed by node U, V

T(v) is trust score of node v.

The most important step in the above algorithm is the calculating trust score for the user node in heterogeneous social graph. Trust score is calculated based on the PageRank algorithm. Initially, Mal is empty and it store the information about the nodes which are less than. Here we are classifying based on the trust score. If the user having score less than the threshold value are classified as malicious. As you

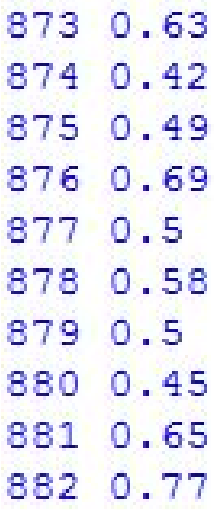

Figure 3.6: User Scores

can observe in Figure 3.6, after implementing the above algorithm we are got user id's with trust score of range 0 to 1.

3.3 Feature Selection

In this approach, we propose a new feature for detecting malicious user. The following are feature used in our classification

User ID: it is numerical value. User assigned with one value when creating an account in Twitter. It is unique value for identifying a user in Twitter.

Followers Count: it means that number of Twitter users are following him in Twitter. If the user having more followers counts, then the user may be celebrities, news channels, politicians etc. Here in our approach we are omitting the users who are having followers count.

Friends Count: it means that to how many number of the user is following. In online social network the spammer having high following count and low followers count. For gaining the more benefit they were sending a friend request more number of peoples in the network and less users are following spammers.

Status Count: status count it stats that how actively the user in Twitter. Mostly the spammers having the large status count because they are sending more malicious URLs to many users.

User location: it shows that the user belongs to which geographical region. There some of the users from particular location are sending more malicious URLs. The URLs having a domain IP addresses based on that from which domain the spam URLs are generated.

Has URL: Some users having URL in profile data.

Spam URLs: spammers are continuously posting malicious URLs to all the users.

Here we are finding the number of spam URLs present in all tweets i.e. count of spam URLs.

Duplicate URLs: the duplicate URLs identifies the number URLs are tweeted repeatedly again and again. Spammers are creating URLs sending many times the same URL for getting the benefit from the lure users. Here the user may clicking on at least one of the URLs. Non spammers creating a URLs on different topics. We are computed this feature by average of URLs posted by the user.

$$DuplicateURLs = \frac{\#Totalnumberof URLs}{\#Totalnumberof uniqueURLs} \tag{3.2}$$

By the above formula 3.2, if the value of Duplicate URLs is more then there is chance that the user is malicious. This metric taken advantage for detection because for creating different malicious URLs the spammer has to incur an extra work or require more money to create URL for same content.

Trust Score: trust score is a special feature, we are calculated based on the short URLs of the hashtags.

Trust score is more important feature it is calculated based on tweets, hashtags etc.

Chapter 4

Evaluation and Results

We presented the evaluation of malicious accounts, by analyzing the collected data of 4820 users information and 380 suspended user information. After calculating the feature values then the feature data feed them to three machine learning algorithms- Decision Tree, Random Forest, Nave Bayes classifiers. For this classification, we have used the most popular Weka software package. In this most of the classification algorithms are implemented. Weka is an open source collection of machine learning classifiers for data mining. The following Figure 4.1 shows the approach for classification.

Now we describe the way of classification of malicious users. Initially dataset is divided into training dataset (80%), testing dataset (20%). In order to assess the most efficient mechanism to detect malicious accounts, we inspected various machine learning algorithm. All below classifiers are the standard classifiers and widely used in solving problems.

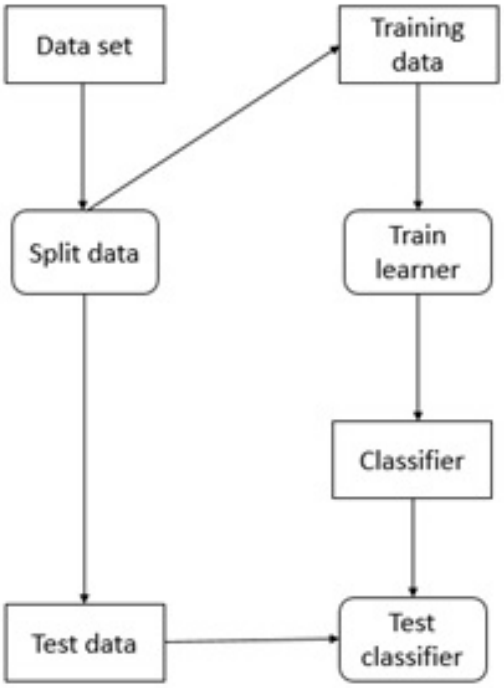

Figure 4.1: Classification Approach

4.1 Supervised Learning Algorithms

The following is the detail description about the classifiers.

4.1.1 Decision Tree Classifier

Decision tree most popular classifier which generates a tree like structure feature names corresponding to internal nodes feature values corresponding to branches, and class labels corresponding to leaf nodes. In this each node represents the test on the attributes i.e. decisions of the attribute. If the attribute is satisfies the required condition based on that it divide the data. Tree display the relationships among

attributes are there in the training data set. Decision tree is predictive model that uses a set of binary rules applied to calculate the target value.

Constructing the decision tree is done by selecting the attributes that splits the training data in proper class i.e. legitimate and malicious classes. Decision trees implemented based on the *information gain*. Which is based on the *entropy*. If the entropy is low then the set is homogeneity of type and if entropy is zero then the set is contains only one type of data. Once identified splitting attribute then rest of the training data are pushing down the tree i.e. data that is satisfying the splitting criteria are thrown into the *true* side of the tree. While, if the data is not satisfies the required criteria are thrown into the *false* side of the tree. The above process is repeated until the each node in the tree contains data of the same class, at that it store the class label.

During the classification, it predicts the class of an unknown data based on criteria defined over the node, starting from the root node. If the attribute in the data satisfies the condition then the classifier follows the *YES* class. If not satisfies then it follows the *NO* class. It checks the each criteria in the right path until reaching the leaf nodes.

4.1.2 Nave Bayes Classifier

Nave based classifiers is based on the probability and based on applying Bayes theorem with strong independence assumption. The descriptive term for the above probability model is *independent feature model*.

Nave Bayes classifier assumes that particular class feature presence or absence is unrelated to the other class feature presence or absence. In this classifier, we have a hypothesis that the given data belongs to the related class. Precise nature of the probability model, in supervised learning settings we can train nave Bayes classifier very efficiently. In many practical applications, it uses maximum likelihood for parameter estimation. In many complex real world situations, nave Bayes classifier works well. The advantage of nave Bayes classifier is that for estimate the parameters it require only the small amount of training data.

Nave Bayes probabilistic model

The probability model is a conditional model over a dependent class variable with limited number of outcomes means classes, conditions on the feature variables F_1 to F_n.

$$P(\frac{C}{F_1, ..., F_n}) \tag{4.1}$$

If the value of n is large, basing a model is infeasible. Then we reformulating the model then it feasible or tractable.

$$P(\frac{C}{F_1, ..., F_n}) = \frac{P(C)P(\frac{F_1, ..., F_n}{C})}{P(F_1, ..., F_n)} \tag{4.2}$$

The above equation can be written plain english as follows

$$posterior = \frac{prior * likelihod}{evidence} \tag{4.3}$$

In reality, we are only concentrating on numerator, because denominator not depending on the class c and values of features F_i.

4.1.3 Random Forest Classifier

During the training period random forest builds many trees. In random forest each node is split using the best among a subset of predictors randomly chosen at the node. It is user-friendly because it has only two parameters. To classify unknown samples, the input queried to every tree in the forest. Here each tree used for predicting unknown sample data. The overall output of a predicted sample data is based on class label with highest number of votes among all the trees.

Random forest is constructed based following steps

- There are N cases in training set. All cases are at random, with replacement, taken from the data set. For growing a tree all the samples will be trained.
- If m variables are selected from the set of M variables at each node ($m<<M$) and m is used for best split the node. During forest growing the value of m is constant.
- The tree is growing up to the large extend as possible, without pruning.

4.1.4 Evaluation Metrics

Accuracy (A) and F-measure are the metrics which are used for the evaluation of the classifier performance. F- Measure is defined in terms of Recall (R) and Precision (P). If evaluation metrics having higher value, then the classifier is best suitable for data set. The evaluation metrics described effectively by confusion matrix Table 4.1.

Table 4.1: Confusion Matrix

	Malicious	Legitimate
Malicious	TP	FN
Legitimate	FP	TN

TP(True Positive) means actual class of a testing data is malicious and it classified as malicious.

FN means actual class is malicious and predicted as non-malicious.

FP means actual class is legitimate and classified as malicious.

TN means actual class is legitimate and classified as non-malicious.

$$P = \frac{TP}{(TP + FP)} \tag{4.4}$$

$$R = \frac{TP}{(TP + FN)} \tag{4.5}$$

$$F - measure = \frac{2 * (P * R)}{(P + R)} \tag{4.6}$$

$$A = \frac{(TP + TN)}{TP + FN + FP + TN} \tag{4.7}$$

4.2 Results

The objective of current study is identifying aberrant behavior of users in Twitter. We have analyzed user suspiciousness based on the trust score. If the calculated trust score is greater than the threshold value Θ then the user is legitimate user. We are taken a threshold value as 0.5. If the user score is less than 0.5 then the user no more trustable as shown in Figure 4.2.

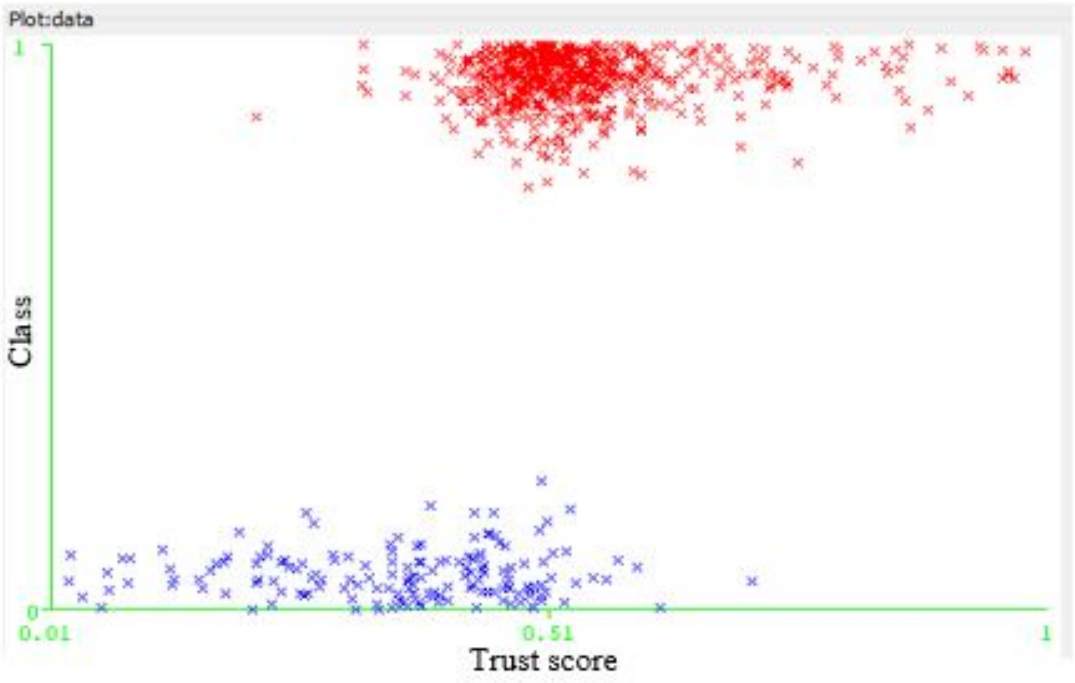

Figure 4.2: Classification of Users

Here, we treat the obtained trust score as a feature along with the all obtained user profile features like followers count, following count, status count etc.

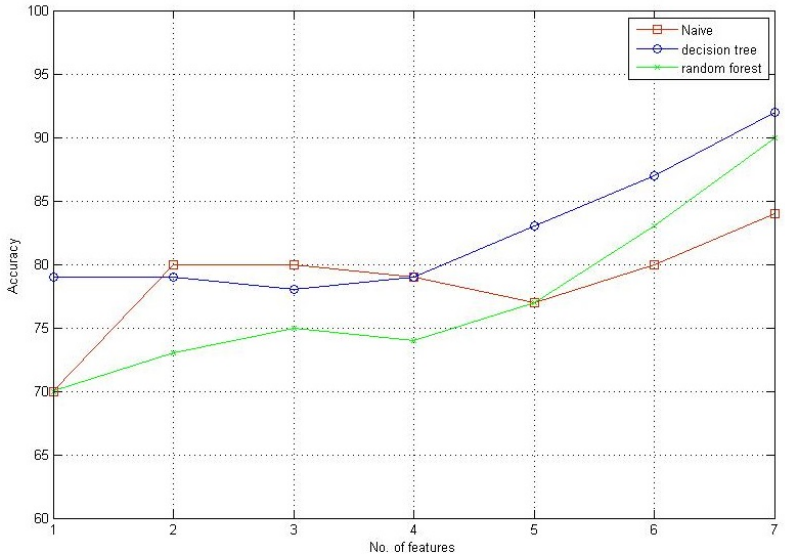

Figure 4.3: Efficiency vs. no of features in training data set

Table 4.3: Comparison of Classifiers

Evalution Metric	**Decision Tree**	**Naive Bayes**	**Random Forest**
Accuracy	**92.6%**	89.9%	90.4%
F-measure(Malicious)	**81.0%**	64.4%	76.3%
F-measure(Legitimate)	**95.5%**	93.4%	94.0%
True Positive Rate	88.2%	80.9%	79.0%
False Positive Rate	93.6%	90.1%	93.0%
Positive Predictive Rate	74.9%	53.5%	73.7%
Negative Predictive Rate	97.3%	97.1%	94.8%

In the Figure 4.3 it shows efficiency of each classifier based on the number of features selected. When we are adding the trust score feature to training data set the efficiency of all the algorithms are increased. In the Table 4.3 it shows that decision tree works better compared with the other classifiers. In our dataset, decision tree correctly classifies 75% malicious users. 25% malicious users are misclassified as legitimate.

Chapter 5

Conclusion and Future Scope

In this thesis, we have developed an algorithm for calculating trust score for each user in heterogeneous social graph for Twitter. The trust score is special a feature that can be used to detect malicious activities in Twitter with high accuracy. Our classifier attains an improved F-measure is 81% and with an accuracy of 92.6%.

In this work, we have successfully detected malicious users. For calculating trust score we have considered only short URLs of trending topics. Based on the backward propagation, we assign trust score to tweets if trending topics present in that tweet and followed by the users.Future work deals with calculation of trust score by considering the short URLs present in the tweet.

Bibliography

[1] Wenjun Jiang, Guojun Wang, and Jie Wu. Generating trusted graphs for trust evaluation in online social networks. *Future generation computer systems*, 31:48–58, 2014.

[2] Dolvara Gunatilaka. A survey of privacy and seucrity issues in social networks. In *Proceedings of the 27th IEEE International Conference on Computer Communications. Washington: IEEE Computer Society*, 2011.

[3] Birhanu Mekuria Eshete. *Effective Analysis, Characterization, and Detection of Malicious Activities on the Web*. PhD thesis, Fondazione Bruno Kessler, Italy, 2013.

[4] The New york Times. `http://www.nytimes.com/2015/02/15/world/bank-hackers-steal-millions-via-malware.html?_r=1`.

[5] Erhardt C Graeff. What we should do before the social bots take over: Online privacy protection and the political economy of our near future. 2014.

[6] Gordon V Cormack. Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4):335–455, 2007.

[7] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 49–62. ACM, 2009.

[8] Hongyu Gao, Yan Chen, Kathy Lee, Diana Palsetia, and Alok Choudhary. Poster: online spam filtering in social networks. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 769–772. ACM, 2011.

[9] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 35–47. ACM, 2010.

[10] Kyumin Lee, James Caverlee, and Steve Webb. The social honeypot project: protecting online communities from spammers. In *Proceedings of the 19th international conference on World wide web*, pages 1139–1140. ACM, 2010.

[11] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics, 2011.

[12] Sarita Yardi, Daniel Romero, Grant Schoenebeck, et al. Detecting spam in a twitter network. *First Monday*, 15(1), 2009.

[13] Sangho Lee and Jong Kim. Warningbird: Detecting suspicious urls in twitter stream. In *NDSS*, 2012.

[14] Pasquale De Meo, Fabrizio Messina, Domenico Rosaci, and Giuseppe ML Sarné. Recommending users in social networks by integrating local and global reputation. In *Internet and Distributed Computing Systems*, pages 437–446. Springer, 2014.

[15] Neha Gupta, Anupama Aggarwal, and Ponnurangam Kumaraguru. bit. ly/malicious: Deep dive into short url based e-crime detection. In *Electronic Crime Research (eCrime), 2014 APWG Symposium on*, pages 14–24. IEEE, 2014.

[16] De Wang, Shamkant B Navathe, Ling Liu, Danesh Irani, Acar Tamersoy, and Calton Pu. Click traffic analysis of short url spam on twitter. In *Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), 2013 9th International Conference Conference on*, pages 250–259. IEEE, 2013.

[17] Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev. What's with the attitude?: identifying sentences with attitude in online discussions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1245–1255. Association for Computational Linguistics, 2010.

[18] Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 13. ACM, 2012.

[19] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics, 2011.

[20] Michael Fire, Dima Kagan, Aviad Elyashar, and Yuval Elovici. Friend or foe? fake profile identification in online social networks. *Social Network Analysis and Mining*, 4(1):1–23, 2014.

[21] NDTV. `http://www.ndtv.com/india-news/fake-handles-keep-union-minister-kiren-rijiju-from-making-twitter-entry-761337`.

[22] Wikipedia. `http://en.wikipedia.org/wiki/PageRank`.