RESEARCH ARTICLE OPEN ACCESS

Tweet Based Bot Detection Using Big Data

Mr D.Purushothaman MCA.,M.E.[1], K Pavan [2]
[1]Asst. Professor, Department of Computer Applications
[2] Student, Department of Computer Applications
[1], [2] Chadalawada Ramanamma Engineering College (Autonomous)

ABSTRACT
Twitter is one of the most popular micro-blogging social media platforms that has millions of users. Due to its popularity, Twitter has been targeted by different attacks such as spreading rumors, phishing links, and malware. Tweet-based botnets represent a serious threat to users as they can launch large-scale attacks and manipulation campaigns. To deal with these threats, big data analytics techniques, particularly shallow and deep learning techniques have been leveraged in order to accurately distinguish between human accounts and tweet-based bot accounts. In this project, we discuss existing techniques, and provide a taxonomy that classifies the state-of-the-art of tweet-based bot detection techniques. We also describe the shallow and deep learning techniques for tweet-based bot detection, along with their performance results. Finally, we present and discuss the challenges and open issues in the area of tweet-based bot detection.
Keywords: - Big Data, Tweets, Bot Detection, Deep Learning, Social Media.

I. INTRODUCTION

In recent research work in the field of Twitter social botnet detection. They provided an analytical review of each proposed method with its limitations and advantages. The techniques were classified into three main categories, namely graph-based, machine learning-based, and crowd sourcing based techniques. The crowd sourcing technique uses human intelligence to identify various patterns, which is stated to be the most error prone out of the three techniques. It was also shown that machine learning methods and, more specifically, random forest classifiers are the most commonly used for detecting social bots in Twitter users. In the existing system provided a short comparative survey of the research work in the field of Twitter spam detection within the year range of 2009-2015. They described different detection methods within four categories: account based, tweet-based, graph-based, and hybrid-based methods. The account-based methods were shown to leverage the user profile's metadata like followers and following count and other derived features such as age of the account. While in graph-based methods, features like distance and strength of connectivity between users were shown to be used for spam detection. However, in tweet-based methods, the survey mainly focused on detecting spam using URL and its derived features, such as length and domain name. To detect a spam user, posted URLs were analyzed and classified as malicious or benign. Besides this, the authors highlighted overlooked features that were argued to improve the spam detection.

Another comparative survey was presented in the field of multiplatform spam user detection. The authors recognized that different platforms, such as e-mails, blogs, or microblogs, require different techniques and features to achieve accurate detection. Therefore, proposed techniques within the year range of 2011-2015 were classified based on the platform that the dataset lies within. A qualitative comparison was conducted for each group of methods under the same platform. The existing system observed that the botnet used a URL network shortening services and redirections to obfuscate the actual landing pages. They disclosed that users clicked on these URLs, found the botmaster establishing the Bursty botnet, and registering landing pages on phishing websites. They confirmed that the botmaster is still successful in owning Twitter bot-related services. This study includes a review and insight into Twitter's cyberspace infrastructure, cybercrime operation, and the dark markets. The system doesn't have technique shallow learning-based detection methods. There is no technique deep neural networks are applied on twitter data to determine the relevant content for users, and hence improve their experience on the platform .

II. RELATEDWORKS

M Imran and S Asalam Khan proposed Toward an optimal solution against denial-of-service attacks in software defined networks. Software Defined Networking (SDN) separates the control logic from data forwarding and shifts the whole decision power to the controller, making the switch a dumb device.

SDNs are becoming more and more important due to the key features like scalability, flexibility and monitoring. The centralized control of SDN makes it vulnerable to different attacks such as Flooding, Spoofing, Denial of Service (DoS), etc. These attacks can degrade the SDN performance by overwhelming its different components such as controller, switch and control channel. This project provides a comprehensive review of different mitigation approaches and categorizes them into three different classes on the basis of their methodology to handle the malicious traffic. In addition to that, we find out limitations in these mitigation approaches and propose the possible features of an optimal solution against DoS attacks.

E Karbab and A Darhab proposed MalDozer: Automatic framework for Android malware detection using deep learning. Android OS experiences a blazing popularity since the last few years. This predominant platform has established itself not only in the mobile world but also in the Internet of Things (IoT) devices. This popularity, however, comes at the expense of security, as it has become a tempting target of malicious apps. Hence, there is an increasing need for sophisticated, automatic, and portable malware detection solutions. In this project, propose MalDozer, an automatic Android malware detection and family attribution framework that relies on sequences classification using deep learning techniques. Starting from the raw sequence of the app's API method calls, MalDozer automatically extracts and learns the malicious and the benign patterns from the actual samples to detect Android malware. MalDozer can serve as a ubiquitous malware detection system that is not only deployed on servers, but also on mobile and even IoT devices.

K Yang and S Wang proposed Arming the public with artificial intelligence to counter social bots. The increased relevance of social media in our daily life has been accompanied by efforts to manipulate online conversations and opinions. Deceptive social bots — automated or semi-automated accounts designed to impersonate humans — have been successfully exploited for these kinds of abuse. Researchers have responded by developing AI tools to arm the public in the fight against social bots. Here we review the literature on different types of bots, their impact, and detection methods. use the case study of Botometer, a popular bot detection tool developed at Indiana University, to illustrate how people interact with AI countermeasures. A user experience survey suggests that bot detection has become an integral part of the social media experience for many users. However, barriers in interpreting the output of AI tools can lead to fundamental misunderstandings. The arms race between machine learning methods to develop sophisticated bots and effective countermeasures makes it necessary to update the training data and features of detection tools.

III. PROPOSED SYSTEM ARCHITECTURE

The proposed model implemented the bidirectional strategy in which tweet sentences are processed both forward and backward for each layer enabling a better understanding of the overall text context. To train the model, a public dataset Cresci-2017 is used that consists of tweets from 3,474 human accounts and 1,455 bots, resulting in 11.4 million tweets in total. Before training, each tweet was preprocessed and tokenized to fit the word embedding model. A pre-trained GloVe model was used to convert text to numerical vectors that are acceptable by the network. The vectors were fed to a three-layer model with a decreasing dropout layer that was initially set to 0.5. Two subsets of testing datasets, composed of 1,982 and 928 accounts respectively, were used to evaluate their model resulting in precision and accuracy of 93% and 95%, respectively.

In the proposed system introduced a novel deep learning model to distinguish bots from humans using their retweet patterns termed RTbust. Before building the model, the authors analyzed the behavioral patterns of bots and humans alike. The analysis demonstrated a distinctive pattern of retweeting in terms of timing, and it was categorized into four patterns. The _rst is the droplet pattern, which corresponds to normal users in which there exists a fair amount of time between the tweet being posted and the retweet operation. The three remaining patterns belonged to potential bots due to their suspicious and rapid retweeting pattern. The proposed system offered several innovative approaches that have vastly increased the efficiency of spam identification. The system is more effective due to presence of tweet-based bot detection.

In this proposed system, there are two modules. They are Tweet server and end user.

1. Tweet Server: In this Module, it Provides following functionalities like Login, data processing, Train and Test Data sets with Logistic Regression and Naïve Bayes, Generate Trained and Tested Accuracy, View Trained and Tested Accuracy Results, View Predicted Tweet Type Details, Generate tweet type ratio, log out.

2. End User: In this Module, it Provides following functionalities like register, login, post tweet dataset, Predict Tweet Type, view profile, Log out.

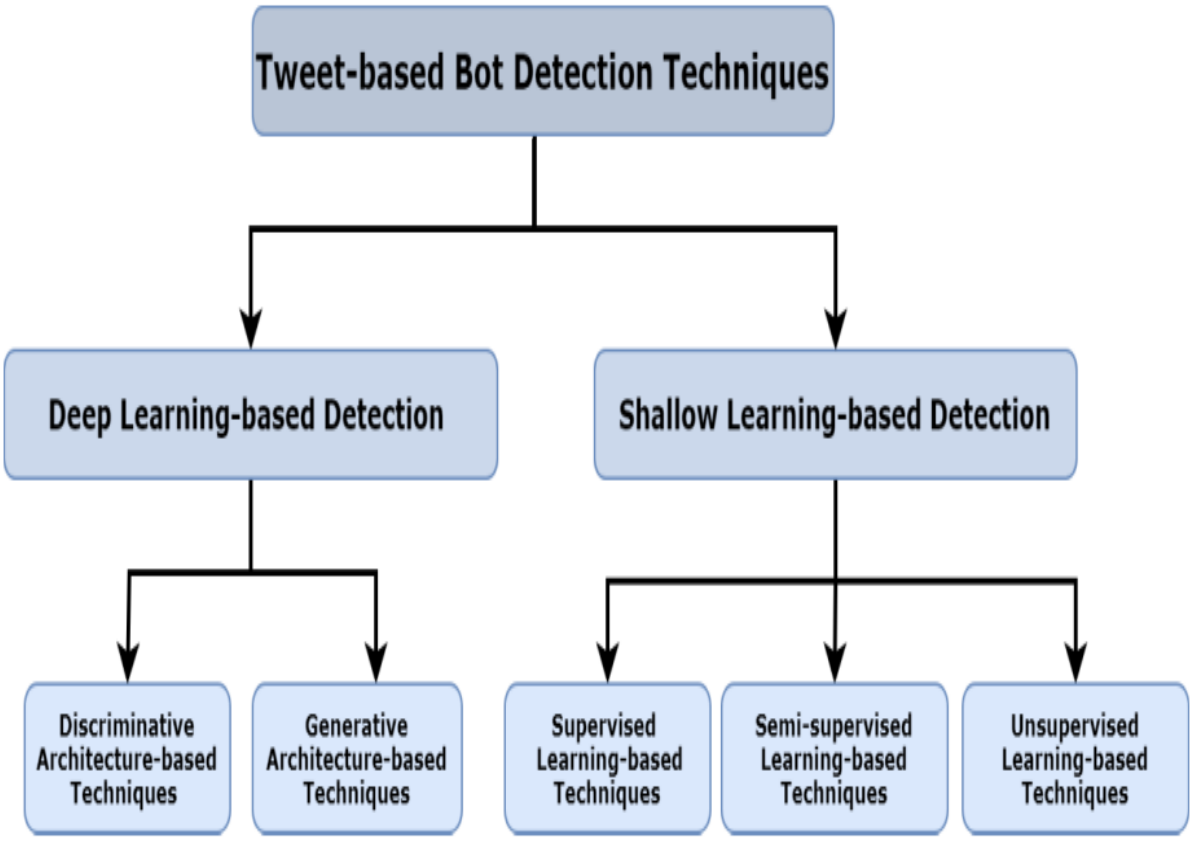

Fig. 1 Proposed system architecture

IV. RESULTS AND DISCUSSION

The output screens obtained after executing and running the system are given from Fig. 2 to Fig.12

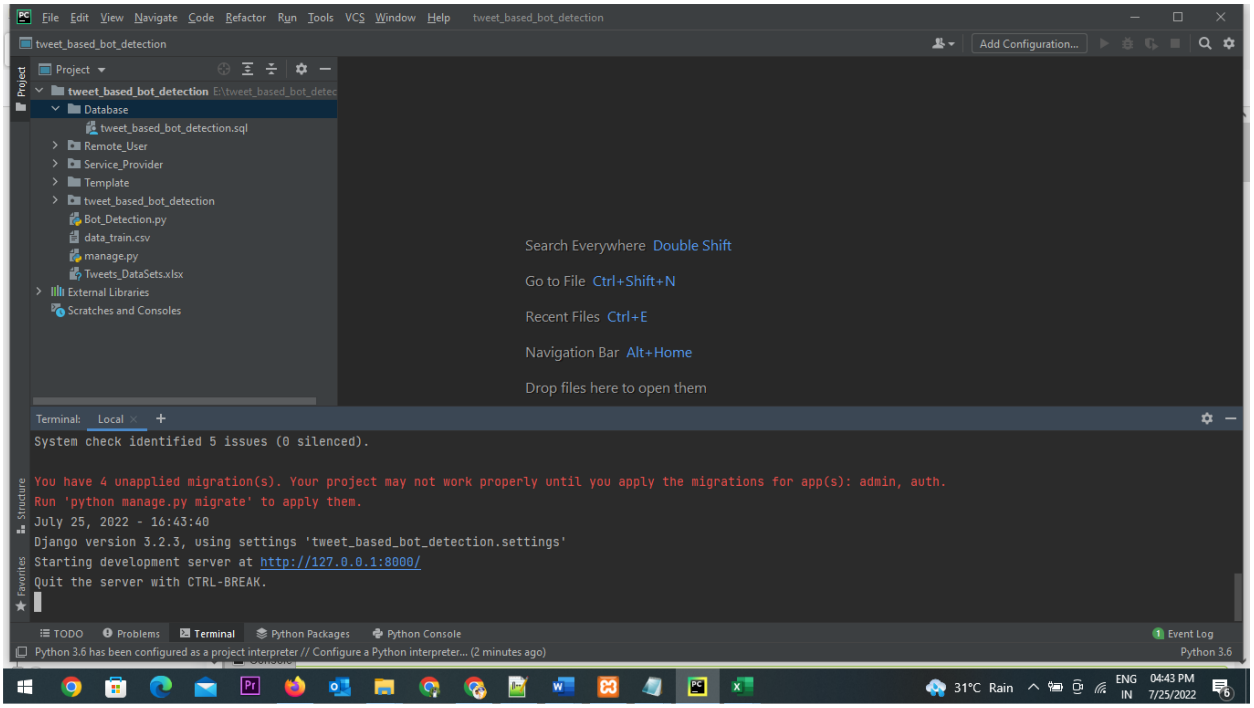

Fig. 2 **Generating Link**

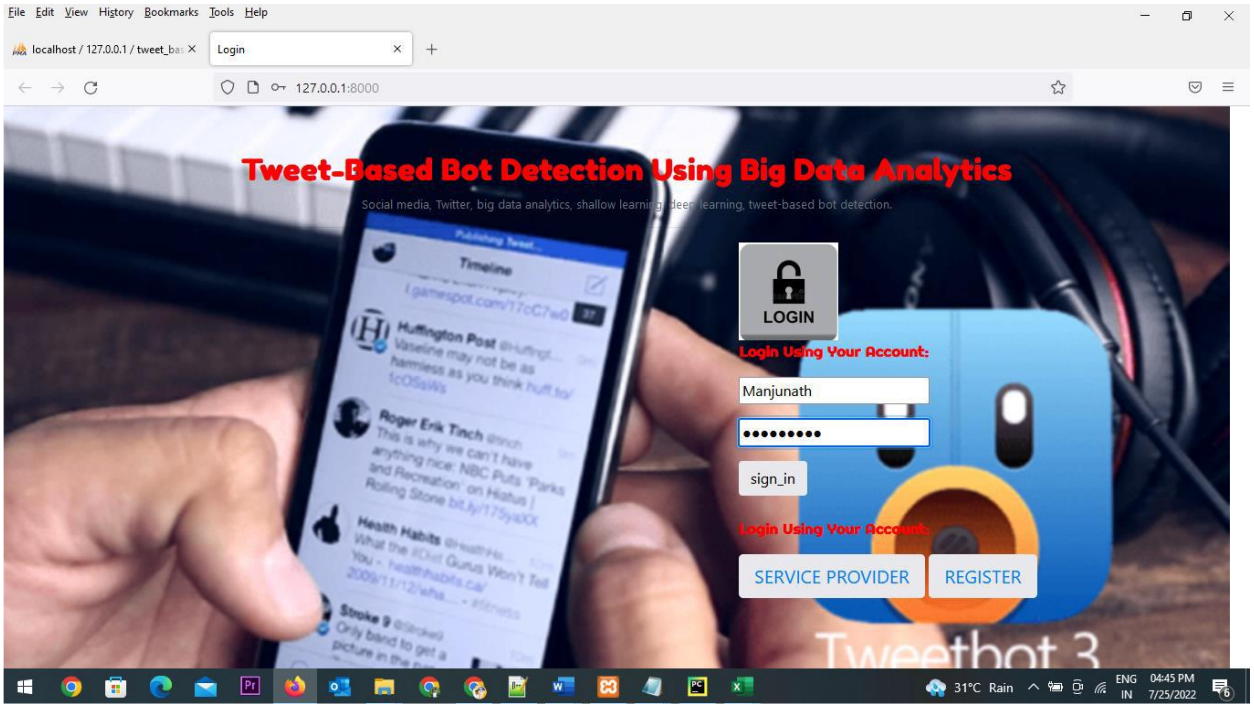

Fig.3 User Login

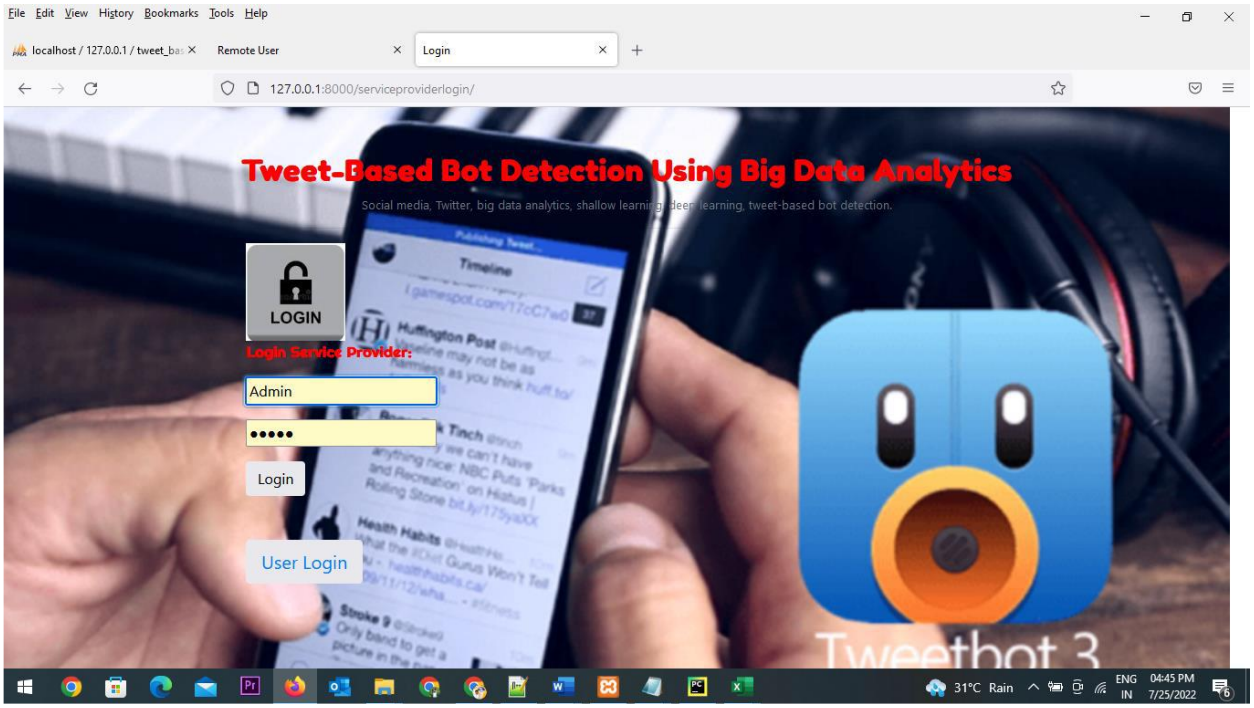

Fig.4 Admin login

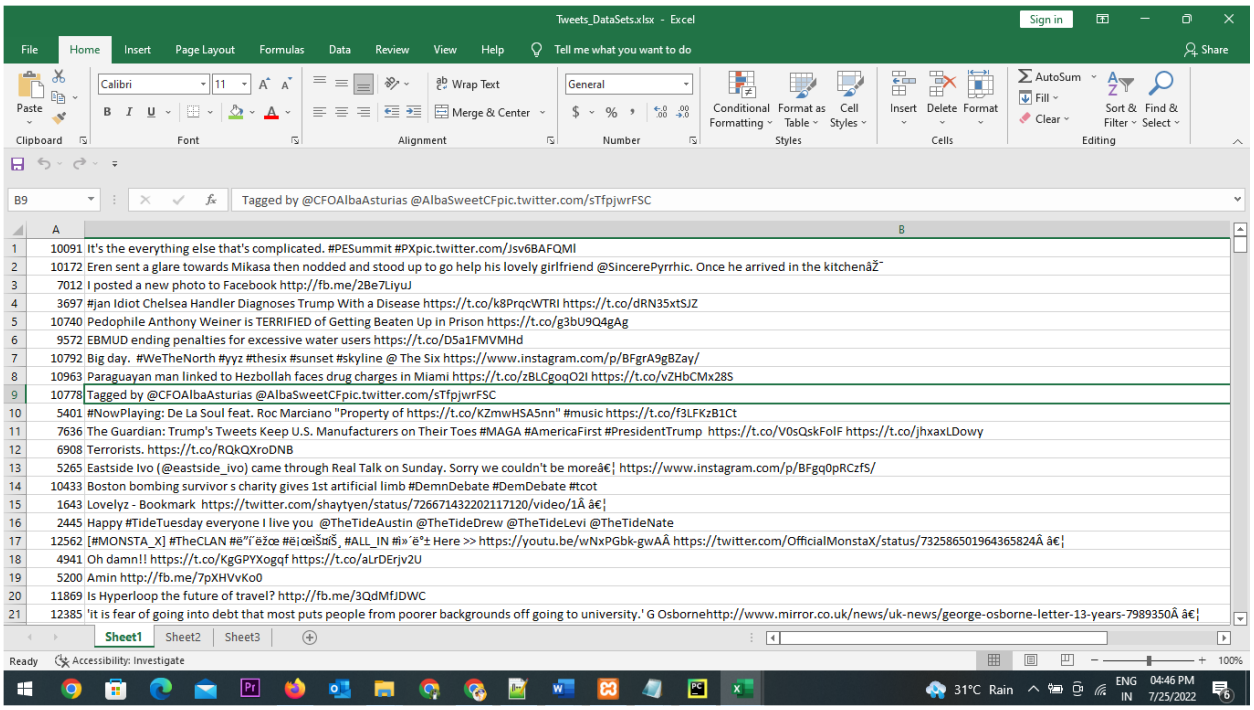

Fig.5 Tweet dataset

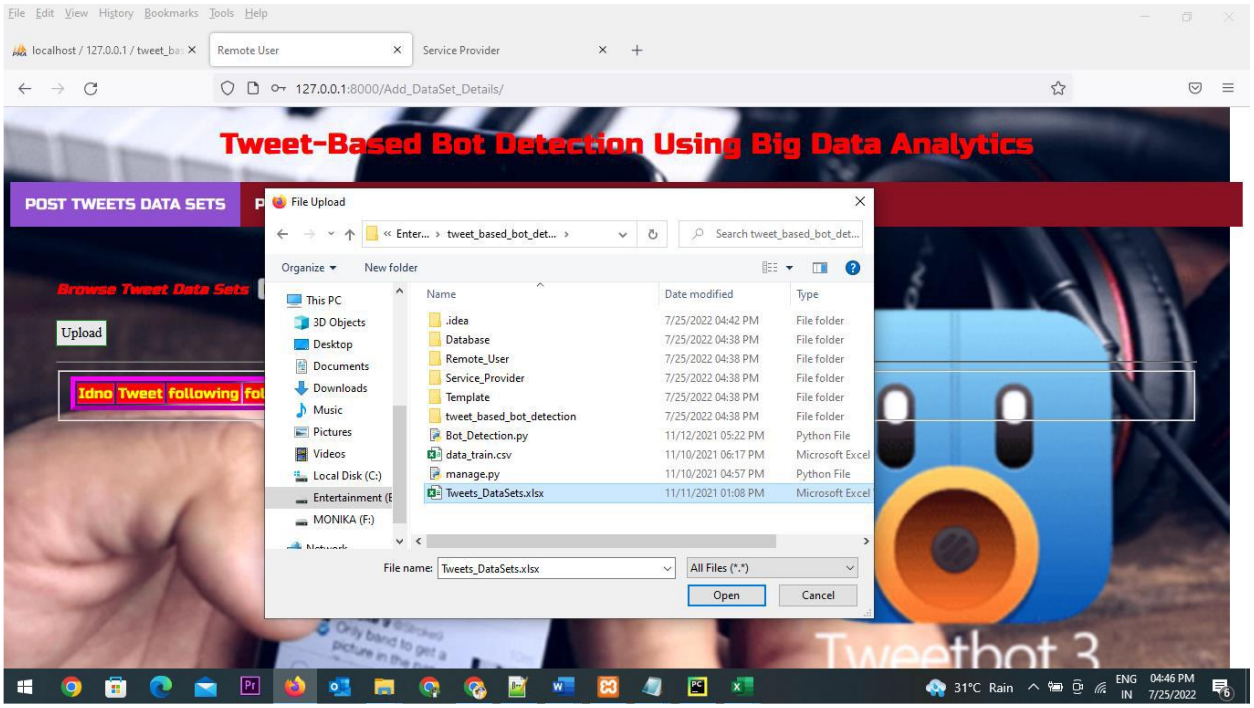

Fig.6 Browse dataset

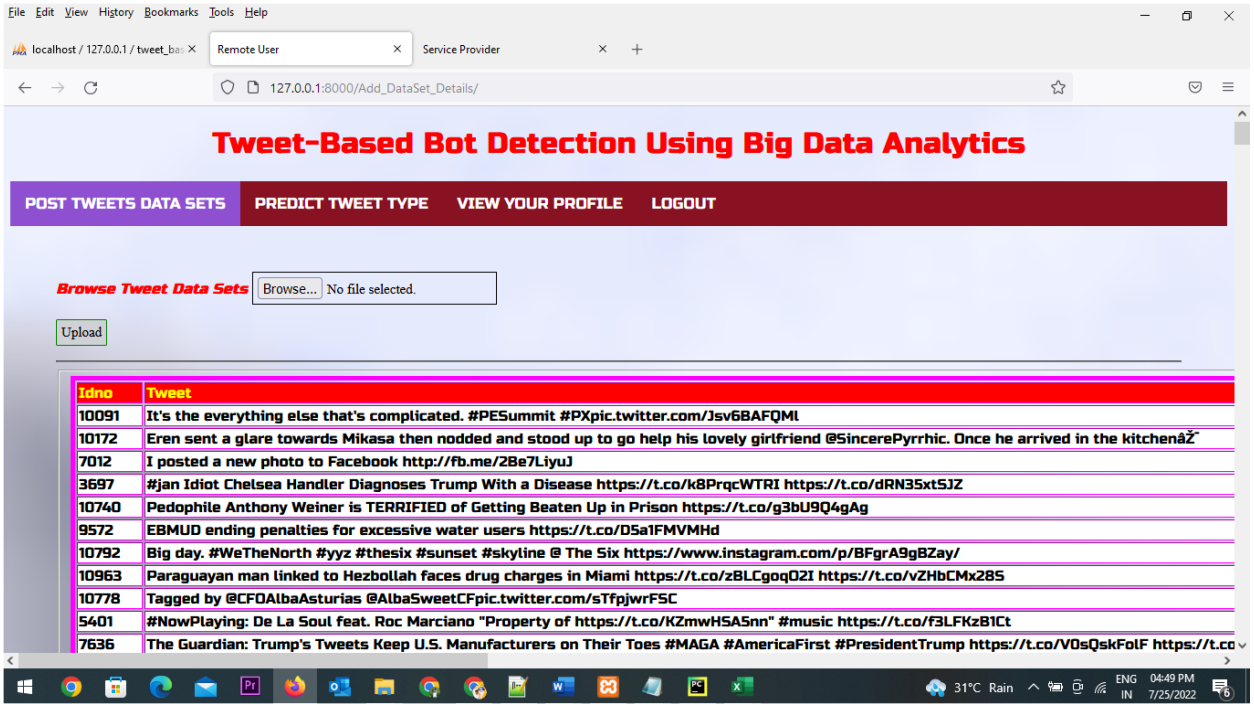

Fig.7 Dataset details

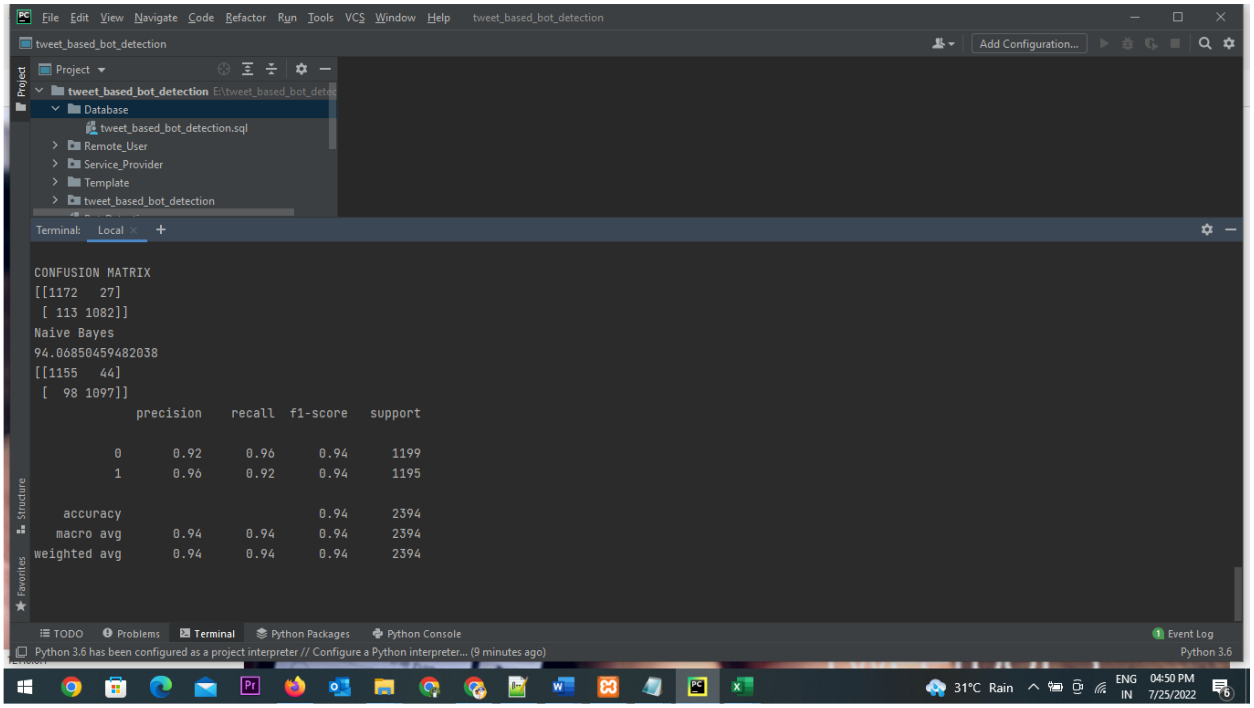

Fig.8 Train dataset with Naïve Bayes

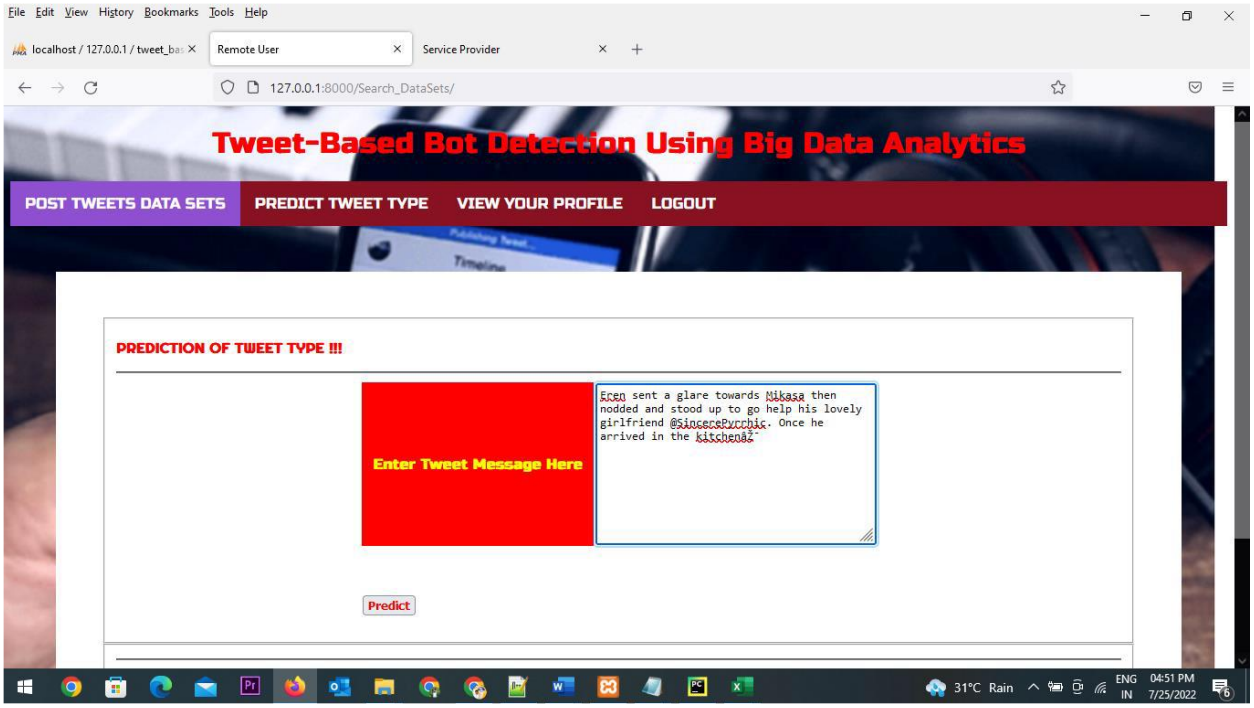

Fig.9 Enter details for prediction

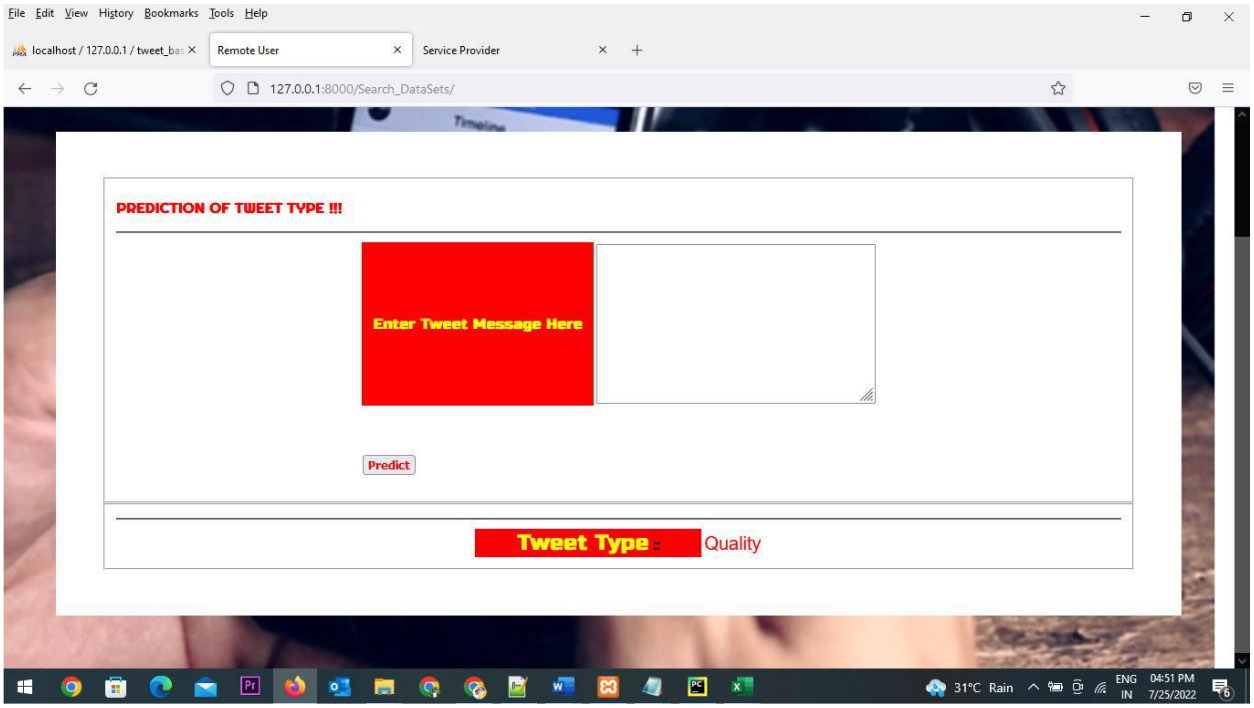

Fig.10 Prediction status

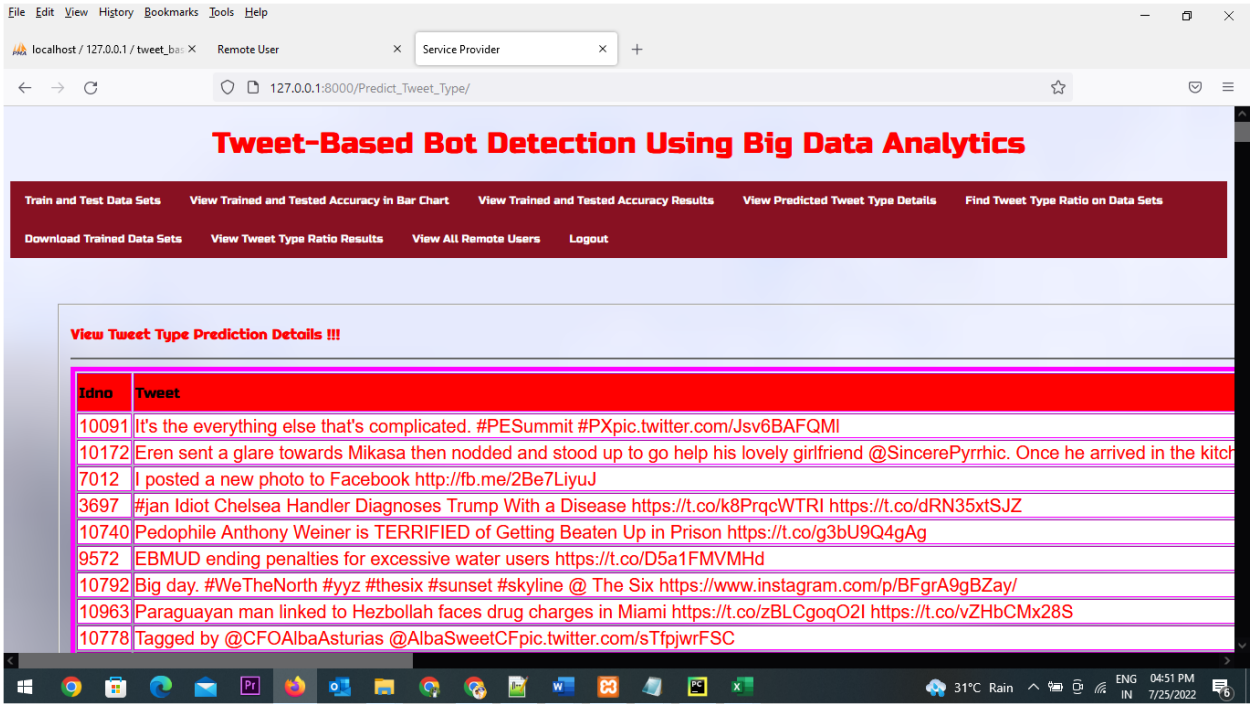

Fig.11 Predicted tweet type details

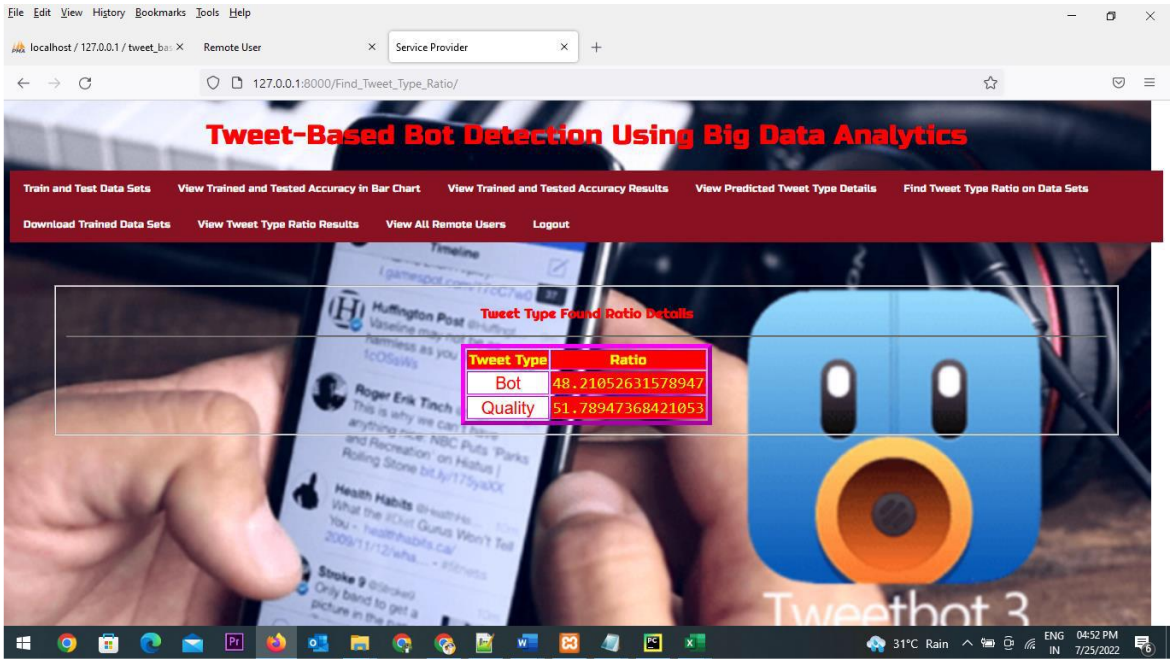

Fig.12 Tweet type ratio details

V. FUTURE SCOPE AND CONCLUSION

Twitter is one of the most popular social media platforms that allows connecting people and helps organizations reaching out to customers. Tweet-based botnet can compromise Twitter and create malicious accounts to launch large-scale attacks and manipulation campaigns. In this review, we have focused on big data analytics, especially shallow and deep learning to fight against tweet-based botnets, and to accurately distinguish between human accounts and tweet-based bot accounts.

REFERENCES

[1] M. Mohsin. (2020). 10 Social Media Statistics You Need to Know in 2021. [Online]. Available: https://www.oberlo.com/blog/social-media-marketing-statistics

[2] I. Arghire. (2020). Twitter Hack: 24 Hours From Phishing Employees to Hijacking Accounts. https://www.securityweek.com/twitter-hack- 24-hours-phishing-employees-hijacking-accounts.

[3] The Rise of Social Media Botnets. Accessed: Feb. 21, 2021. [Online].Available: https://www.darkreading.com/attacks-breaches/the-rise-of-social-media-botnets/a/d-id/1321177

[4] M. Imran, M. H. Durad, F. A. Khan, and A. Derhab, ‘‘Toward an optimal solution against denial of service attacks in software defined networks,’’ Future Gener. Comput. Syst., vol. 92, pp. 444–453, Mar. 2019.

[5] M. S. Savell. (2018). Protect Your Company’s Reputation From Threats by Social Bots. [Online]. Available: https://zignallabs.com/blog/protect-your-companys-reputation-fromthreats- by-social-bots/

[6] S. Aslam. (2021). Twitter by the Numbers: Stats, Demographics &Fun Facts. [Online]. Available: https://www.omnicoreagency.com/twitter-statistics/

[7] A. Aldweesh, A. Derhab, and A. Z. Emam, ‘‘Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and openissues,’’ Knowl.-Based Syst., vol. 189, Feb. 2020, Art. no. 105124