International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/IJSRCSEIT

Tool for identifying and categorizing the Twitter bot's using Machine Learning

Prof. J. N. Ekatpure[1], Harshad Barve[2], Aditya Bhange[2], Sanket Patil[2], Aadit Yadav[2]

[1]Assistant Professor, [2]UG Student

Department of B.E. Computer, SBPCOE, Indapur, Maharashtra, India

ARTICLEINFO

Article History:

Accepted: 10 Oct 2023

Published: 30 Oct 2023

Publication Issue

Volume 9, Issue 10

September-October -2023

Page Number

41-50

ABSTRACT

In today's day-to-day life, plenty of people use Twitter in order to get new information as well as for entertainment purposes. Many of us are aware of the bots on Twitter. Some bots are very useful and help users to do some activities like retweeting, replaying a particular message, and so on. On the other hand, there are some bots on Twitter which are harmful to users' sensitive information. Those bots cause some crucial damage to the user's privacy. Also, some third-party users use the sensitive information of the user gathered for harmful bots for many harmful purposes.

In this project, we discuss a tool that helps the user to identify which are useful bots and which are harmful bots. This tool also helps to classify bots on Twitter into two categories as follows:

1. Useful bots
2. Harmful Bots

The purpose of this tool is to provide safety measures to users while using Twitter. Also, this helps to classify the useful as well as harmful bots on Twitter in order to use Twitter safely and use the useful bots for their useful purpose and save their time as well as sensitive information.

Keyword- Summarization, Classification, Identification, Machine Learning, Analyze, Data Analysis

I. INTRODUCTION

In the vast landscape of social media, Twitter stands as a dynamic platform for global conversations, news publish, etc. However, there are some softwares that can interact with systems or users commonly known as bots. The nature of these bots, coupled with their potential to influence narratives and manipulate tweets, makes it necessary to develop advanced tools for their identification and categorization. This research introduces a

robust machine learning framework designed to reveal the difficulty of Twitter's bot ecosystem. With a focus on both identification and categorization, our tool provides platform administrators, cybersecurity experts, and researchers with the means to discern between genuine users and automated bots, while also revealing the purposes behind the activities of these bots.

This research helps us in following task:

1. Identification: Develop an advanced machine learning model capable of identifying Twitter bots by analyzing behavioral patterns, engagement count and other key features.
2. Categorization: Implement a comprehensive categorization system to classify detected bots into distinct groups based on their intent, such as spamming, or engagement manipulation.
3. Our approach is to combine machine learning algorithms, natural language processing techniques, and data analytics to sift through the massive volume of Twitter data. Leveraging labeled datasets and a meticulous training regimen, the model endeavors to discern subtle distinctions between human and bot behavior.
4. Extract relevant features from user profiles, tweets, and engagement data to improve the accuracy of bot detection.
5. Use machine learning to detect unusual behavior patterns that are characteristic of bot accounts, such as high-frequency posting or identical content sharing.
6. LPR in modern transport systems identifies vehicles via computer vision. Our novel SR algorithm improves license plate legibility in traffic videos.[11]

II. REVIEW / LITERATURE SURVEY

Sr No.	Paper Title	Author Name	Year of Publication	Problem solved in this paper: Existing Problem Statement	The technique used to solve the problem: Existing Problem Solution	What will be future work: Future Scope
1	Twitter Bot Detection Using Diverse Content Features and Applying Machine Learning Algorithms [1]	F.K. Alarfaj, H. Ahmad, H.U. Khan, A.M. Alomair, N. Almusallam, M. Ahmed	2023	To detect Twitter bots based on diverse content-specific feature sets and explore the use of state-of-the-art machine learning classifiers. The real-world	The approach uses the content analysis, special characters, word frequency, part-of-speech, and sentiments. Among several machine learning techniques, random forest (RF), naïve	The consideration of images in detecting Twitter bots because images can contain valuable information and analyzing them could potentially improve the accuracy of bot detection models.

				data from Twitter is scrapped using Twitter API and is pre-processed using standard procedure.	Bayes (NB) and rule-based classification (RBC) are used. The content-related features are used to identify Twitter users as human or bot.	
2	Enhanced Twitter bot detection using ensemble machine learning [2]	Hrushikesh Shukla,Balaji patil,Nakshtra jagtap	2021	To detect social media bots on Twitter, we utilized metadata of Twitter profiles and applied a unique feature selection method, and also explored the power of ensemble learning to make a robust classifier	Data preprocessing techniques are useful to increase the usability of the data and to extract more knowledge from the data. five machine learning algorithms random forest, KNN, AdaBoost, logistic regression, and naïve Bayes For automatic feature selection, we used three commonly used algorithms, namely Principal Component Analysis (PCA)	Moving beyond traditional feature-based detection, future systems may focus on analyzing the behavioral patterns of accounts to detect bots more effectively. This could include looking at posting frequency, content similarity, and engagement patterns.
3	Twitter Bot Detection Using Machine Learning	P. Sai Karthik Reddy , P. Sai Nath , Dr. J. Vijayashree	2023	To develop a system that could identify whether a Twitter	a bot detection system using machine learning with the random	Need to developed few more models and with using some other features which

	Algorithms [3]			account was a bot or not. Through our system development and machine learning evaluation process,.	forest classifier. To achieve this, we designed and implemented a system that takes an account's username, ID, status, verification, listed count, and number of followers as input from the user account and classifies it as either a human user or a bot.The Random Forest classification algorithm was used to build the model, which achieved an accuracy rate of 95%.	helps to find the bot in a more précised and accurate way.
4	Detecting Malicious Twitter Bots Using Machine Learning [4]	Sopinti Chaitanya Raj, B. Srinivas, S.P. Kumar	2022	Detecting people on Twitter who mask their identities for malicious reasons. Because it poses a risk towards other users, it is important towards recognise Twitter bots.it is crucial that	An algorithm in our research that can identify Twitter bots. towards identify BOTS from tweets, logistic regression was used. It contributed towards a decrease in cybercrime. In comparison towards decision tree,	Future implementations could provide real-time data, which would allow Twitter towards incorporate this function into their app. Additionally, it can be integrated among all other market-available social networking programmes. in

				tweets are posted through real people and not Twitter bots. A twitter bot posts spam-related topics.	Multinomial Navie Bayes, and random forest, bag about words algorithm was shown towards be best learning model.	this project, dataset used for detection is provided through us, it is entirely manual. However, in future, I may upgrade project so that model can use dataset needed for bot detection on its own.
5	TwiBot-22: Towards Graph-Based Twitter Bot Detection [5]	Shangbin Feng , Zhaoxuan Tan , Herun Wan1 , Ningnan Wang , Zilong Chen, Binchi Zhang	2022	TwiBot-22, a comprehensive graph-based Twitter bot detection benchmark that presents the largest dataset to date, provides diversified entities and relations on the Twitter network, and has considerably better annotation quality than existing datasets	Employ a two-stage data collection process and adopt the weak supervision learning strategy for data annotation. Then re-implement 35 representative Twitter bot detection models and evaluate them on 9 datasets, including TwiBot-22,to promote a holistic understanding of research progress. We further examine the role of graphs in graph-based methods and the	Graph-based bot detection methods demand significantly more computation resources and execution time than feature-based models. Given that the Twitter network is rapidly expanding, aim to further explore scalable and graph-based bot detection methods.

					generalization ability of competitive bot detection baselines.	
6	Social media bot detection with deep learning methods [6]	Kadhim Hayawi , Susmita Saha , Mohammad Mehedy Masud , Sujith Samuel Mathew , Mohammed Kaosar	2023	Developing deep learning techniques for computation and time efficient techniques for social bot detection with better or compatible performance.	A systematic approach to review the DL applications, as a state-of-the-art and highly advanced technology, in the social media bot detection research to assess the current status and critical challenges. Our review shows that DL-based techniques can be much effective and may potentially outperform traditional ML approaches, with few exceptions that definitely represent a great room for further research.	Retrainable models through real-time processing would be another solution to this issue. Finally, most of the models are confined on twitter now. Leveraging the DL solutions to overcome similar issues in other platforms may potentially increase the usability and impact of this research to a great extent..
7	Naive-Bayesian Classification for Bot Detection in Twitter [7]	Pablo Gamallo and Sattam Almatarneh	2019	To identify influence bots supporting a pro-vaccination discussion on Twitter. The	Linguistic features extracted from tweets and lexical information from external	we will use those features as heuristics of an unsupervised system aimed at ranking Twitter accounts from

				final results of the competition were much more discrete as the best systems did not reach.	resources may help the classification process by improving baseline feature configurations. The experiments also showed that the selected features have a better behavior in the task of identifying bots than in gender profiling transformation speed.	more human to less human. This ranked list of accounts will be revised by annotators so that a reliable gold-standard dataset is obtained at the end.
8	Detection Of Bots In Twitter Network Using Machine Learning Algorithm [8]	Rajnish K. Prince , Snehal S. Thube , Rahul Ranjan , Akash L.Sakat , V. A. Yaduvanshi.	2022	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.	Machine learning algorithm such as Random Forest (RF) and Support Vector Machine (SVM)..	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..
9	Detecting bots in social-networks using node and structural embeddings [9]	Ashkan Dehghan , Kinga Siuta, Agata Skorupka , Akshat Dubey , Andrei Betlen, David Miller , Wei	2022	Whether graph embeddings extract information from the associated network that can be successfully	Examined four distinct feature-sets extracted from the Twitter social network for identifying bot accounts.Divided the features into those captured	The future research question would be to study the impact of combining features gathered various network defnitions, for example one built on

		Xu , Paweł Prałat		used for node classifcation task, what is the relative value of classical vs. structural node embeddings for bot detection,.	directly from the Twitter network, NLP and user-profle data (NLP and P), and those derived from the underlying network structure, node-features (GF) and embeddings (EMB)..	follower/following and another on user-user interact.ion, for identifying bots..
10	Online Twitter Bot Detection: A Comparison Study of Vectorization and Classification Methods on Balanced and Imbalanced Data [10]	Yicong Chen and Jiahe Ling	2023	Aim to classify whether a Tweet comes from a human or a bot. We are particularly interested in comparing the performances of different word embedding methods and classification models under both imbalance and balanced data through the f1-score and confusion matrix.	Bag of Word (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), Doc2Vec, BERT, and fastText. Then, we trained three classification models including Support Vector Machine (SVM), Logistic Regression (LR), and Naive Bayes (NB)..	Implement neural network-based classification models and carry the comparison between each combination of embedding method and classification model. Introduce more variables like users' social network, gender, age, etc. to identify whether the tweet is from a Human or a Bot. Finally, some data augmentation methods for text could be explored to handle the imbalanced data.

III. LIMITATIONS IN EXISTING SYSTEM

1. 1.Evolution of Bots:Bots are designed to adapt, and as detection algorithms improve, so do the techniques used by bots to evade detection. This creates a constant cat-and-mouse game.
2. False Positives: Detection algorithms may generate false positives, flagging legitimate accounts as bots. This can happen when a user's behavior resembles bot-like patterns, leading to the unintentional blocking or suspension of genuine accounts.
3. Sophisticated Bots: Advanced bots can mimic human-like behavior, making it challenging to distinguish them from real users. These bots may have more realistic profiles, varied posting times, and engage in conversations, making them harder to identify.
4. Legitimate Automation: Some accounts use automation for legitimate purposes, such as scheduled posting or automatic updates. Distinguishing between malicious automation and benign automation can be difficult.
5. Use of Human Operators:Bots operated by humans, rather than running on automated scripts, can be challenging to detect. Human-operated bots can mimic natural behavior, making it harder to distinguish them from genuine users.
6. Limited Data Access: Twitter's API (Application Programming Interface) may have limitations on the types of data and signals available for analysis. Limited access to certain data may hinder the effectiveness of bot detection systems..
7. Rapid Scale and Volume: The sheer volume of data on Twitter makes it challenging to analyze and detect bots in real-time. The speed at which information is disseminated on the platform can also make it difficult to keep up with emerging bot strategies.
8. Account Age and Activity: Some detection systems rely on patterns related to account creation date and posting activity. However, this information can be manipulated, and bots may exhibit behavior consistent with genuine users.
9. Collusion and Coordination:Bots may coordinate with each other to appear more legitimate. They may retweet each other's content, engage in coordinated campaigns, or exhibit other behaviors that mimic genuine user interaction.
10. Dynamic IP Addresses: Bots can use dynamic IP addresses, making it harder to trace and block them. They may also use proxies or virtual private networks (VPNs) to mask their true location

IV. CONCLUSION

The development and implementation of bot detection & classification systems represent a significant step forward in safe use of Twitter, offering promising solutions to address the challenges posed by Twitter bots.Bots adapt, evolve, and find new ways to mimic authentic user behavior. As we navigate the complexities of identifying and categorizing Twitter bots, we acknowledge the dynamic nature of these digital bots. . In conclusion, As we look to the future, the path forward involves not only refining our technological tools but also a collaborative and ethically grounded approach to face the challenges that lie ahead. We're always looking to make it better based on what you and other users tell us. We also take your privacy and security very seriously. You can access our tool on the web, and it's designed to be user-friendly.

V. REFERENCES

[1]. F.K. Alarfaj, H. Ahmad,H.U. Khan, A.M. Alomair,N. Almusallam, M. Ahmed " Twitter Bot Detection Using Diverse Content Features and Applying Machine Learning Algorithms" 2023.

[2]. Hrushikesh Shukla,Balaji patil,Nakshtra jagtap "Enhanced Twitter bot detection using ensemble machine learning"2021.

[3]. P. Sai Karthik Reddy , P. Sai Nath , Dr. J. Vijayashree "Twitter Bot Detection Using Machine Learning Algorithms" 2023.

[4]. Sopinti Chaitanya Raj, B. Srinivas, S.P. Kumar "Detecting Malicious Twitter Bots Using Machine Learning"2022.

[5]. Shangbin Feng , Zhaoxuan Tan , Herun Wan1 , Ningnan Wang , Zilong Chen, Binchi Zhang "TwiBot-22: Towards Graph-Based Twitter Bot Detection" 2022.

[6]. Kadhim Hayawi , Susmita Saha , Mohammad Mehedy Masud , Sujith Samuel Mathew , Mohammed Kaosar "Social media bot detection with deep learning methods" 2023.

[7]. Pablo Gamallo and Sattam Almatarneh "Naive-Bayesian Classification for Bot Detection in Twitter" 2019.

[8]. Rajnish K. Prince , Snehal S. Thube , Rahul Ranjan , Akash L.Sakat , V. A. Yaduvanshi. "Detection Of Bots In Twitter Network Using Machine Learning Algorithm" 2022.

[9]. Ashkan Dehghan , Kinga Siuta, Agata Skorupka , Akshat Dubey , Andrei Betlen, David Miller , Wei Xu , Paweł Prałat "Detecting bots in social-networks using node and structural embeddings" 2022.

[10]. Yicong Chen and Jiahe Ling "Online Twitter Bot Detection: A Comparison Study of Vectorization and Classification Methods on Balanced and Imbalanced Data" 2023.

[11]. Dhakane, Vikas Nivrutti, and Jalinder Nivrutti Ekatpure. "Super Resolution of License Plates Using Generalized DAMRF Image Modeling."

[12]. Parlewar, P. ., Jagtap, V. ., Pujeri, U. ., Kulkarni, M. M. S. ., Shirkande, S. T. ., & Tripathi, A. . (2023). An Efficient Low-Loss Data Transmission Model for Noisy Networks. International Journal of Intelligent Systems and Applications in Engineering, 11(9s), 267–276