Detecting Malicious Twitter Bots Using Machine Learning

Tanu Satija(✉) and Nirmalya Kar

Computer Science and Engineering Department, NIT Agartala, Jirania, India
tanu0994@gmail.com, nirmalya@nita.ac.in

Abstract. Cybercrimes and phishing scams have increased multi-folds over the past few years. Now a days, hackers are coming up with new techniques to hack accounts and gain sensitive information about people and organisations. Social networking site like Twitter is one such tool. And due to its large audience hackers use such sites to reach large number of people. They use such sites to circulate malicious URLs, phishing mails etc. which serve as the entry point into the target system. And with the introduction of Twitter Bots, this work got even easier. Twitter bots can send tweets without any human intervention after a fixed regular interval of time. Also their frequency of tweets is much more than humans and therefore they are frequently used by hackers to spread malicious URLs. And due to large number of active members, these malicious URLs are reaching out to more people, therefore increasing the phishing scams and frauds. So this paper proposes a model which will use different algorithms of machine learning, first to detect twitter bots and then find out which of them is posting malicious URLs. In the proposed model, some features have been suggested which distinguishes a twitter bot account from a benign account. Based on those statistical features, model will be trained. The model will help us to filter out the malicious bots which are harmful for legitimate users.

Keywords: Malicious bots · Twitter bots · Twitter mining · Malicious URL

1 Introduction

Twitter is a very famous social networking service where users from all over the world post and interact with messages which are known as "tweets". Founded in 2006 by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams [1] twitter has seen some remarkable growth over the years. According to statistics, twitter had around 30 million users which shoot up to 335 million active users in the 2nd quarter of 2018 (Fig. 1) [2].

Twitter has become one of the most used social media platforms because it is both personal and rapid. It gives people from different spheres of lives to express themselves and build relationship with their followers.

Different news channels have twitter accounts to share latest news and developments. Celebrities use Twitter to build a personal connection with their fans.

A. Saha et al. (Eds.): ICCISIoT 2019, CCIS 1192, pp. 182–194, 2020.
https://doi.org/10.1007/978-981-15-3666-3_16

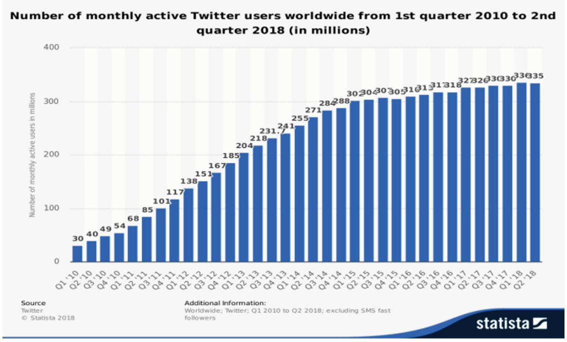

Fig. 1. Comparison between active users from 2010 to 2018 on Twitter [2]

Twitter is being used by many brands to market their products and get feedback from people about their brand. It helps them to improve their product. People with same taste and interest form communities on twitter and share their ideas and hold discussions online. Also people share interesting photographs, new researches. Authors use twitters to promote their books and the list is endless.

Another interesting feature of twitter is bots. A bot software which controls a twitter account with the help of Twitter Application program interface is called Twitter bots [3]. It is like the automation of a twitter account. Automations rules are set by twitter only. These bots can tweet, re-tweet, like, follow, unfollow, or direct message any accounts without any human intervention. According to study conducted by University of Southern California and Indiana University around 15% of the active twitter accounts are bots [4]. These bots behave just as humans and therefore it is difficult to identify them. These bots chats with random users or posts poetry, news or photographs at regular intervals. Some of the famous bots are:

1. @HundredZeros: Twitterbot that posts URLs to the eBooks that are freely available on Amazon.
2. @DearAssistant: Gives detailed response to any question asked quickly.
3. @netflix_bot: Tweets about new shows on Netflix.
4. @nicetipsbot: Twitter bot which posts fun little life tweets multiple times a day and many more [5].

Owing to the popularity of twitter, its huge active audience and such interesting features, it also grabs the attention of malicious and unethical users like hackers, cyber criminals, cyber bullies. There are many malicious and political bots on twitter which post sensitive contents and malicious URLs. Cyber criminals use these bots to circulate malicious and Phishing URLs through automation and regular intervals in order to gain sensitive information of various people leading to phishing scams and frauds. Twitter has seen an increase in the number of malicious bots in recent years that have distributed fake news and distorted images, malicious URLs.

These malicious twitter bots regularly post malicious URLs, fake news and try to reach as many people as possible. These malicious URLs can be of some malicious website or it can be a part of phishing scam asking for some person sensitive information. These bots serve as a very helpful tool for hackers to carry out data breaches. Since 2013 till today, total data breaches recorded is 14,717,618,286 [6]. Twitter has been trying to identify such accounts but due to their human like behaviour, it is difficult to identify them and filter them out (Fig. 2).

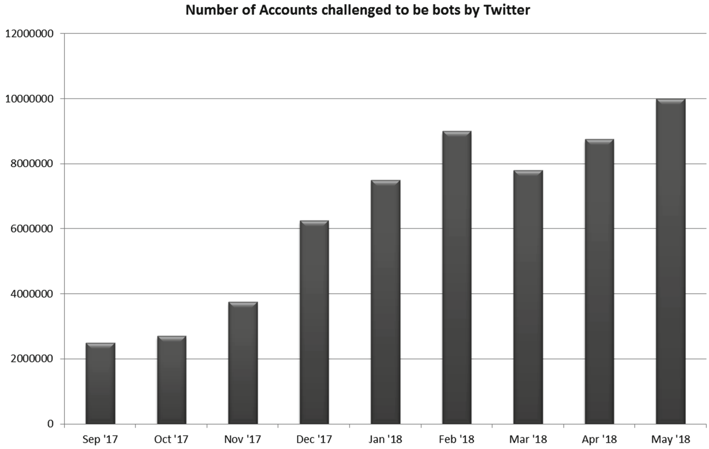

Fig. 2. Number of accounts challenged to be bots by twitter from Sept 2017 to May 2018 [7]

In order to detect such malicious bots, machine learning can be considered to be a useful aid. Machine learning is a part of artificial intelligence. It works on the idea that system can be trained to learn and identify patterns and then take decisions based on that knowledge with very less human intervention [8]. There are like millions of twitter bots and identifying them manually is nearly impossible. But with machine learning algorithms, we can make models which will automatically analyse the twitter accounts and help in identifying twitter bots. So in this paper, a model is being proposed which will, with the help of machine learning algorithms, filter out those twitter accounts which have a bot like activity and are posting malicious URLs.

2 Background

Machine learning is considered to be a fast evolving branch which has increase the pace of automation. Machine learning algorithms like logistic regression, SVM helps in handling and understanding large amounts of data. These algorithms help use to develop models which can identify patterns and take decisions with less human involvement [9]. There are several machine learning algorithms which are used to train models according to our requirements like linear regression is used when we have one dependent and one independent variable [10]. Logistic regression is used when more than one independent variable is there [11].

In order to implement machine learning algorithms and train our models, we need large amount of dataset. In this paper, we are taking datasets to train our model from Kaggle [12]. It is an online community which provides many different kinds of datasets in different formats which can be used by machine learners and data scientist. It allows its members to use published dataset to build different machine learning models. Users can even publish their own dataset on this platform which can be used by others.

Since machine learning is powerful and useful tool, researchers have been trying to incorporate it with security. Lee, Sangho, and Jong Kim have proposed a suspicious URL detection system for Twitter, WARNINGBIRD. They have considered correlated redirect chains of URLs in a number of tweets. As hackers have fewer resources and therefore they will have to use them again , a portion of their redirect chains will be shared. It's a real time system. WARNINGBIRD consists of 4 modules: data collection, feature extraction, training, and classification. Data was collected from the collection of tweets with URLs and crawling for URL redirections. From collected data, useful and unique features were extracted which were used for training the classifier to identify malicious URLs. In WARNINGBIRD 12 features has been identified for classifying suspicious URLs on Twitter. WARNINGBIRD uses a static crawler implemented in Python [13]. In a survey paper published in IJRASET, authors have used the WARNINGBIRD mechanism and implemented it and have given the results [14].

Short URLs (Uniform Resource Locators) are now a days very popular in Social media networks but at the same time detecting whether they are malicious or not becomes even difficult. Raj Nepali, et.al. proposed a mechanism using machine learning to develop a classifier to detect malicious short URLs with visible content features, tweet context, and social features from Twitter. The system will be implemented as Firefox's web browser extension and will be programmed with XML User Interface Language and Javascript to automate the processes. The system will fetch tweets on its own and extract the features from the data and for classification submit it to the classifier. The result will be displayed to the user [15].

Another paper published in 2017, provides basic understanding of how machine learning can be used to detect malicious URLs and presents information about the related work which has been done in this domain [16]. Nikan chavoshi et.al. proposed a warped correlation finder to identify correlated user accounts in social media websites such as Twitter. The authors suggested that humans are not highly synchronous for a long duration, therefore highly synchronous accounts are most likely to be bots. The proposed model works on activity correlation and does not require a labeled data set as compared to twitter suspension technique and per-user technique [17]. Authors Novotny and Jan proposed a machine learning approach to distinguish between sophisticate and less sophisticated twitter bots. They defined 4 categories in which twitter accounts can be separated: social bots, traditional spambots, and fake followers and actual human accounts. They used 4 machine learning methods to train different models and showed a comparison between them. As per their work,random forest performs slightly better than the rest in all performance measures used [18]. Clayton A. Davis et al. proposed a supervised machine learning based model to identify twitter bots. Since it is a supervised learning based model therefore authors used Twitter's REST API to extract data from twitter and converted into labelled training data. The features taken into consideration in the proposed model were divided into 6 main classes: Network features, User features, Friends features, Temporal Features, Content Features, Sentiment features [19]. Zhouhan Chen et al. proposed an approach to detect malicious bot groups on Twitter using features like the use of URL shortening services by account, duplicate tweets and content coordination between accounts over extended periods of time. Since the approached proposed is unsupervised, therefore there is no need for labelled data for training [20].

3 Proposed Model and Discussion

There are millions of bots operating on twitter and not all of them are malicious. The goal of this paper is to develop a system which is able to detect those twitter bots which are posting malicious URLs. The proposed system, called Find_Malicious_Bot can be divided into 3 modules. Figure 3 gives a brief about the 3 modules.

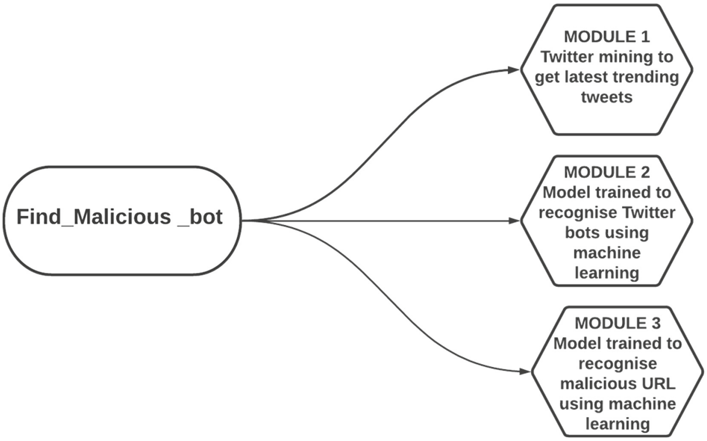
Find_Malicious _bot
MODULE 1
Twitter mining to get latest trending tweets
MODULE 2
Model trained to recognise Twitter bots using machine learning
MODULE 3
Model trained to recognise malicious URL using machine learning

Fig. 3. Proposed system Find_Malicious_Bot with its 3 modules

Following are the details about each module:

3.1 MODULE 1: Mining Twitter Data to Extract Tweets

Under this module a Twitter API is created and trending tweets are extracted according to the required location using WOEID. A WOEID (Where On Earth IDentifier) is a unique 32-bit reference identifier, assigned by Yahoo!, that identifies the place on Earth [21,22]. In this paper,tweets related to India and New York were collected from Twitter using WOEID of India : 23424848 and that of New York is 2459115.

For over a week, top 1000 treading tweets were collected daily using both WOEID, resulting in a database of 7000 tweets for each. From all those tweets,tweets containing any kind of URL were filtered out and all the details of those tweets were saved in a separate database. And the URLs present in the tweets were extracted and saved in a separate database. Details extracted from the tweets are given in Table 1.

Table 1. Details extracted from tweets

Details extracted from Tweets
Id
id_str
screen_name
created_at
Location
Description
url
followers_count
listed_count
friends_count
favourites_count
Verified
statuses_count
Lang
Status
default_profile
default_profile_image
has_extended_profile
Name

Proposed Algorithm

```
Step 1: Create Twitter API
Step 2: WOEID = 23424975 for India
Step 3: WOEID1 = 2459115 for New York
Step 4: Extract tweets using trends_place(woeid) function for India and New York
Step 5: Now from Extracted tweets for India
    Search for tweets which contain any URLs
    If URL present in tweet
    Extract all the details mentioned in Table 1 from tweet
Step 6: Save it in a csv file.
Step 7: Also save the URL in another file with twitter id of that
Step 8: Function to identify URL in tweet text
def Find(string):
    $url = re.findall('http[s]? : //(? : [a - zA - Z]||[0 - 9]||[\$-$
    $\_@.\&+]||[!*, ]||(? : \%[0 - 9a - fA - F][0 - 9a - fA - F]))+', string)$
    return url
```

The trending tweets were collected from Twitter because bots post tweets very frequently (sometimes 700 times in a day) as compared to humans and therefore those tweets become a trend. As per the tweets extracted, tweets from India have more URLs as compared to that of New York.

3.2 MODULE 2: Model Trained to Recognize Twitter Bots Using Machine Learning

In order to identify bot like behaviour several factors need to be considered. There are three most important factors which can help to identify any bot [23] (Fig. 4).

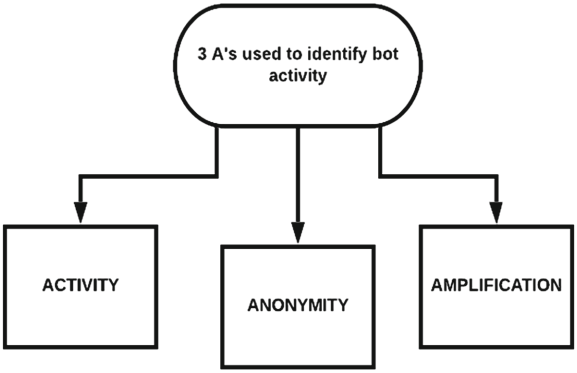

Fig. 4. The 3 A's to identify the bot like activity

3.2.1 Activity

Since bots are automated software, so the frequency of their tweets is much more as compared to humans. There are some bots which are posting tweets every minute. This is not human pattern of behaviour. Therefore by tracking the number of tweets posted by account, bot like activity can be detected. For example: Netflix Bot(@netflix_bot). It will post a new tweet every minute about some new show, gives updates about going on shows and posts URLs of the same whereas a legitimate human account will post 2–3 tweets in a day.

3.2.2 Anonymity

Another factor is the degree of anonymity maintained by account. Usually the less personal information the account gives, the more likely it is to be a bot. If it will be a legitimate human account, then it will have personal information like reading their education, jobs and profile pictures whereas bots won't have such information. For example: Museum Bot(@MuseumBot). The profile of this bot doesn't contain any profile photo.

3.2.3 Amplification

Bots tend to boost the signal from other users by retweeting, liking or quoting them. Therefore, if timeline of an account consists of a procession of retweets and word-for-word quotes of news headlines and very few or no original posts, then there is a very strong possibility that it is a bot. For example: Dear Assistant(@DearAssistant). The timeline of this bot will be filled with retweets and replies to the questions asked. It contains very less original posts. Another example: Museum Bot(@MuseumBot). Its timeline contains URLs and images from museum always. It will not post anything personal.

Keeping above three points in mind and using the details extracted from twitter a model is trained to identify whether the twitter account is a bot or not. The machine learning algorithm used is Bag of Words. This algorithm counts how many times a word appears in a document. Then these counts are used in comparison of documents and find their similarities. It can be used in application like searching, document classification and topic modelling [24]. This algorithm is used to extract features from text and then those features are used for modelling purpose. In this algorithm, occurrence of each word is used as a feature for training a classifier.

Proposed algorithm

Step 1: Dataset about twitter bots is taken from Kaggle and split in 75:25
Step 2: Using 75% of data model is trained on given parameters:

1. Whether account is verified: if verified chances of being bot being malicious is less
2. Tweets name, screen name, description and status is check for some specific words like
   bot—bot—cannabis—tweet me—mishear—follow me—updates—every—gorilla—yes_ofc—forget etc. – these words are usually used in name, description etc. of bots

3. Bag of words algorithm is used to check how many times these words appear
   If frequency is more than 50 ; possibly a bot
4. Then listed_count is check. If greater than 16000; less chances of a bot
5. Check for number of retweets; if more than 10,000 and followers less than 200 ; chance of being a bot
   Like this several other features are tested and data is saved.

Step 3: Then this trained model is tested on rest 20% data and comparison is made with predicted data and original data to calculate accuracy.
Step 4: True positive rate (TPR) and false positive rate (FPR) calculated
Step 5: A ROC curve is plotted between true positive rate (TPR) and false positive rate (FPR) with threshold value of 45°.

The data set for malicious bots and legitimate human accounts is taken from Kaggle and divided in ratio of 75:25 ratios to get optimal result for training and testing purposes. The accuracy of this model is: 95.44% as per the data set used. A Receiver Operating Characteristic (ROC) curve is plotted between the true positive rate (TPR) and false positive rate (FPR) with threshold value of 45°. ROC is a curve which is dependent on probability and tells how well model can distinguish between two classes. Figure 5 shows the ROC plot for the model trained. Area under curve is near to 1 which means model has a good measure of separability Equations to calculate TPR and FPR:

$$TPR = TP/P = TP/(TP + FN) = 1 - FNR \tag{1}$$

$$FPR = FP/N = FP/(FP + TN) = 1 - TNR \tag{2}$$

Where

condition positive (P): the number of real positive cases in the data
condition negative (N): the number of real negative cases in the data
true positive (TP): hit
true negative (TN): correct rejection
false positive (FP): false alarm
false negative (FN): miss
false negative rate: FNR
true negative Rate: TNR [25]

3.3 MODULE 3: Model Trained to Recognize Malicious URLS Using Machine Learning

Under this module, a model is trained using machine learning for detecting whether the URL is malicious or not. The machine learning algorithm used is logistic regression and dataset is taken from Kaggle. Logistic regression is algorithm which is used when there is only one dependent variable but more than one independent variable. The value of this dependent variable is calculated using all these independent variables. It is used in this model because here there

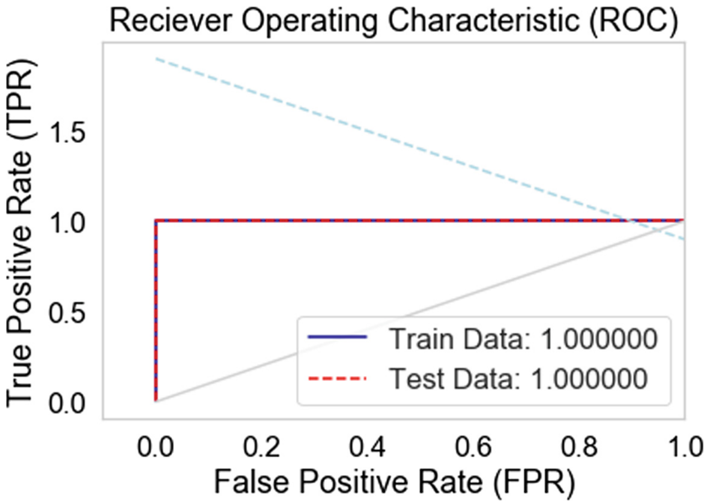

Fig. 5. ROC curve for between TPR and FPR

is only one dependent variable that is, whether the twitter account is bot or not, which is dependent on many other parameters like twitter account name, when it was created, account's status, number of followers etc. which are all independent of each other.

Proposed Algorithm

Step 1: Dataset from Kaggle is taken
Step 2: From sklearn.linear_model package import LogisticRegression
Step 3: Data from CSV file is split into 80:20 ratio for training and testing
Step 4: Model is trained using features from csv file.
Step 5: Then the trained model is tested on the 20% of data.
Step 6: Then score is calculated based on test result which tells the accuracy of model.

The data is divided in ratios of 80:20 in order to get optimal result. If the dataset is divided in 60:40 ratio, then accuracy was coming to be 80.05%. Therefore, Eighty percent of the data is used for training the model and twenty percent of data for testing the accuracy of the model. As per the dataset, the model is able to achieve 93% (approx.) of accuracy. It is a static model just like WARNINGBIRD [13]. It works on the predefined dataset only. Figure 6 shows the complete flow of the proposed system and how all modules shall work together.

Figures 7 and 8 shows the details about Module 2 and Module 3.

The proposed system can help in identifying twitter bots which are posting malicious URLs and report them to twitter. Also it can help in reducing APT attacks, phishing attacks etc. as the entry point for all these attacks is usually the malicious URL circulated either through social media or emails. Once if the source of these malicious URLs are detected then these attacks can be prevented.

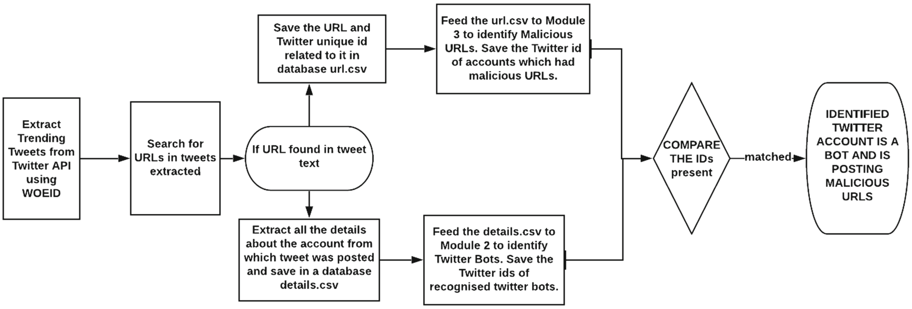
Extract Trending Tweets from Twitter API using WOEID
Search for URLs in tweets extracted
If URL found in tweet text
Save the URL and Twitter unique id related to it in database url.csv
Feed the url.csv to Module 3 to identify Malicious URLs. Save the Twitter id of accounts which had malicious URLs.
Extract all the details about the account from which tweet was posted and save in a database details.csv
Feed the details.csv to Module 2 to identify Twitter Bots. Save the Twitter ids of recognised twitter bots.
COMPARE THE IDs present
matched
IDENTIFIED TWITTER ACCOUNT IS A BOT AND IS POSTING MALICIOUS URLS

Fig. 6. Flow of proposed system

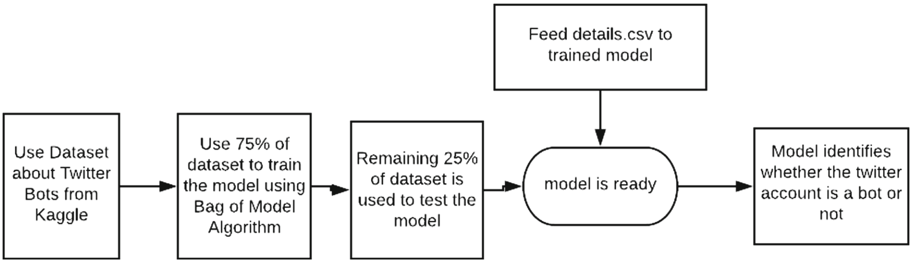
Feed details.csv to trained model
Use Dataset about Twitter Bots from Kaggle
Use 75% of dataset to train the model using Bag of Model Algorithm
Remaining 25% of dataset is used to test the model
model is ready
Model identifies whether the twitter account is a bot or not

Fig. 7. Module 2 of proposed system

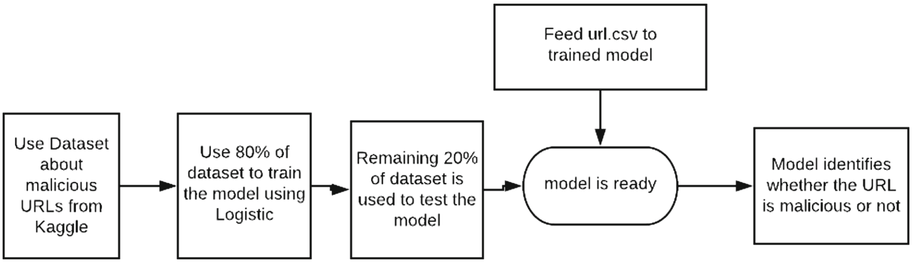
Feed url.csv to trained model
Use Dataset about malicious URLs from Kaggle
Use 80% of dataset to train the model using Logistic
Remaining 20% of dataset is used to test the model
model is ready
Model identifies whether the URL is malicious or not

Fig. 8. Module 3 of proposed system

4 Conclusion

In present day scenario, twitter bots can be very dangerous if they are broadcasting malicious URLs. They can very easily reach out to many people and can lead to many data breaches and phishing scams. So in this paper a model is proposed, which by using machine learning algorithms can successfully identify twitter bots which are posting malicious URLs. The proposed model has been tested and so far it is achieving good accuracy. It is able to distinguish between twitter bots and legitimate human accounts. Also, the model is able to detect malicious URLs. The proposed system can be used to prevent attacks where the entry point is malicious URL. There is only one limitation. Proposed model is trained to work with long URLs for now. But now a days these social networking sites are using URL shortening facility which can reduce the URL to minimum size possible. So in future, another module will be added to this system which will be able to work on short URLS and detect whether they are malicious or not.

References

1. https://en.wikipedia.org/wiki/Twitter
2. https://medium.com/the-startup-growth/how-many-people-are-on-twitter-662d797d5958
3. Mønsted, B., Sapieżyński, P., Ferrara, E., Lehmann, S.: Evidence of complex contagion of information in social media: an experiment using Twitter bots. PLoS ONE **12**(9), e0184148 (2017)
4. https://beebom.com/best-twitter-bots/
5. https://www.welivesecurity.com/2018/06/28/twitter-bots-disassemble
6. https://breachlevelindex.com/
7. Data from Twitter blog. https://blog.twitter.com/
8. Chang, S., Cohen, T., Ostdiek, B.: What is the machine learning? Phys. Rev. D **97**(5), 056009 (2018)
9. Marx, V.: Machine learning, practically speaking. Nat. Meth. **16**(6), 463 (2019)
10. Kan, H.J., Kharrazi, H., Chang, H.-Y., Bodycombe, D., Lemke, K., Weiner, J.P.: Exploring the use of machine learning for risk adjustment: a comparison of standard and penalized linear regression models in predicting health care costs in older adults. PLoS ONE **14**(3), e0213258 (2019)
11. Kuha, J., Mills, C.: On group comparisons with logistic regression models. Sociol. Meth. Res. **47**(1), 0049124117747306 (2018)
12. Kainkaryam, S., Ong, C., Sen, S., Sharma, A.: Crowdsourcing salt model building: Kaggle-TGS salt identification challenge. In: 81st EAGE Conference and Exhibition 2019 (2019)
13. Lee, S., Kim, J.: WarningBird: detecting suspicious URLs in Twitter stream. NDSS **12**, 1–13 (2012)
14. Chaudhary, M., Hingoliwala, H.A.: Warning Tweet: a detection system for suspicious URLs in Twitter stream. Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET) **2**, 297–305 (2014). ISSN: 2321–9653
15. Alshboul, Y., Nepali, R., Wang, Y.: Detecting malicious short URLs on Twitter (2015)
16. Sahoo, D., Liu, C., Hoi, S.C.H.: Malicious URL detection using machine learning: a survey. arXiv preprint arXiv:1701.07179 (2017)
17. Chavoshi, N., Hamooni, H., Mueen, A.: DeBot: Twitter bot detection via warped correlation. In: ICDM, pp. 817–822 (2016)
18. Novotny, J.: Twitter bot detection & categorization-a comparative study of machine learning methods (2019)
19. Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: BotOrNot: a system to evaluate social bots. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 273–274. International World Wide Web Conferences Steering Committee (2016)

20. Chen, Z., Tanash, R.S., Stoll, R., Subramanian, D.: Hunting malicious bots on Twitter: an unsupervised approach. In: Ciampaglia, G.L., Mashhadi, A., Yasseri, T. (eds.) SocInfo 2017. LNCS, vol. 10540, pp. 501–510. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67256-4_40
21. Devadoss, A.K.V., Thirulokachander, V.R., Devadoss, A.K.V.: Efficient daily news platform generation using natural language processing. Int. J. Inf. Technol. **11**(2), 295–311 (2019)
22. Chen, B.-C., Davis, L.S.: Deep representation learning for metadata verification. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), pp. 73–82. IEEE (2019)
23. https://medium.com/dfrlab/botspot-twelve-ways-to-spot-a-bot-aedc7d9c110c
24. Zeng, J., Liu, M., Xiang, F., Ruiyu, G., Leng, L.: Curvature bag of words model for shape recognition. IEEE Access **7**, 57163–57171 (2019)
25. https://en.wikipedia.org/wiki/Receiver_operating_characteristic