



Detect Spam Messages and Fake Users in Online Social Networks Using Support Vector Machines

CHAVALA SIREESHA ^{#1}, K.RAMBABU ^{#2}

^{#1} MSC Student, Master of Computer Science,

D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

^{#2} Head & Assistant Professor, Master of Computer Applications,

D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

ABSTRACT

In current days social media plays a vital role on each and every human being for sharing their updates from one person to other person within certain location. These social media will try to share reviews and feedback that is posted by the end users and hence it greatly help the online users to take decision about any product or tweet based on that topic reviews. In general different users try to post different types of reviews based on their individual opinions for an appropriate product or post. Hence it laid a way for the intruders to post spam reviews and spam comments for the appropriate post. Identifying these spammers and the spam content is a hot topic of research and also some people try to create duplicate tweets based on same subject name to recommend that as positive tweet, hence the system should able to capable to identify such a duplicate tweets and they need to block that duplicate tweets and tag that user as Fake User. Here we used Naïve Bayes Classification algorithm to classify the tweets based on spam keywords. By conducting various experiments on our proposed model we finally came to a conclusion that our proposed approach is best in finding the solution to block spammers and fake users to misuse the social media.

Key Words: Social Media, Spammer, Fake User, Naïve Bayes Classification

1. INTRODUCTION

In today's world almost any kind of information is mainly obtained using internet from any source across the world. Due to the increase of social media sites the users are permitted to collect the bulk amount of data and share such huge amount of data from one location to other. This huge amount of data may sometimes contain the duplicate or fake content which is uploaded by the end user[1]. Twitter is one of the micro blog which has rapidly become an online source for acquiring real-time information about users. Twitter is one kind of OSN network, in which all the users can share information like news, current affairs and new updates around the world, climate conditions, predictions and a lot more. Most of the people are connected with twitter because the information is instantly conveyed to his/her followers, without any restrictions. With the invention of several social networks in the real world, there is a huge need to study about each and every individual user and their behaviors in the social media. Without the proper knowledge about OSN usage, the users can be easily tricked by the intruders or fraudulent[2].



Figure. 1. Represent the Tips and Functionalities of Social Media

From the above figure 1, we can see the tips and functionality of social media like Search for new topic, Follow Others Chat with different persons around the world, Post Own Tweet, Share the Tweet posted by others, Like the Tweets posted by other users, Find out the Updated News Information, Send Friend Request and Accept Friends Request[3] and a lot more.

Normally the social media applications operate in a dialogic transmission system, where the term dialogic means multiple sources will be operated by the multiple receivers from multiple locations[4]. This is very much quite opposite to the primitive social media applications which try to follow under mono-logic transmission model. Some of the best social media web sites which became popular in very less time include:

1. Facebook
2. Twitter,
3. TikTok,
4. WeChat,
5. Instagram,
6. LinkedIn and a lot more.

Recently, a lot of researchers are attracted to the concept of the detection of spam users and spam content in social networking sites. Almost it is very difficult task to identify the Spam detection in social media because of its wide usage. There is one term hazardous maneuvers, which is mostly adopted by several spammers in order to cause massive destruction of the community in the real world.

2. BACKGROUND WORK

In this section we will mainly discuss about the background work that is carried out in order to prove the performance of our proposed spammer and fake user identification in social media.

MOTIVATION

The main motivation for starting this current paper came from the following examples:

First we try to observe the evidence collected from primitive OSN topic manipulation, in general the online user try to use an influence model for analyzing the dynamic nature of an endogenous hash tag and identify the total manipulation of endogenous diffusion. Also we try to study the online social media usage by level by level based on some topics. If we consider a Post at topic level, the OSN User[5] will try to cover the topics' popularity, topic coverage, its transmission speed, wide coverage of content and its reputation[6].

Here we try to apply some classification algorithm in the data mining and machine learning domain like Naïve Bayes Classification algorithm and then try to observe each factor which is extracted from that tweet. Here the bayes classification technique is very efficient in finding the spam content accuracy from a given tweet [7]. Here the naïve bayes classification technique is able o find the transmission factor that is required to transmit the data from one location to other location. Here we try to illustrate the We further illustrate the interaction pattern between malicious accounts and authenticated accounts, with respect to posted message in OSN.

For the comparison of normal messages and spam messages,we try to collect a list of messages in online social networks and try to apply classification technique on those set of messages. Those messages which don't have any vulgar meaning or abused content is placed in one list and those which contain abused meaning and abused content will be placed in another list. Here we try to identify the malicious users who don't have a registered account but try to enter illegally into the OSN account and try to create some fake

messages and spam content related post on others accounts[8].Hence our proposed naïve bayes classification algorithm will accurately identify such a user's and try to avoid those users not to login into the account[9].

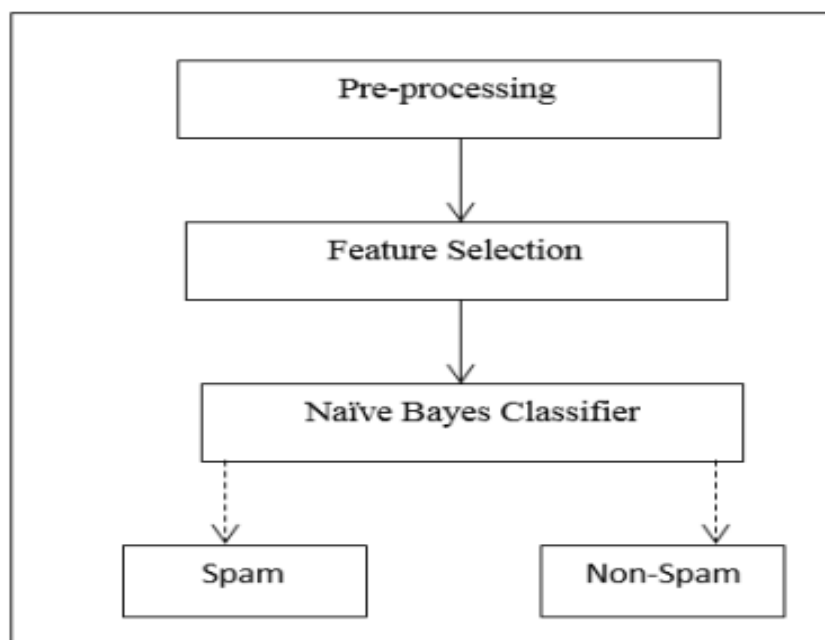


Figure. 2. Represent the Flow of Spam Identification Using Naïve Bayes Classification Algorithm in Social Media

3. PROPOSED SPAMMER AND FAKE USER IDENTIFICATION TECHNIQUE USING RANDOM FOREST CLASSIFIER

In this section we mainly define about the proposed **spammer and fake user identification** technique over a social media using Random Forest Classifier technique.

PRELIMINARY KNOWLEDGE

Random forest has nearly the same hyper parameters as a decision tree classification algorithms. This RF model adds some additional randomness to the input data model, in which the tree grows bigger and bigger[10]. This is best in searching the features which are gathered from the message and then they are identified based on the features extracted from the dataset.

The proposed RF classifier algorithm working steps are shown in figure 4 :

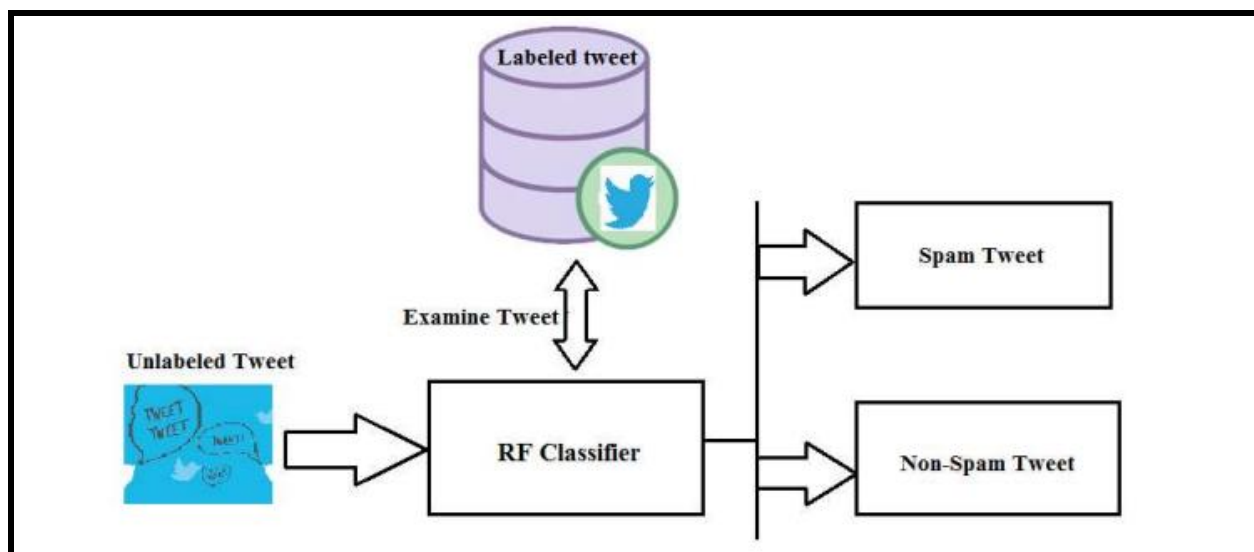


Figure 3. Represent the Flow of Proposed Random Forest Classifier Algorithm in Finding Spammer and Fake Users in the Social Media

STEP 1:

Initially the application need to collect a set of raw tweets from several OSN users and place all those tweets in a separate data storage location. These tweets may contain some useful information as well as some may be contain spam content and abused in its internal meaning, All these tweets are kept as unlabeled tweets.

STEP 2:

Now the RF classifier try to maintain a set of spam words by gathering from various internet sources, this will place all the words in a separate database in order to match these words with the content features which are extracted from the raw tweets[11].

STEP 3:

Now the OSN users try to create new tweets and share among the set of friends who are in followers and following list. Once the tweets are shared the end users try to share comment and feedback for this owners post. Now the RF classifier tries to extract the main features from the tweet messages and try to classify these features with the list of spam words[12].

STEP 4:

Here the classification of tweets is processed and those which are containing positive meaning will be identified and classified as positive or Non Spam tweets and those which are matched with Spam words are identified as Spam tweets.

STEP 5:

Now the users who try to post such spam posts are automatically classified and maintained in a separate list and those users who try to post always positive tweets are kept in separate list.

For analysing the anomalous behaviour of Twitter based on URL, the dataset is prepared by accumulating 1000 tweets of a user. By using RF classifier on some twitter dataset, we try to observe the 5 main functions which cause the spam and fake content in OSN service. They are as follows:

URL RANK GENERATION:

This is the primary function which is used in order to get the URL for the posted tweet. This URL is basically send to the website of ALEXA where the source code is obtained and the tree is generated by the help of web scraper from the given source code.

TWEET SIMILARITY

This is the most important function which is classified by using RF classifier in which the tweet is identified not based on its content but also examined by analysing the total URL of that tweet.

MALWARE URL

This is the third function which is classified and identified based on RF classifier, in which this is used to share the malware data from a user account to other accessors account. This will be identified by the API which is present for accessing the twitter application like WebOfTrust (WOT) API is used to check the reputation of the URL that whether it is a good URL or contains some malware.

TIME DIFFERENCE

This is the fourth function which is identified based on RF classifier in which we try to find out the time difference between each and every tweet which is posted on user wall. Here the calculation is done mostly on all the tweets with its previous three tweets and the next three tweets, and forms the cluster of seven tweets.

ADULT CONTENT IDENTIFICATION

This is last function which is classified in order to see if there is any bad or vulgar adult content is present in the tweet or in its corresponding message. If there is any such content present immediately the RF classifier need to identify such a tweets and try to mark them as spam content.

4. IMPLEMENTATION PHASE

Implementation is the stage where the theoretical design is converted into programmatically manner. In this stage we will divide the application into a number of modules and then coded for deployment. The front end of the application takes JSP,HTML and Java Beans and as a Back-End Data base we took My SQL data base. The application is divided mainly into following 2 modules and inside these there are several other sub modules available. They are as follows:

1. Tweet Server Module
2. OSN User Module

1. TWEET SERVER MODULE

In this module, the tweet server is one who acts like a admin for this application. Initially the tweet server need to login by using valid user name and password. After login successful he can do some operations such as

- View and Authorize Users,
- Add and View Spam Filters ,
- View All User Posted Tweets,
- View All User Tweets Based On URLs,
- View Friend Request and Response,
- View All Tweets with Re-Tweets,
- View All Tweets ,
- Re-Tweets and Comments,
- View All Spammers Detection,
- View All Fake User Identification,
- View Fake User Identification Results,
- View Fake Tweet Identification Results.



2. OSN USER MODULE

In this module, there are n numbers of users are present. User should register first before doing any operations. Once the user gets registered ,he need to get activation permission from the web server and once if he gets the activation.Now he can login into his account with his valid id and password and once he gets login successful he will do some operations like

- View User Profile,
- Search Friends ,
- Create Tweets,
- View My Friends,
- View Friend Requests,

- Search Tweets and Comment ,
- View My Tweets and Comments,
- View Friend's Retweets and Give Comments.

5. EXPERIMENTAL RESULTS

1) Main WINDOW

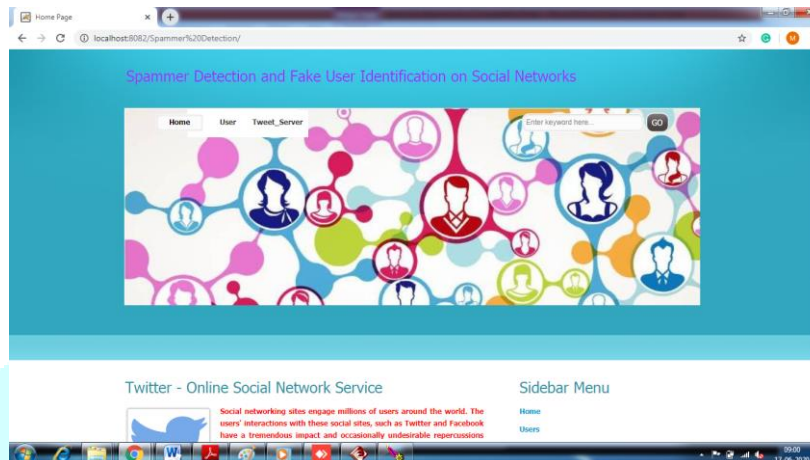


Figure . Represents the Source Window

2) Tweet Server Window

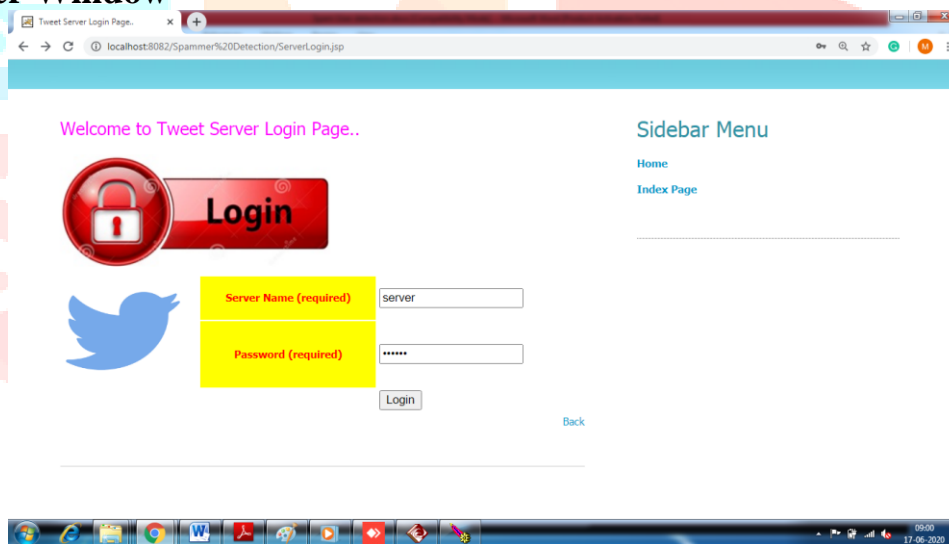


Figure .Represents the Tweet Server Window

3) Tweet Server Main Page



Figure . Represents the Tweet Server Main Page

4) User Registration Page

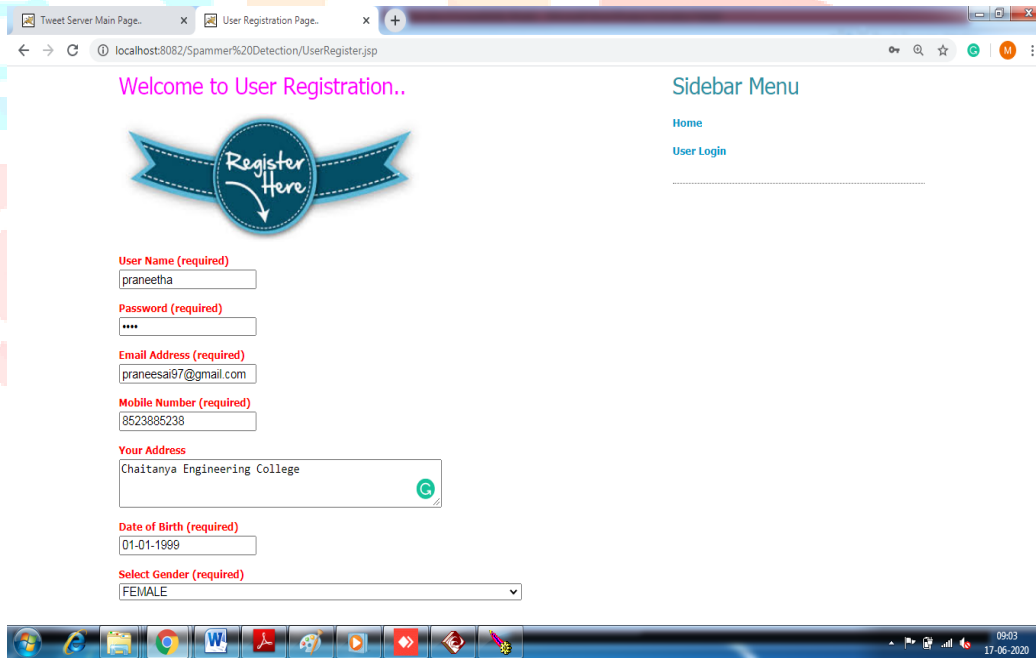


Figure . Represents the User Registration Page

5) Admin Try to Authorize the Users


ID	Name	Email	Phone	Address	Status	Role
5	Manjunath	tmksmanju13@gmail.com	9535866270	#27,4th Cross,Rajajinagar	Authorized	Normal
6	tmksmanju	tmksmanju13@gmail.com	9535866270	#7827,4th Cross,Rajajinagar	Authorized	Fake User
7	praneetha	praneesa197@gmail.com	8523885238	Chaitanya Engineering College	waiting	Normal

[Back](#)

Figure . Represents the Admin try to Authorize Users

6) User Login Page

Welcome User **praneetha ..**

 Social networking sites engage millions of users around the world. The users' interactions with these social sites, such as Twitter and Facebook have a tremendous impact and occasionally undesirable repercussions for daily life. The prominent social networking sites have turned into a target platform for the spammers to disperse a huge amount of irrelevant and deleterious information. Twitter, for example, has become one of the most extravagantly used platforms of all times and therefore allows an unreasonable amount of spam. Fake users send undesired tweets to users to promote services or websites that not only affect legitimate users but also disrupt resource consumption. Moreover, the possibility of expanding invalid information to users through fake identities has increased that results in the unrolling of harmful content. Recently, the detection of spammers and identification of fake users on Twitter has become a common area of research in contemporary online social Networks (OSNs). In this paper, we perform a review of techniques used for detecting spammers on Twitter. Moreover, a taxonomy of the Twitter spam detection approaches is presented that classifies the techniques based on their ability to detect: (i) fake content, (ii) spam based on URL, (iii) spam in trending topics, and (iv) fake users. The presented techniques are also compared based on various features, such as user features, content features, graph features, structure features, and time features. We are hopeful that the presented study will be a useful resource for researchers to find the highlights of recent developments in Twitter spam detection on a single platform.

User Menu

- Home
- My Profile
- Search Friends
- Create Tweets
- View My Friends
- View Friend Requests
- Search Tweets and Comment
- View My Tweets and Comments
- View Friend's Retweets and Give Comments
- Log Out

Figure . Represents the User Login Page

7) Admin add Spam Words



Figure . Represents the Add Spam Words

8) Admin View All Filter Details

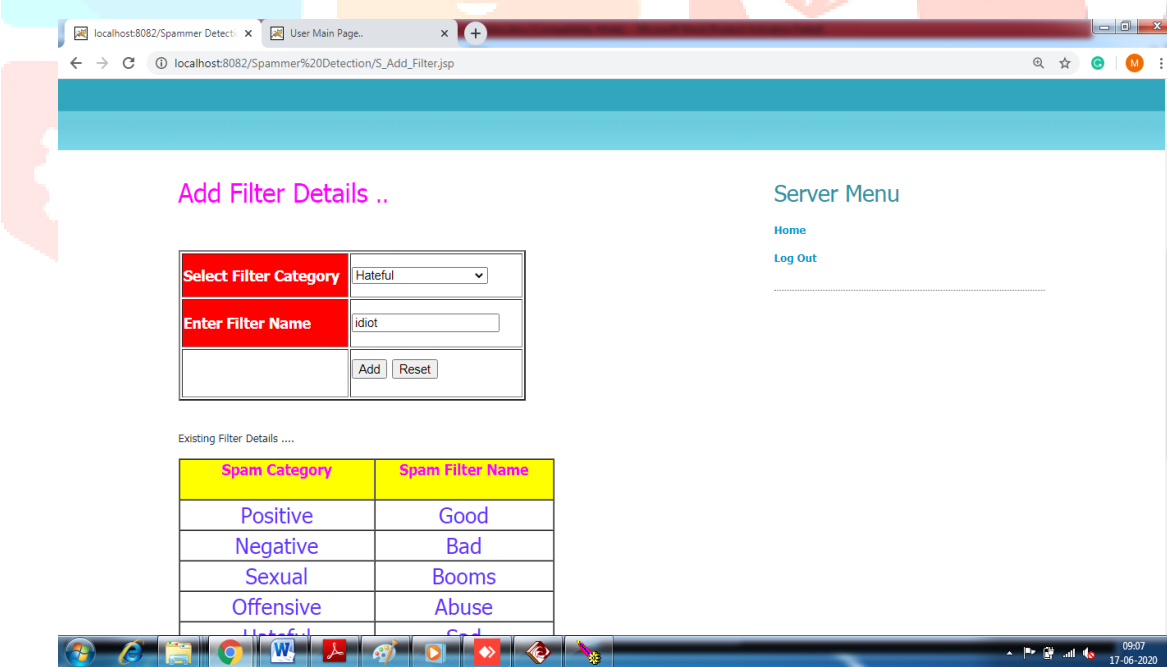
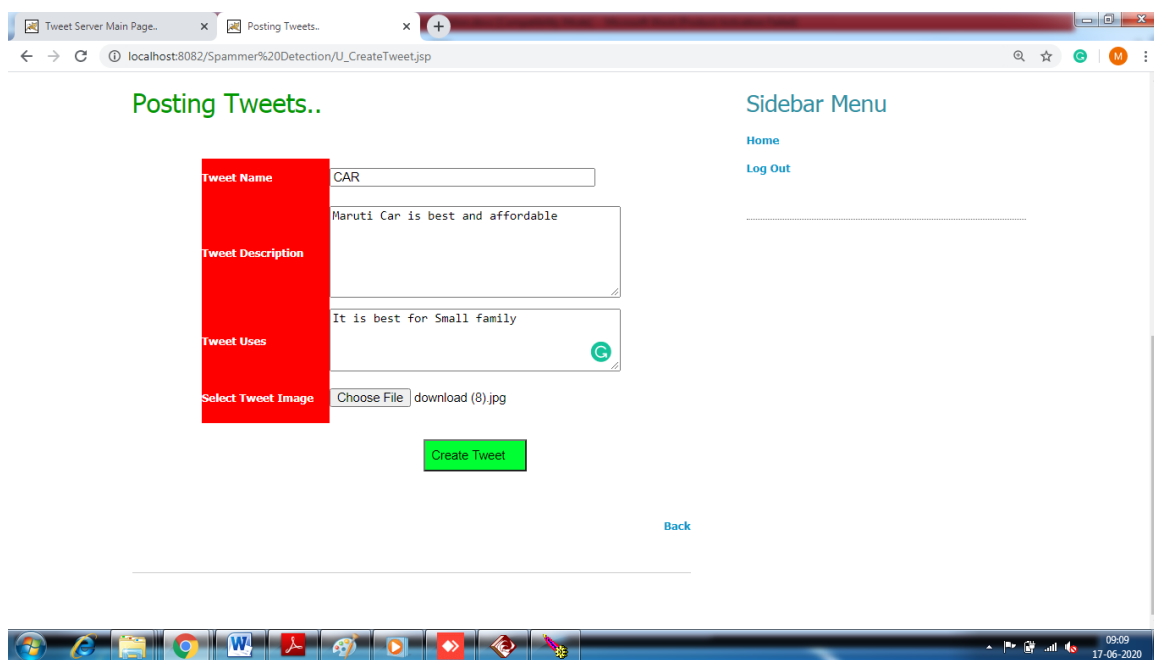


Figure . Represents the All Filter Details

9) User Post a Tweet



6. CONCLUSION

In this paper we for the first time have construct a identify the Spammer based on the reviews given by the end users based on spam keywords and also identify the fake user who tries to create duplicate tweets in social media. Here we used Random Forest Classifier inorder to identify the tweets as spam and non-spam and also to identify the fake users who try to create duplicate tweets with same number and same URL. By conducting various experiments on our proposed model we finally came to a conclusion that our proposed approach is best in finding the solution to block spammers and fake users to misuse the social media.

7. REFERENCES

- [1] A well-known authors, N. Jindal and B. Liu, has written a paper on “Opinion spam and analysis””, published in WSDM, 2008.
- [2] A well-known authors, Ch. Xu and J. Zhang, has written a paper on “Combating product review spam campaigns via multiple heterogeneous pairwise features”, published in SIAM International Conference on Data Mining, 2014.
- [3] A well-known authors, F. Li, M. Huang, Y. Yang, and X. Zhu, has written a paper on “Learning to identify review spam”, published in Proceedings of the 22nd International Joint Conference on Artificial Intelligence; IJCAI, 2011.
- [4] A well-known authors, S. Feng, R. Banerjee and Y. Choi, has written a paper on “Syntactic stylometry for deception detection”, published in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers; ACL, 2012.

- [5] A well-known authors ,N. Jindal, B. Liu, and E.-P. Lim, has written a paper on “Finding unusual review patterns using unexpected rules”, published In ACM CIKM, 2012.
- [6] A well-known authors,M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, has written a paper on “Finding deceptive opinion spam by any stretch of the imagination”, published in ACL, 2011.
- [7] A well-known authors, G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, has written a paper on “Exploiting burstiness in reviews for review spammer detection”, published in ICWSM, 2013.
- [8] A well-known authors, A. j. Minnich, N. Chavoshi, A. Mueen, S. Luan, and M. Faloutsos, has written a paper on “Trueview: Harnessing the power of multiple review sites”, published in ACM WWW, 2015.
- [9] A well-known authors, B. Viswanath, M. Ahmad Bashir, M. Crovella, S. Guah, K. P. Gummadi, B. Krishnamurthy, and A. Mislove, has written a paper on “Towards detecting anomalous user behavior in online social networks”, published in USENIX, 2014.
- [10] A well-known authors, H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, has written a paper on “Spotting fake reviews via collective PU learning ”, published in ICDM, 2014.
- [11] A well-known authors, L. Akoglu, R. Chandy, and C. Faloutsos, has written a paper on” Opinion fraud detection in online reviews by network effects”, published in ICWSM, 2018.
- [12] A well-known authors, R. Shebuti and L. Akoglu, has written a paper on” Collective opinion spam detection: bridging review networks and metadata”, published in ACM KDD, 2019.

