Detecting Malicious Twitter Bot Using Machine Learning And URL Analysis

Bhavika Talele[1], Kuntal Rane[2], Abhishek Pohare[3], Prof.Satyajit Sirsat[4]

Computer Engineering Department [1 2 3 4]

Nutan Maharashtra Institute of Engineering and Technology, Pune, Maharashtra [1 2 3 4]

Abstract— Social media platforms like Twitter face a growing challenge with the proliferation of malicious Twitter bots. These bots spread disinformation, manipulate public opinion, and engage in fraudulent activities, undermining trust in online spaces. To address this issue, our project focuses on detecting malicious Twitter bots using advanced machine learning techniques and URL analysis. By examining various features, including URL characteristics, we've developed a robust framework. We've trained models like Logistic Regression, Random Forest, Naive Bayes, and Decision Trees to classify Twitter accounts as malicious or benign. Our work contributes to enhancing social media security, protecting user trust, and combating fake accounts.

Keywords— Malicious Twitter bots, Machine learning, URL analysis, Social media security, Classification, Logistic Regression, Random Forest, Naive Bayes, Decision Trees, User trust, Fake accounts.

I. INTRODUCTION

In the realm of today's interconnected world, social media platforms have evolved into vital conduits for global communication, information sharing, and interaction. Twitter, in particular, has emerged as a powerful medium, shaping public opinion and fostering engagement. However, this influence has also attracted a darker presence—malicious Twitter bots. These automated entities pose a significant threat to the authenticity and trustworthiness of social media, engaging in activities that range from spreading disinformation to orchestrating cyberattacks.

Detecting and mitigating the influence of these malicious actors has become a complex challenge, as they adeptly emulate human behaviour, making their identification a formidable task. Compounding this challenge is the malicious dissemination of URLs, leading users to phishing sites, malware downloads, or other harmful destinations. Hence, our project, "Detecting Malicious Twitter Bots using URL Analysis and Machine Learning," addresses this pressing issue by leveraging advanced artificial intelligence and machine learning techniques to enhance the security and trustworthiness of the Twitter platform.

II. PROJECT DOMAIN

Within the expansive domain of social media security and authenticity, our project seeks to provide robust solutions to safeguard the integrity of these platforms. The focal points include social media security, the identification of malicious Twitter bots, and the application of machine learning coupled with URL analysis. Together, these elements aim to contribute to the overall trustworthiness of these online spaces.

2.1 Social Media Security:

- Why it matters: Visualize social media platforms as your cherished online communities—safe havens where connecting with friends, staying informed, and sharing thoughts should be devoid of concerns about unscrupulous activities [8,10].
- What we're looking at: Our emphasis lies in fortifying the security and trustworthiness of social media, ensuring users can navigate these platforms with confidence and without second-guessing their interactions.

2.2 Malicious Twitter Bots:

- What they are: Envision sly robots masquerading as regular Twitter users, spreading lies or assuming false identities to manipulate online discourse [9].
- What we're doing: We're dedicated to developing methods to identify and expose these deceptive robot accounts, enhancing the authenticity and honesty of your Twitter experience. [12]

2.3 Machine Learning and AI:

- The cool tech: Harnessing smart computer programs capable of learning from examples and deciphering vast datasets, akin to teaching a computer to discern between genuine users and malicious entities on Twitter.
- Why it's neat: This technology serves as a powerful tool to sift through millions of Twitter accounts, enabling us to pinpoint and neutralize these elusive malicious bots.

2.4 URL Analysis:

- What's a URL: Those web links like "www.somewebsite.com." Our focus is on scrutinizing the links shared on Twitter to identify any potentially harmful content.
- Why it's important: By meticulously examining these links, we aim to safeguard users from inadvertently clicking on malicious content, ensuring a secure and protected online experience.

2.5User Trust and Security:

- Your peace of mind: Our overarching goal is to ensure you can trust your social media platforms. We want you to be confident that what you see and who you interact with is genuine, not someone pretending to be someone else or sharing risky links.
- Why you'll love it: When successful, our project allows you to enjoy your social media experience with heightened confidence and reduced worry, offering a space where authenticity and security are paramount.

By encompassing these critical elements, our project aims to contribute to the broader objectives of social media security. We aspire to provide users with a more secure and genuine online environment, free from the influence of malicious actors, and fortified by advanced machine learning and URL analysis techniques.

III. Literature Survey

An in the domain of detecting malicious bots on social media platforms like Twitter, several significant studies have eased the way for innovative approaches to mark the evolving challenge of detecting social bots. These studies have supported various methodologies to advance the understanding and detection of social media bots.

In Eiman Alothali, proposed a supervised machine learning approach coupled with network analysis to distinguish between human users and bots based on their online behaviour. Their method relied on known bot behaviour and characteristics, able to classify Twitter accounts effectively [1]. In a more recent attempt, Clayton A. Davis, aimed to build a system for evaluating social media bots, with an emphasis on the features and characteristics that differentiate them from the genuine users [2].

Feng Wei explored into the use of Recurrent Neural Networks (RNNs) for detecting the malicious social bots. Their approach was a combination of the analysis of textual content and network behaviour of Twitter accounts. Their study is notable for its application of deep learning techniques [3]. In this paper, the authors research centres around the application of deep learning techniques for bot detection. The concept of deep learning offers sophisticated methods for analysing user behaviour and the fraudulent bot activities [4,11].

Investigates the impact of bots on the tweet sentiment and content exposure. They focus primarily on sentiment analysis and URL-based analysis to detect the activity of malicious bots[5]. The authors explore the concept of bot detection using reduced feature set. The study dives into feature engineering and selection ultimately enhancing the efficiency of bot detection models[6].

Moreover, Alex Hai Wang, focused on the spam bots in the social media networking sites. The methodology he used centers around the machine learning approach in addressing bot related issues. These studies collectively represent the diverse methodologies, from the supervised machine learning to the feature-based bot detection to sentiment analysis and recurrent neural networks [7].

IV. Proposed System

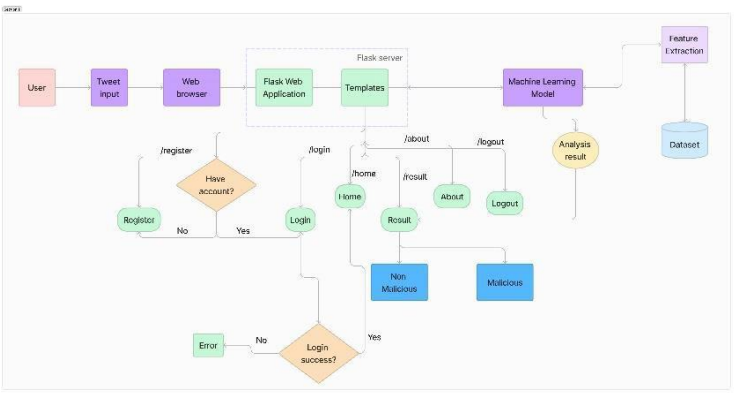

Fig. 1 System Overview

We can use a variety of supervised classification approaches, such as Random Forest, Decision Tree, SVM, Naive-Bayes, and K-Nearest Neighbors (KNN), to determine whether a bot on Twitter is harmful or not. The following procedures outline how to use these strategies to achieve the system classification for malicious bot detection:

1. Preparing the data and gathering it: Gather a sizable dataset of tweets from various Twitter accounts. Tweets from bot and human accounts can be extracted using the Twitter API. Pre-process the data by sanitizing and screening the tweets to exclude any stop words, URLs, mentions, and other unwanted content.

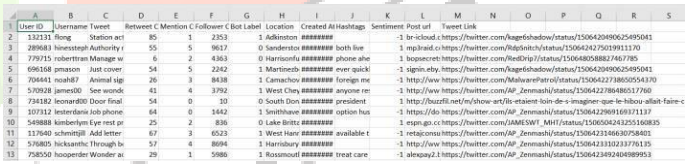

Fig. 2 Dataset Used

2. Feature extraction: Take relevant characteristics that can be utilized to train the classifiers out of the pre-processed twitter data. Metadata like the account age, follower count, and total number of tweets posted are examples of these elements. They may also consist of linguistic elements like tone, word frequency, and the use of hashtags or mentions.
3. Labeling: Indicate whether the account has been found to be a bot propagating spam or other harmful content by classifying the dataset as malicious or non-malicious.
4. Splitting the dataset: The labelled dataset was divided into training and testing sets. Utilize the

testing dataset to assess the classifiers' performance after training them on the training dataset.

5. Model evaluation: Assess each classifier's performance using metrics including accuracy, precision, recall, and F1-score. Examine the outcomes of every method and select the most effective one.
6. Application: The application of our project holds significant implications for social media platforms striving to enhance their security measures and protect user trust. Specifically, our solution stands as a pivotal asset for the identification and containment of malicious Twitter bots, known for orchestrating disinformation campaigns and cyberattacks through deceptive URLs.

   By delving into the analysis of URLs shared within tweets, our solution empowers platforms to take proactive measures against disinformation and fraudulent activities. In doing so, we contribute to the preservation of user trust and the creation of a secure online environment. Our approach aligns with the broader industry trends, echoing the sentiment highlighted in the referenced work regarding the rising need to address malicious activities on Twitter.

   the referenced work regarding the rising need to address malicious activities on Twitter.

   Our endeavour not only focuses on detection but also aims to provide deeper insights into user behaviours, classifications, and the manipulation of hashtags. This multifaceted approach, drawing inspiration from the techniques outlined in the reference, positions our project as a comprehensive solution to the evolving challenges posed by malicious Twitter bots. By leveraging machine learning and URL analysis, we aim to provide social media platforms with the tools necessary to navigate the intricate landscape of online security, fostering user confidence and ensuring the authenticity of the Twitter ecosystem.

V. ALGORITHM

Random_Forest: Our Twitter account classification project is based on Random Forest, a potent ensemble learning technique. It does a thorough analysis of the behaviour of malicious and benign accounts, as well as the URLs published in tweets, to effectively distinguish between them. The power of this paradigm resides in its capacity to build several decision trees. Every decision tree explores different aspects of Twitter account activity, including the content of shared URLs, engagement trends, and the frequency of postings. By combining the outcomes of these distinct trees, Random Forest produces a strong and accurate estimate of the legitimacy of an account.[13]

Because of its adaptability, Random Forest can identify complex and subtle trends in Twitter bot activity. For example, it may detect minute deviations in behaviour associated to URLs, spotting trends that point to malevolent intent. The model can identify fraudulent accounts more accurately than individual decision trees because to this level of investigation. Furthermore, Random Forest's ensemble nature reduces the possibility of overfitting while simultaneously enhancing accuracy, guaranteeing that our approach will continue to be flexible and efficient in a dynamic online context.

VI. PERFORMANCE ASSESSMENT

Performance measures are required to assess how well our system uses machine learning classification to identify dangerous bots on Twitter. Metrics for measuring performance can tell us how well models, algorithms, or evaluation processes are working. The success of the suggested model was assessed using a number of metrics, one of which was a two-dimensional matrix that showed the actual and assigned class. To characterize the makeup of this matrix, true positives, false positives, true negatives, and false negatives are included.

```
Accuracy: 1.0
Confusion Matrix:
 [[4 0]
 [0 4]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         4
           1       1.00      1.00      1.00         4

    accuracy                           1.00         8
   macro avg       1.00      1.00      1.00         8
weighted avg       1.00      1.00      1.00         8
```

Fig. 3 Confusion Matrix Results

An FP, or finding inappropriate tweets unrelated to phony Twitter accounts, is a type 1 error. The TP, on the other hand, is a statistic that shows the effectiveness of finding inappropriate tweets from phony Twitter accounts. However, a type 2 mistake, sometimes referred to as a FN, occurs when the detection algorithm is unable to recognize incorrect tweets from phony Twitter accounts. TN occurs when incorrect tweets that aren't connected to phony Twitter accounts are appropriately ignored by the detection mechanism. A confusion matrix is one tool that can be used to assess the detection system's accuracy. The error rate can be used to compute the accuracy (A), which is a measurement of how closely the detection process matches the real value. When the current rate is divided by the detection rate and given as a percentage, the result is the mistake rate. The accuracy is determined by dividing the total number of tweets by the sum of TP, TN, and NP.

$$Error\ rate = \frac{X - Y}{Z\ x\ 100}$$

Where, x = Detection rate,
y = Current rate,
z= rate

Recall, sensitivity, specificity, and precision/sensitivity are further performance metrics. Specificity is determined by dividing TN by the sum of TN and FP, whereas sensitivity is

determined by dividing TP by the sum of TP and FN. Calculating the detection rate involves dividing the total of TP, TN, FP, and FN by TP. These metrics can be used to assess how well the detection system performs in spotting offensive messages from phony Twitter accounts.

VII. RESULT AND ANALYSIS

In the results and analysis section, we conducted exploratory data analysis (EDA) to gain insights into the relationships and patterns within our dataset. We utilized visualization techniques such as a correlation matrix, pairplot, and bar plot to provide a comprehensive understanding of the features and their impact on bot detection.

The correlation matrix visually represents the pairwise correlations between numerical features in our dataset. A heatmap was generated using the Seaborn library, allowing us to observe the strength and direction of these correlations. The annotation on the heatmap provides correlation coefficients, helping identify potential multicollinearity and guiding feature selection for our predictive model. The results from the correlation matrix contribute to feature engineering and model interpretation.

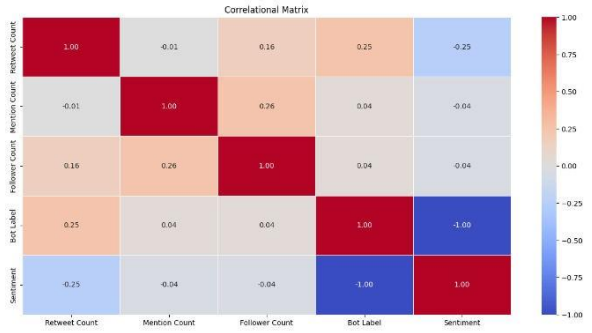

Fig. 4 Correlation Matrix

Additionally, a pairplot was constructed to visualize the relationships among selected features, including 'Retweet Count,' 'Mention Count,' 'Follower Count,' 'Verified,' 'Sentiment,' and the target variable 'Bot Label.' This pairplot facilitates the examination of feature distributions and their variations based on the bot label. Different markers and hues were used to distinguish between bot and non-bot instances, providing a clear overview of feature relationships in the dataset.

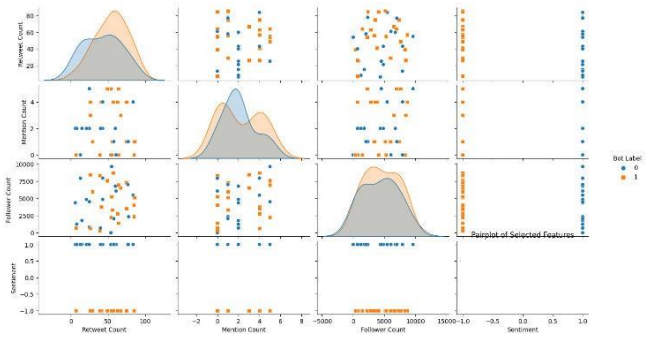

Fig. 5 Pair plot

Furthermore, a bar plot of important features was created to highlight the significance of each feature in influencing the bot detection process. This visualization aids in identifying key features that contribute most to distinguishing between bot and non-bot instances, guiding the interpretation of model predictions.

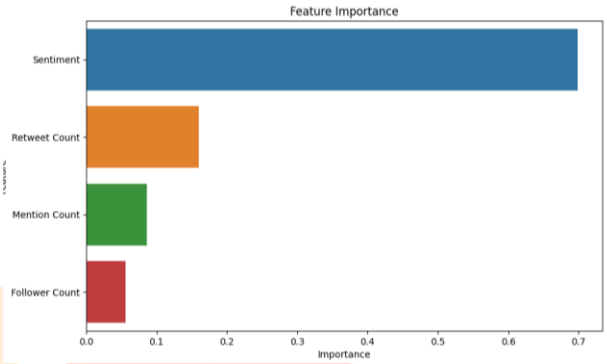

Fig. 6 Bar plot

The login page is the initial point of interaction, where users are required to enter their unique user ID and password for secure access to the platform. This authentication step ensures that only authorized users can enter the system, protecting their personal information and maintaining the integrity of their accounts.

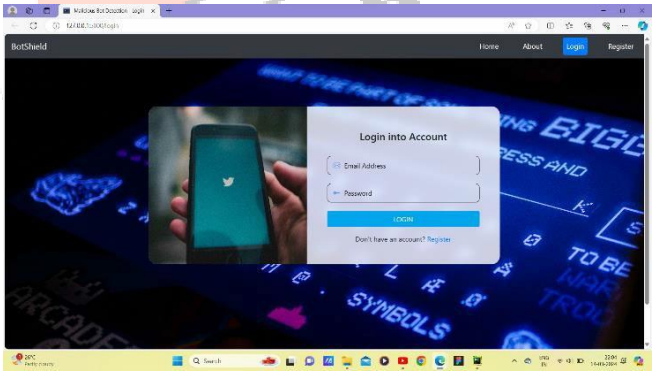

Fig. 7 Login page

Upon successful login, users are directed to the home page—a central hub for various activities. Here, users can seamlessly navigate through the platform's features. One key functionality on the home page is the input section for tweet links and

associated post URLs. Users can conveniently provide this information as part of the analysis process.

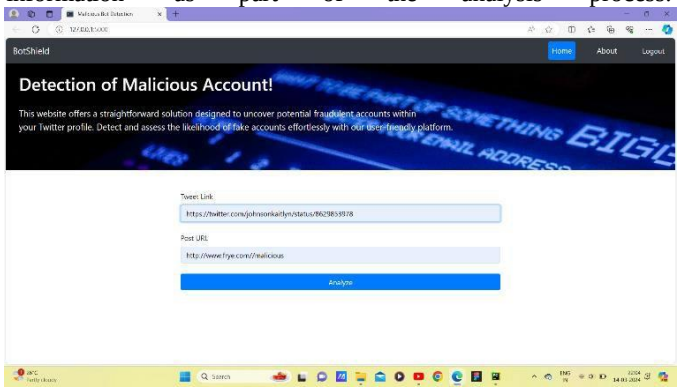

Fig. 8 Home page

After submitting the required information on the home page, users are redirected to the result page. This dynamic page unveils the analysis report, offering crucial insights into the nature of the provided Twitter account. Users receive a comprehensive overview, indicating whether the account is classified as malicious or not. The result page empowers users with valuable information, aiding them in making informed decisions regarding the analysed Twitter account

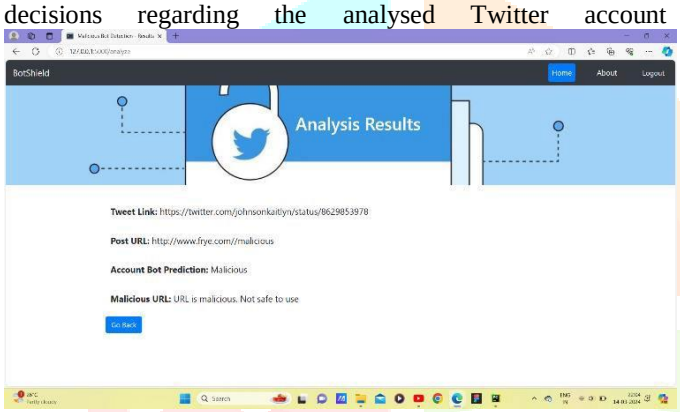

Fig. 9 Result page

VIII. FUTURE SCOPE

The project's future scope involves advanced deep learning and neural networks for better bot detection, real-time monitoring to respond swiftly to threats, and collaboration with social media platforms like Twitter.

Future versions will leverage deep learning and neural networks for more precise analysis of user behavior and URLs, enhancing the detection of malicious Twitter bots.

Ethical considerations and user data privacy are important. Extending the scope to multiple platforms will contribute significantly to fighting malicious bots in the digital age.

IX. CONCLUSION

Our research is a crucial answer to maintain online trustworthiness in a digital world where the spread of malicious Twitter bots constitutes a persistent and developing danger to social media platforms' legitimacy. We have set out on a mission to tackle the various issues that these dishonest digital entities pose by utilizing sophisticated machine learning techniques and thorough URL analysis. The project is a solid framework that identifies and categorizes any risks within the Twitter ecosystem by accounting for a wide range of variables, such as IP addresses, URL length, URL shortening, and more. Using a variety of machine learning models, such as Random Forest, Naive Bayes, Decision Tree, and Logistic Regression, we discovered that Random Forest was the most effective and We have successfully completed a comprehensive investigation of Twitter account legitimacy, concentrating on URLs shared within tweets.

REFERENCES

[1] Alothali, E., Zaki, N., Mohamed, E. A., & Alashwal, H. (2018). *Detecting Social Bots on Twitter*: A Literature Review. 2018 International Conference on Innovations in Information Technology (IIT).

[2] Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). *BotOrNot. Proceedings* of the 25th International Conference Companion on World Wide Web - WWW '16 Companion.

[3] Wei, F., & Nguyen, U. T. (2019). *Twitter Bot Detection Using Bidirectional Long Short-Term Memory Neural Networks and Word Embeddings*. 2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA).

[4] Adam Kenyeres1, György Kovács1Luleå University of Technology (2022). *Twitter bot detection using deep learning*. XVIII. Conference on Hungarian Computational Linguistics.

[5] Stella, M., Ferrara, E., & De Domenico, M. (2018). *Bots increase exposure to negative and inflammatory content in online social systems*. Proceedings of the National Academy of Sciences, 201803470.

[6] Fonseca Abreu, J. V., Ghedini Ralha, C., & Costa Gondim, J. J. (2020). *Twitter Bot Detection with Reduced Feature Set*. 2020 IEEE International Conference on Intelligence and Security Informatics (ISI).

[7] Wang, A. H. (2010). *Detecting Spam Bots in Online Social Networking Sites*: A Machine Learning Approach. Data and Applications Security and Privacy XXIV, 335–342.

[8] Y. Zhou et al ProGuard Detecting vicious accounts in a social network- grounded online elevations, IEEE Access, vol. 5, pp. 1990- 1999, 2017.

[9] F. Morstatter, L.Wu,T.H.Nazer,K.M.Carley andH. Liu, A new approach to bot discovery Striking the balance between perfection and recall, in Proc. IEEE/ ACM Int. Conf.Adv. Social Network. Anal. Mining, San Francisco, CA, USA, Aug. 2016, pp. 533-540.

[10] Discovery of humans, licit bots, and vicious bots in online social networks grounded on ripples, ACM Trans. Multimedia Comput, Commun, Appl, vol. 14, no. 1s, Feb. 2018, Art.no. 26.

[11] M.Sahlabadi,R.C.Muniyandi andZ.Shukur- Detecting abnormal geste in social network Websites by using a process mining- Fashion-J.Comput.Sci. -vol. 10, no. 3, pp. 393- 402, 2014.

[12] M.Al- Qurishi, M.S.Hossain, M.Alrubaian,S.M.M.Rahman andA.Alamri- using analysis of stoner geste to identify vicious conditioning in

large-scale social networks- IEEETrans.Ind.Informat., vol. 14, no. 2, pp. 799-
813, Feb. 2018.

[13] Phillips. Efthimion1, Scott Payne1, Nick Proferes2, ‖ Supervised Machine Learning Bot Discovery ways to Identify Social Twitter Bots ‖, Master of Science in Data Science, Southern Methodist University, 6425 Boaz Lane, Dallas, TX 75205, 2018.S. Barbon,Jr.G.F.C.Campos,G.M.Tavares,R.A.Iga wa,M.L.Proen ¸ ca, Jr andR.C.Guido..