Detection Of Bots In Twitter Network Using Machine Learning Algorithm

Rajnish K. Prince[1], Snehal S. Thube[2] , Rahul Ranjan[3], Akash L.Sakat[4] ,Ms. V. A.Yaduvanshi[5]

UG student, Department of Electronics and Telecommunications, SKNCOE, SPPU, Pune Assistant Professor, Department of Electronics and Telecommunications, SKNCOE, SPPU, Pune

[1]Rajnishkumarprince90@gmail.com [2]snehal.thube_skncoe@sinhgad.edu [3]Rahulrsd1234@gmail.com, [4]Akashsakat60@gmail.com [5]vijaya.yaduvanshi_skncoe@sinhgad.edu

Abstract— Malicious social bots are a very common issue in online social networks. These malicious social bots are being used for a many of purposes such as artificially increasing the popularity of a people or movement,manipulating financial markets, amplifying phishing attacks, spreading spam, and shutting down free speech. Therefor detection of these bots in online social networks is of great importance. Social media platforms are not able to apply maximum stringent requirements for account creation because for a variety of reasons such as it may prevent some legitimate users from signing up, it will lack the ability to maintain some anonymity for protestors under oppressive regimes, it may cause inconvenience to real users. So another machine learning algorithms were used to find these malicious social bots. In proposed system we are using various machine learning algorithm such as Random Forest (RF) and Support Vector Machine (SVM).

Keywords— *SVM, Machine Learning, Malicious, Twitter Dataset.*

I. INTRODUCTION

Social media has played a more important role in our daily life. With billions of users producing and consuming information every day, it is a natural extension that people turn to this medium to read and disseminate news. Social media bots are programs that change in size depending on their function, capability, and layout and can be used on social media platforms to do different useful and malicious tasks while stimulating human behaviour. Some social media bots provide needful services, such as climate updates and match score. With so many people turning to social media, malicious users like bots have begun to sway the conversations in whatever direction their creators want .These malicious bots have been used for malicious tasks such as spreading false information about political candidates , inflating the perceived popularity of celebrities, deliberately pushing down the messages of protestors and activists, illicitly advertising by spamming the social web with links to commercial websites and influencing financial markets in an attempt to manipulate the direction of stock prices. Furthermore, these bots can change the results of common analyses performed on social media.Trend jacking -use of top trending topics to focus on an intended audience for targeting purposes, watering hole attack-attacker guesses or observes which websites an organization often uses and infects one or more of them with malware, hash tag hijacking-use of hash tags to focus an attack(e.g. spam ,malicious links)on a specific audience using the same hash tag and click farming or like farming-inflate fame or popularity on a website through liking or reposting of content via click farms. Bot detection is an important task in social media. While these methods are fast (requiring a simple database lookup), and are expected to have low False Positive rates, a major shortcoming is that they fail against newly generated URLs. This is a severe limitation as new URLs are generated every day. To address these limitations, there have been many attempts to resolve this problem through the help of machine learning.

II. RELATED WORK

Deep Neural Networks for Bot Detection:-

Proposed a deep neural network based on contextual LSTM (Long ShortTerm Memory) architecture allowing the use of both tweet data and metadata to find bots at the tweet level. The contextual features are extracted from user metadata and fed as auxiliary input to LSTM deep nets processing the tweet text.

By using supervised learning algorithms to detect doubtful URLs in online social networks:-

Proposed a supervised machine learning classification model to detect the distribution of malicious content in online social networks (ONSs).A random forest classification was used with a combination of features derived from a area of sources. The random forest model without any tuning and feature selection produced a recall value of 0.89.

To examining social bots on Twitter from big data approach:-

Proposed System presented an important findings on social bots from an econometric analysis of the weekly panel data set. Twitter is a viable platform to study social bots and big data of user-generated content, further insights from other platforms, such as Facebook, will broaden our understanding of how social bots can impact information quality and virality. Second,

sentiment analysis is a helpful tool for automatically classifying textual big data, but it has some inherent limitations due to the complexities and intricacies of human language.

III. METHODOLOGY

SVM ARCHITECTURE:-

Support vector machines are an part of supervised learning algorithms which related to both the regression and classification types of machine learning algorithms.SVM is a collection of machine learning algorithms that can be used to recognize patterns in given data. Given A set of training data it would like to classify. A classification task usually involves separating data into training and testing sets. It has been employed in a wide area of actual world problems such as text differentiation, hand-written, digit detection, tone recognition, image sorting and object detection, micro-array gene expression data analysis, data classification.

RF ALGORITHM:-

Random forest is the ensemble differentiator, which collects the output of many decision trees by majority vote . In ensemble learning, the results of multiple classifiers are brought together, and a single decision is made on behalf of the community. Each decision tree in the forest is created by selecting different samples from the original data set using the bootstrap technique . Then, the decisions taken by many different individual trees are subject to voting and present the class with the more number of votes as the class estimate of the committee. In the RF method, trees are created by CART (classification and regression trees) algorithms and boot bagging combination method. The data set is divided into training and test data. From the training data set, samples are selected as bootstrap (resampled and sampled) technique, which will form trees (in a bag) and data that will not build trees (out of the bag).

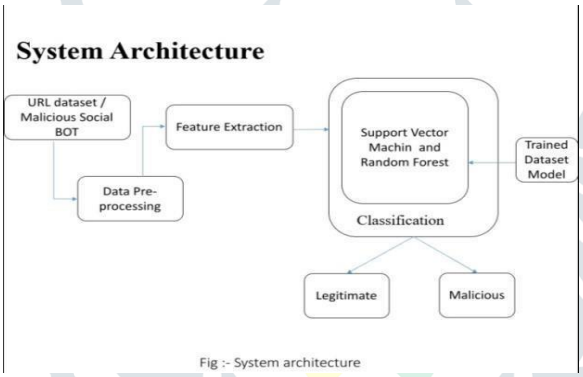

Fig. Block Diagram

The connectivity is done using MySQL database. And proposed model is website based applications and it is build using php and css as front end and for back end python is used.We are Collecting various MALICIOUS SOCIAL BOT dataset from twitter social media. Once it's collected it's divided into 80% for training and 20%for testing.The proposed system undergoes some techniques such as Preprocessing

Feature Extraction Classification:-

Data preprocessing: - It is a technique used in data mining that involves transforming raw data into an understandable format. The data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data. It is contains missing value of dataset is cleaned, and decimal values are converted into exact float values.

Data splitting: - The current dataset is divide into two, training data set and testing data set. The data splitting is done in an 80 and 20 ratio. 80% of the dataset is taken as the Training Set which is used to train the model. The remaining 20% becomes the Test Set which is used to test the model, to analyze its accuracy. The testing set is not used for training, which could otherwise lead to overfitting the mode.

Feature Selection: - The data features that used to train machine learning models have a huge influence on the performance of the model.

Classification: - The model is trained by fitting the training set to the classifier model. The classifier model upon testing, classifies the air quality into good or bad. The most of classifications are fairly related to the testing set.

Support Vector Machine (SVM):
Examples of SVM boundaries:-
Selecting best hyperplane for our classification. We will show data from two classes. The classes represented by triangle and circle.

Case 1**:** Think about the case in Fig 1, with the data is from two different classes. Now, to find out the best hyperplane which can differntiate the two classes. Please check Fig. 1. On the right to find which hyperplane best suit this use case. In SVM, we try to maximize the distance between hyperplane & nearest data point. This is known as margin. Since 1st decision border is maximizing the distance between classes on left and right. So, our maximum margin hyperplane will be 1st.

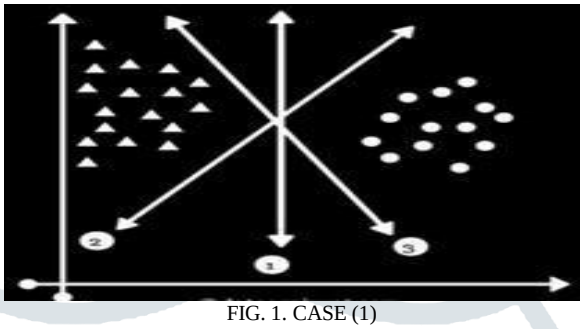

FIG. 1. CASE (1)

Case 2: - Consider the case in Fig 2, with data from two different classes. Now, to find the best hyperplane which can separate the two classes. As data of each class is distributed either on left or right. The aim is to select hyperplane which can separate the classes with maximum margin. In this case, all the decision boundaries are separating classes but only 1st decision boundaryis showing.

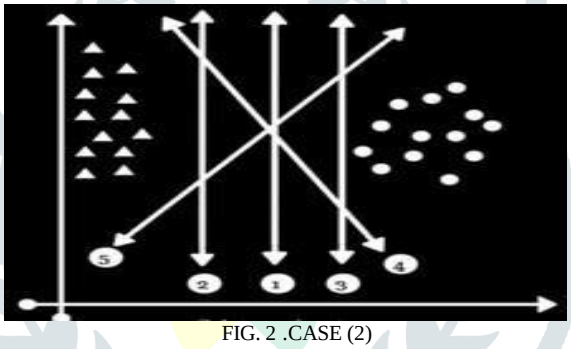

FIG. 2 .CASE (2)

Case 3: - Consider the case in Fig 3, with data from two different classes. Now, we wish to find the best hyperplane which can separate the two classes. Data is not evenly distributed on left and right. Some of the are on right too. You may feel we can ignore the two data points above 3rd hyperplane but that would be incorrect. SVM tries to find out maximum margin hyperplane but gives first priority to correct classification. 1st decision boundary is separating some from but not all. It's not even showing good margin. 2nd decision boundary is separating the data points similar to 1st boundary but here margin between boundary and data points is larger than the previous case. 3rd decision boundary is separating all from all classes. So, SVM will select 3rd hyperplane.

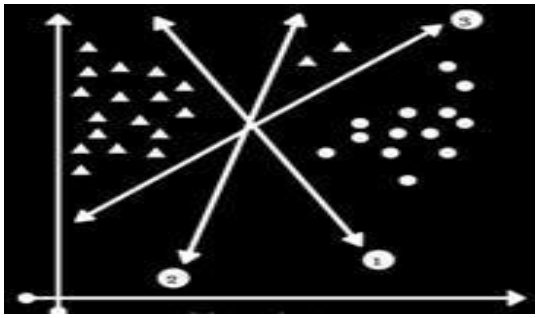

FIG. 3. CASE (3)

Case 4: - Consider the figure 4, we will learn about outliers in SVM. We wish to find the best hyperplane which can separate the two classes. Data is not evenly distributed on left and right. Some of the are on right too. In the real world, you may find few values that correspond to extreme cases i.e., exceptions. These exceptions are known as Outliers. SVM have the capability to detect and ignore outliers. In the image, 2 are in between the group of. These are outliers. While selecting hyperplane, SVM will

automatically ignore these and select best-performing hyperplane.1st & 2nd decision boundaries are separating classes but 1st decision boundary shows maximum margin in between boundary and support vectors.

Case 5**: -** We will learn about non-linear classifiers.
Please check the figure 5 on right. It's showing that data can't be separated by any straight line, i.e, data is not linearly separable.

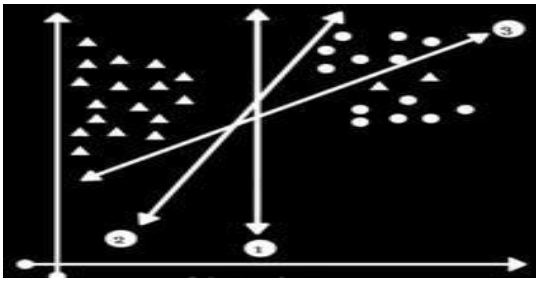

FIG. 4. CASE (4)

SVM possess the option of using Non- Linear classifier. We can use different types of kernels like Radial Basis Function Kernel, Polynomial kernel etc. We have shown a decision boundary separating both the classes. This decision boundary resembles a parabola.

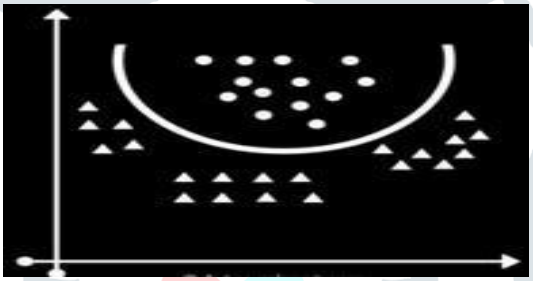

FIG. 5.CASE (5)

IV. RESULTS AND DISCUSSIONS

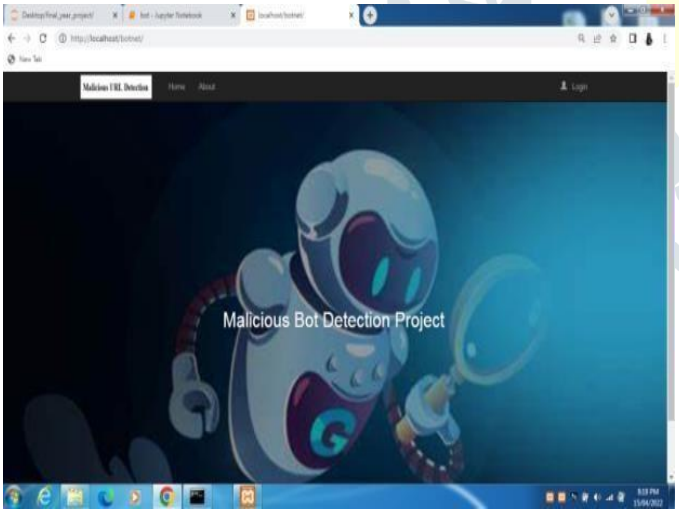

FIG. 6 . MALICIOUS BOT DETECTION WINDOW

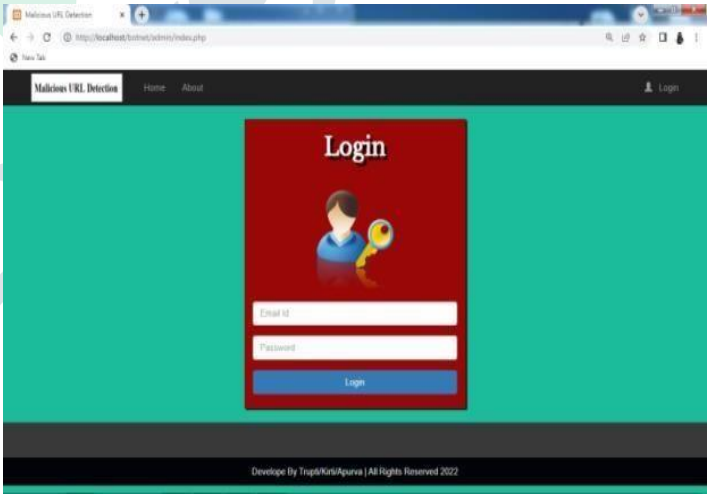

FIG. 7.LOGIN WINDOW

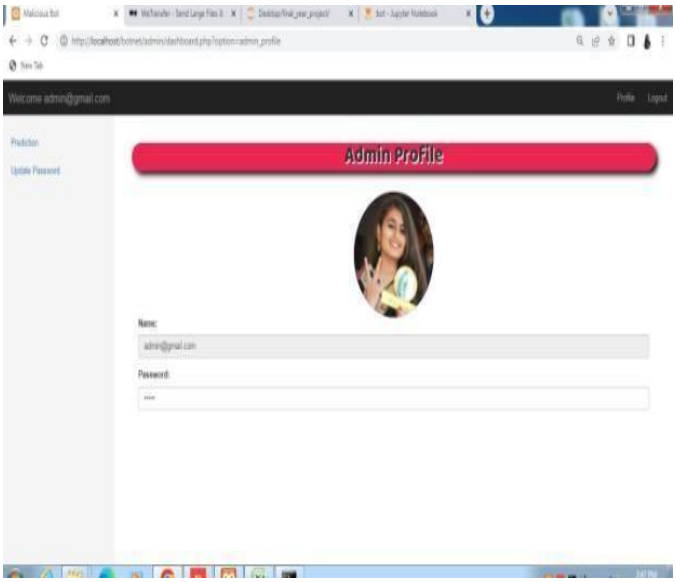

FIG. 8 .ADMIN PROFILE WINDOW

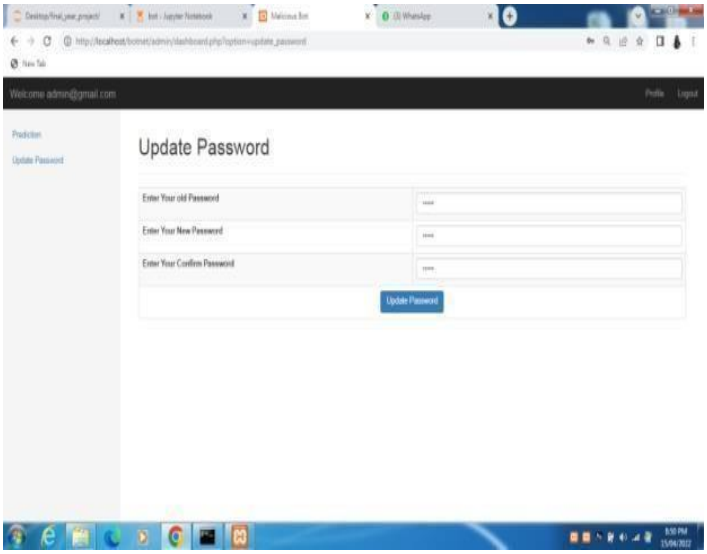

FIG. 9.PASSWORD UPDATION WINDOW

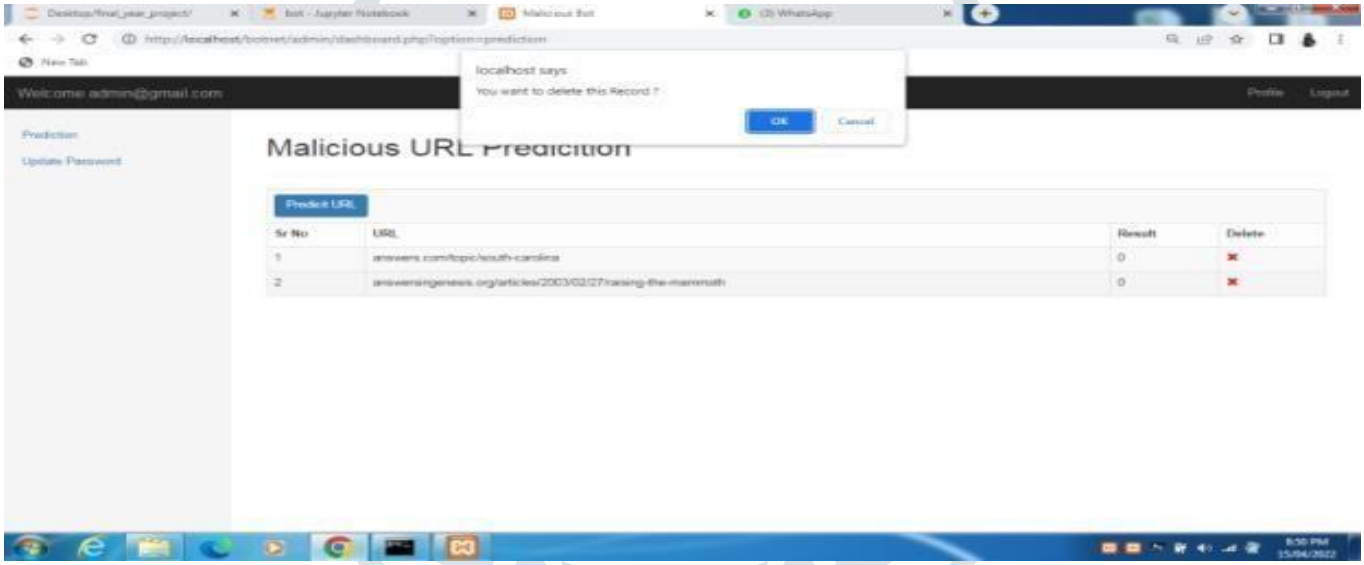

FIG.10.MALICIOUS URL PREDICTION RESULT

V. CONCLUSION

The prevalence of detecting malicious bots on social media platforms such as Twitter, the need for improved, inexpensive Bot detection methods is apparent. We proposed a Support vector Machine (SVM) and Random Forest (RF) algorithm allowing us to detect the tweets or url which may be malicious or harmful for users. In the proposed system till now we have downloaded and installed all the software which are required for system. The dataset has been collected from kaggle site and preprocessing step have been processed. In next phase the features of preprocessed data will be extracted and the algorithm will be implementedand a model will be saved which can be used for classifying the data.

ACKNOWLEDGMENT

It gives us great pleasure and satisfaction in presenting this project report on “Detection of bots in twitter network using machine learning algorithm”. We thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of [E & TC Dept] which helped us in successfully completing our project work. Also, We would like to extend our sincere esteems to all staff in laboratory for their timely support.

VII. REFERENCES

[1] Sneha Kudugunta, Emilio Ferrara,“ Deep Neural Networks For Bot Detection ”,IEEE 201

[2] Mohammed Fadhil And , Peter Andras,“ Using Supervised Machine Learning Algorithms To Detect Suspicious Urls In Online Social Networks”, IEEE 2021

[3] Xia Liu, “ A Big Data Approach To Examining Social Bots On Twitter”, IEEE 2019

[4] Sylvio Barbon Jr, Gabriel F. C. Campos,“ Detection Of Human, Legitimate Bot, And Malicious Bot In Online Social Networks Based On Wavelets ”, IEEE 2018

[5] Greeshma Lingam, Rashmi Ranjan Rout And Dvln Somayajulu,“ Detection Of Social Botnet Using A Trust Model Based On Spam Content In Twitter Network”,, IEEE 2018

[6] Chongzhen Zhang, Yanli Chen, YangMeng ,“ A Novel Framework Design of Network Intrusion DetectionBased on Machine Learning Techniques ”,IEEE 2021

[8] Linhao Luo,X,Zhang, Xiaofei Yang and Weihuang Yang,“Deepbot: A Deep Neural Network.based approach for Detecting Twitter Bots”, IEEE 2020.

[10] Peining Shi,Z.Zhang,“Detecting Malicious Social Bots Based on Clickstream Sequences”,IEEEAccess 2019

[11] Heng Ping , Sujuan Qin,“ A Social Bots Detection Model Based on Deep Learning Algorithm ”,IEEE 2018.

[12] F. Morstatter, L. Wu, T. H. Nazer, K. M. Carley, and H. Liu, ‘‘A new approach to bot detection::Striking the balance between precision and recall,’’ in Proc. IEEE/ACM Int. Conf. Adv. Anal. Mining, San Francisco, CA, USA, Aug. 2016, pp. 533–540.

[13] C. A. De Lima Salge and N. Berente, ‘‘Is that social bot behaving unethi- cally?’’ Commun. ACM,vol. 60, no. 9, pp. 29–31, Sep. 2017.

[14] M. Sahlabadi, R. C. Muniyandi, and Z. Shukur, ‘‘Detecting abnormal behavior in social network Websites by using a process mining technique,’’J. Comput. Sci., vol. 10, no. 3, pp. 393–402, 2014.

[15] F. Brito, I. Petiz, P. Salvador, A. Nogueira, and E. Rocha, ‘‘Detecting social-network bots basedon multiscale behavioral analysis,’’ in Proc.7th Int. Conf. Emerg. Secur.Inf., Syst. Technol. (SECURWARE), Barcelona,Spain, 2013, pp. 81–85.

[16] Y. Zhou et al., ‘‘ProGuard : Detecting malicious accounts in social- network-based onlinepromotions,’’ IEEE Access, vol. 5, pp. 1990–1999, 2017.