Machine Learning Classifiers for Social Media Bots Detection on Twitter using Explainable AI

Sarishty Gupta
Department of Computer Science & Engineering and Information Technology
Jaypee Institute of Information Technology
Noida, India
sarishty.gupta@gmail.com

Adwitiya Sinha
Department of Computer Science & Engineering and Information Technology
Jaypee Institute of Information Technology
Noida, India
adwitiya.sinha@jiit.ac.in

2023 Second International Conference on Informatics (ICI) | 979-8-3503-4383-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICI60088.2023.10421550

***Abstract*—Social networks have shaped modern society by revolutionizing how people connect and communicate with others around the world. Social media bots have become more prevalent as a result of social networks' phenomenal development and influence, nevertheless. These automated systems that are guided by algorithms or malicious actors can mimic human behavior and provide the impression that actual people are there. Growing concerns about the integrity of online debate, the dissemination of false information, and potential public opinion manipulation have been sparked by the use of social media bots. It is now necessary to create efficient procedures for identifying and thwarting social media threats in order to tackle this expanding danger. Such efforts use a variety of methods, including as examining profile metadata, observing behavioral patterns, and utilizing the strength of machine learning algorithms. In this paper, we propose an approach which uses multiple features of a user entity on Twitter to train a model for social media bot detection. We employed Twibot-22 dataset to conduct the experiment and compared the performance of our approach with established benchmark models in order to assess its efficacy. Remarkably, our Random Forest model outperformed the benchmarks, achieving an accuracy rate of 0.96. We utilized explainable AI technique to enhance the interpretability of random forest model. This study provides insights that are of immense value in refining the model and delving deeper into the distinguishing features of social bots and human users within the Twitter environment.**

Keywords— Social bots; social network analysis; machine learning classifiers; explainable AI; cybersecurity; LIME.

I. INTRODUCTION

Social networks have become integral platforms for communication, cooperation, and information dissemination in the age of digital connectivity and information sharing. Through the quick interchange of ideas, attitudes, and content, these virtual spaces have not only transformed how people communicate but also changed the communication landscape. But this unheard-of increase in social network usage has also introduced new difficulties and complexity, the most notable of which is the spread of social bots [1].

Social media bots are automated computer programs created to interact with users on social media sites by mimicking human behavior [2]. These bots have the ability to create and distribute material, engage with individuals, follow them, and even have an impact on discussions and trends. While some social bots are helpful and fulfill legitimate needs, such automated chatbots for customer support, others are made with evil intent and work to sway public opinion, disseminate false information, or carry out fraudulent operations[3,4].

As an instance, in Mumbai, social media platforms were inundated with rumors asserting that government-backed vaccines were part of a conspiracy to render Muslim children sterile.

Consequently, this misinformation campaign led to a significant reduction in the number of vaccinations administered, with only 50% of the expected recipients actually receiving the vaccine [5].

Moreover, a study [6] conducted during the 2020 US Presidential Election uncovered a notable surge in the volume of election-related tweets per capita generated by bots. These automated accounts predominantly disseminated content aligned with specific political ideologies.

Additionally, social bots were involved in the propagation of less credible information during the Covid-19 [7]. Furthermore, social bots have emerged as prominent purveyors of disinformation related to climate change, potentially undermining support for policies aimed at mitigating rising global temperatures.

These nefarious activities collectively pose a significant challenge to information security, jeopardizing the wholesome progression of social networking platforms [8].There has never been a more pressing need for social bot detection. The reliability, security, and integrity of social networks are seriously threatened by the complexity and adaptability of these automated entities. Unchecked social bot activity has far-reaching effects on public health, politics, e-commerce and cybersecurity, among other fields.

Various disciplines, including machine learning, data science, network analysis, and cybersecurity, are used to discover social bots[9,10]. In order to ensure the legitimacy and authenticity of online interactions as well as to defend people and organizations from potential harm, it is crucial to identify these bots [11]. It makes it possible for social media platforms to keep a secure and dependable environment for users, stops the propagation of rumors and propaganda, and ultimately upholds the values of accountability and transparency in the digital era [12].

In our proposed approach, we conducted training using the Twibot-22 dataset. We have used multiple features which are extracted for user entities and applied Random forest model to identify social bots. The model performance demonstrated noteworthy results when compared to other established benchmark models applied to the Twibot-22 dataset. Further we employed explainable AI method to comprehend the model and understand the impact of certain features on model performance.

This paper is further organized as follows. Section II provides the related work on social media bot detection. Section III discusses about the proposed methodology including the data collection and preprocessing process, machine learning models and model explainability using LIME. Section IV displays the experimental analysis and results performed to evaluate the proposed approach. Section V discussed the future research and concludes the paper.

979-8-3503-4383-0/23/$31.00 ©2023 IEEE

II. RELATED WORK

Various machine learning techniques have been designed that curb the impact of social bot profiles, as these automated profiles on social media platforms mirror human traits with the intention to manipulate public perception or spread misleading information.

A profile-based framework was established for the detection of social bots. This framework utilized random forest and attained an accuracy of 89% by working on a set of six features obtained via hybrid feature selection [13]. In a similar vein, the authors conducted the localization of automated accounts by leveraging profile metadata. This process involved the incorporation of weight-of-evidence encoding and employed an extra tree classifier for the purpose of feature selection. This approach showcased commendable performance, achieving an accuracy rate of 88% with the application of the random forest algorithm [14].

In a manner akin to other techniques for identifying bots, reference [15] implemented a network graph approach that relies on behavioral similarity and trust values among participants. They employed Deep Autoencoder-based methods for detecting social botnet communities. Additionally, the study explored the relationship between a user's social interactions and the network graph associated with the account for social bot detection. The study provided an effective semi-supervised technique that demonstrates a surprising capacity to accurately distinguish between bots and authentic users. [16].

In a study by the author in [17], a temporal-based approach was executed to identify social bot accounts. This methodology hinges on analyzing variations in user behavior during specific time intervals, leveraging statistical diversity measures. Furthermore, the research introduced a mechanism for detecting malicious tweets based on temporal attributes. This aspect involves the examination of pairwise similarities in individual retweeting behavior [18].

In addition to the overarching goal of identifying social bots, distinct methods have emerged. These approaches encompass a range of facets, comprising profile attributes, content analysis, temporal characteristics, behavioral tendencies, and network-based methodologies, all geared toward the detection of social bots. However, it's noteworthy that these approaches often place lesser emphasis on the holistic integration of attributes like timing related, interaction related and metadata which collectively contribute to a more robust social bot detection framework [19].

The authors in [20] introduced an innovative and transparent method for identifying bots, which provides interpretable, accountable, and AI-driven bot detection on the Twitter platform. Additionally, they have implemented a publicly accessible bot detection web service. The service seamlessly incorporated an interpretable machine learning framework and a user feedback mechanism within an efficient crowdsourcing framework.

In a recent investigation conducted in [21], the authors leveraged Twitter profile metadata and implemented a distinctive feature selection technique to identify social media bots on Twitter. Additionally, they created a resilient classifier using the potential of ensemble learning. To choose the best ensembling approaches they conducted a thorough comparative analysis.

III. PROPOSED METHODOLOGY

A. Data Collection and Preprocessing

Initially our research involved collecting and preprocessing the dataset, which contains information about social media users. We have used Twibot-22 dataset [22]. There are 1,000,000 different users in the data, comprising139,943 bots and 860,057 humans. Twibot-22 displays 14 types of relations and 4 types of entities.

We have used 5000 user entries from the dataset. Each row of our data represents a user, and the columns include features such as the count of followers' accounts, description, and count of following accounts, the number of tweets, listed count, username, URL, location, verified status, and the target variable, bot label.

Numerical characteristics including follower count, following count, and tweet count were retrieved as features from the dataset. Feature engineering was used to create derived attributes namely the ratio of followers to followers, the length of the description, length of username, and the existence of URL and location information. We have represented human account as 0 and bot accounts as 1.

B. Machine Learning Models

To classify users as bots or humans, several machine learning models were employed to the dataframes after performing the preprocessing. The following models were considered:

1) Support Vector Machine (SVM): It is a strong machine learning medhod which is utilized in our study for classification. SVM works by identifying the appropriate hyperplane for separating data points from various classes. The purpose is maximising the marginbetween the two classes (bots or humans). This is done by calculating the distance measured between the hyperplane and data points located closest from each class [23].

For a binary classification problem, SVM aims to discover a hyperplane represented by the following equation:

$$w^t y + c = 0 \quad (1)$$

In equation (1), w denotes the weight vector, y refers to the input data vector and c denotes the bias term.

A Support Vector Machine classifier was executed on the data to predict bot labels. The classification reports, including some performance metrics were computed to assess its effectiveness.

2) Decision Tree: It is a non-parametric supervised learning approach which is also employed for classification in our work. They describe decisions as a tree-like structure, where every internal node indicates a decision based on attribute, each leaf node represents a final decision, and every branch node denoting the the decision's result [24].

A splitting criterion is used in decision trees to select the feature and split point at each internal node. Gini lmpurity is used as criterion for splitting.

$$\mathrm{Gini}(x) = 1 - \sum_{j=1}^{n} \mathrm{prob}(j|x) \quad (2)$$

In equation (2), n symbolizes number of classes and $prob(j|x)$ is the probability of class j at node x.

3) Random Forest: Multiple decision trees are incorporated in Random Forest which is an ensemble learning method, to enhance the prediction accuracy. It operatesby constructinga large number of decision trees during training and combining their predicted outcomes during testing.

The final prediction of a Random Forest for classification is usually dependent on the vote of the individual decision trees, either in a weighted or majority fashion [25]. Given N decision trees, the predicted class for a data point y is represented as:

$$\text{Class(y)} = \text{argmax}_l \sum_{i=1}^{N} F(p_t(y) = l) \quad (3)$$

In equation (3), $p_t(y)$ denotes the prediction of the i^{th} decision tree, $F()$ is the indicator function and l is a class label. A Random Forest classifier was trained with 700 decision trees on the dataset.

C. Model Explainability using LIME

We employed Local Interpretable Model-agnostic Explanations (LIME) technique to improve the transparency and comprehension of ML models, specifically for our dataset and the selected Random Forest model. LIME provides an effective framework for producing contextualized and understandable justifications for model predictions [26-27].

LIME functions according to the idea of approximating the behavior of a complex model with a simpler and easier to understand counterpart, known as a "local surrogate model." The surrogate model is built around a particular piece of data, enabling us to examine the underlying reasoning behind the model prediction in a constrained, clear context. LIME is significant in our study since it can make the Random Forest model more transparent.

IV. EXPERIMENTAL RESULTS

A. Exploratory Data Analysis

In order to comprehend more about the aspects of the dataset and their relationships, exploratory data analysis was carried out. Several visualizations were created to better understand the data.

To see the pairwise correlations between the numerical features, a correlation heatmap was created. Strong correlations and possible multicollinearity problems were found as a result.

Fig. 1 shows the correlation heatmap for all the features. The heatmap shows a strong correlation between features listed_count and followers_count, url_presence and description_length, location_presence and description length, location_presence and url_presence. Also, there is negative correlation of bot_label with decription length, url_presence, location_presence and verified_status.

Pair plots were used to visualize relationships between numerical features, with the data points differentiated by the bot label. Fig. 2 shows the pair plots between features following count, followers count and tweet count. The pair plot shows that followers count has positive correlation with following count and tweet count.

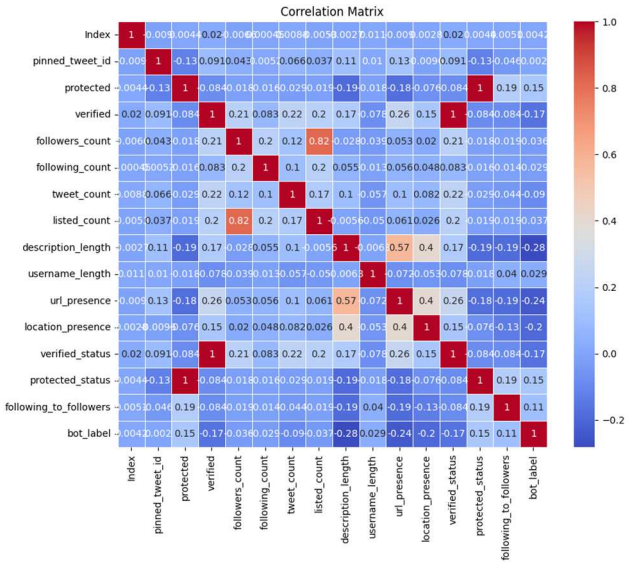

Fig.1. Correlation Matrix of Features

B. Results

We trained different machine learning models on our Twibot-22 dataset. The random forest model exhibits an accuracy of 0.96, signifying its capability to properly identify 96% of input data. The precision value of 0.96 underscores the model accuracy for determining social media bots. The ability to detect a sizable portion of the dataset's social media bots is demonstrated by its recall value of 0.95.Furthermore, it attains a F1 value of 0.95, indicating a harmonic mean between the model recall and precision. These outcomes demonstrate the random forest algorithm's efficacy in identifying social media bots, a crucial factor in upholding the integrity of social media platforms.

We have compared the random forest model with other models. Table I show the results and performance comparison among the models used for study. The SVM model shows the accuracy of 53%. For decision tree, an accuracy of 92% was achieved. In the context of the Random Forest technique, 700 trees were utilized to ascertain the most suitable node for division, ultimately yielding an accuracy rate of 0.96.

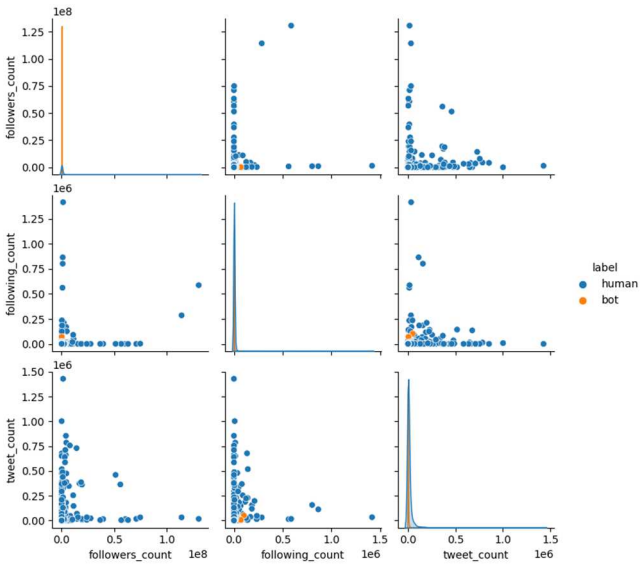

Fig.2. Pair Plot between Following, Followers and Tweet count

TABLE I. COMPARISON OF MODEL PERFORMANCE

Models	Precision	Recall	F1-Score	Accuracy
SVM	0.73	0.54	0.42	0.53
Decision Tree	0.93	0.92	0.92	0.92
Random Forest	**0.96**	**0.95**	**0.95**	**0.96**

C. Explaining model using LIME

We considered three specific instances from our dataset. Lime presents significant comprehensions into the behavior of our implemented random forest model on individual data points, shedding light on why certain predictions were made.

Fig. 3 shows the first instance as one twitter user. This user is classified as a bot by our model. Lime reveals that one of the key features contributing to this classification is the large amount of accounts the user follows. Our model seems to heavily rely on this feature when making predictions, which indicates that a large following count might be a strong indicator of bot-like behaviour in our dataset.

Fig. 4 represents the second instance. Our second instance focuses on a user who is classified as a human by our model. Lime's explanation highlights the importance of the description_length feature in this classification. It appears that humans in our dataset tend to have a higher description length, which distinguishes them from bots. This interpretation aligns with our understanding of human behaviour on social media platforms.

Fig. 5 shows the third instance. The third instance is predicted as bot account by our model. Lime explains the presence of "tweet count" feature is important in the classification. It indicates that more tweet count can be factor to consider in social bot detection. These three instances exemplify the power of Lime in providing transparent and interpretable explanations for our model predictions. They highlight the specific features that influence the outcomes and help us understand the decision-making process of our machine learning model.

Such insights are invaluable for refining our model and gaining deeper insights into the characteristics of humans and social media bot accounts on Twitter.

V. CONCLUSION

Bots on social media have aided in the quick spread of information, affected political movements, and even shaped public conversation. It is necessary to create efficient techniques for identifying and minimizing social media bots in order to combat this expanding menace. This study explored the crucial area of detection of social media bots using the Twibot-22 dataset. To concentrate on the complex matter of recognizing social media bots, our comprehensive approach utilized multiple features of users and applied benchmark methods of machine learning like Decision Tree, SVM, and Random Forest. In relation to other models, Random Forest outscored them. The accuracy of the Random Forest model was 96%, with 95% for recall, 96% for precision, and 96% for F1.

The utility of models in the discipline of machine learning is not just determined by their performance and accuracy. It is equally crucial to understand why a model predicts a certain way. Building trust, maintaining justice, and facilitating regulatory compliance all depend on a model explainability, or its capacity to explain how decisions were made. The employing of LIME on our Random Forest model yielded invaluable insights into the reasoning behind particular predictions. It allowed us to comprehend that certain features like following count, tweet count, description length and followers count had a significant impact on the model's output for specific occurrences within our dataset.

In future research, we can investigate more sophisticated ensemble models, including fusing deep learning architectures with conventional machine learning models. Further we can concentrate on creating bot detection systems that can adjust to changing bot behaviors and actively defend against adversary attacks.

Fig.3. First Instance Analysis Predicted as Bot

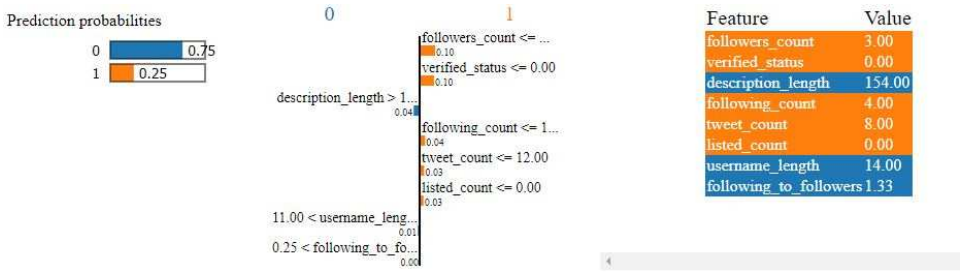

Fig. 4. Second Instance Analysis Predicted as Human

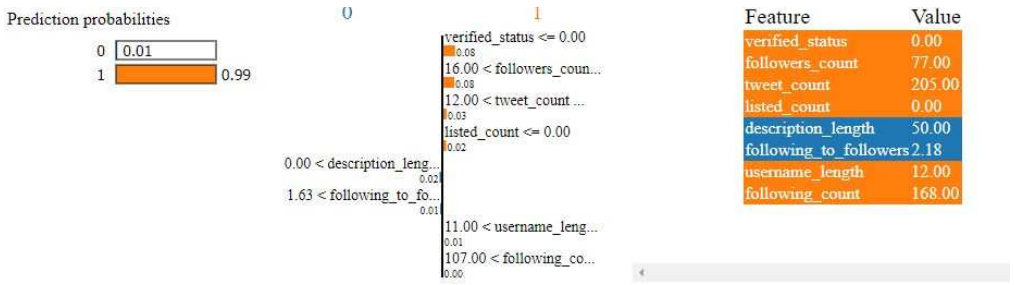

Fig. 5. Third Instance Analysis Predicted as Bot

REFERENCES

[1] H. Liu, J. Han, and H. Motoda, "Uncovering deception in social media," Social Network Analysis & Mining, 2014, vol.4, pp. 1-2.

[2] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. "The rise of social bots." Communications of the ACM 59, no. 7: 96-104, 2016.

[3] M. Jiang, P. Cui, and C. Faloutsos, "Suspicious Behavior Detection: Current Trends and Future Directions," IEEE Intelligent Systems, 2016, vol. 31, pp. 31-39.

[4] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," SIGKDD Explorations, vol. 19, no. 1, pp. 22–36, 2017.

[5] H. J. Larson. "Stuck: How Vaccine Rumors Start--and Why They Don't Go Away". Oxford University Press, 2020.

[6] H. C. H. Chang, E. Chen, M. Zhang, G. Muric, and E. Ferrara. "Social bots and social media manipulation in 2020: the year in review." arXiv preprint arXiv:2102.08436, 2021.

[7] K.C. Yang, C. Torres-Lugo, and F. Menczer. "Prevalence of low-credibility information on twitter during the covid-19 outbreak." arXiv preprint arXiv:2004.14484, 2020.

[8] M. Workman, "An empirical study of social media exchanges about a controversial topic: Confirmation bias and participant characteristics," Social Media in Society, pp. 381–400, 2018.

[9] M. Heidari, S. Zad, and S. Rafatirad, "Ensemble of supervised and unsupervised learning models to predict a profitable business decision," in IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021, 2021.

[10] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "Botornot: A system to evaluate social bots," in Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume, pp. 273–274, 2016.

[11] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017, pp. 963– 972, 2017

[12] L. Madahali and M. Hall, "Application of the benford's law to social bots and information operations activities," in 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), pp. 1–8, CyberSA, 2020.

[13] E. Alothali, K. Hayawi, and H. Alashwal. "Hybrid feature selection approach to identify optimal features of profile metadata to detect social bots in Twitter." Social Network Analysis and Mining 11.2021: 1-15.

[14] H. Shukla, N. Jagtap, and B. Patil. "Enhanced twitter bot detection using ensemble machine learning." In 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 930-936. IEEE, 2021.

[15] G. Lingam, R. R. Rout, D. VLN Somayajulu, and S. K. Das. "Social botnet community detection: a novel approach based on behavioral similarity in twitter network using deep learning." In Proceedings of the 15th ACM Asia Conference on Computer and Communications Security, pp. 708-718.,2020.

[16] M. Mendoza, M. Tesconi, and S. Cresci. "Bots in social and interaction networks: detection and impact estimation." ACM Transactions on Information Systems (TOIS) 39, no. 1: 1-32, 2020.

[17] D. Kosmajac, and V. Keselj. "Twitter bot detection using diversity measures." In Proceedings of the 3rd International Conference on Natural Language and Speech Processing, pp. 1-8. 2019.

[18] N. Vo, K. Lee, C. Cao, T. Tran, and H. Choi. "Revealing and detecting malicious retweeter groups." In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 363-368. 2017.

[19] Y. Wu, Y. Fang, S. Shang, J. Jin, L. Wei, and H. Wang. "A novel framework for detecting social bots with deep neural networks and active learning." Knowledge-Based Systems 211: 106525, 2021.

[20] M. Kouvela, I. Dimitriadis, and A. Vakali. "Bot-Detective: An explainable Twitter bot detection service with crowdsourcing functionalities." In Proceedings of the 12th International Conference on Management of Digital EcoSystems, pp. 55-63. 2020.

[21] H. Shukla, N. Jagtap and B. Patil, "Enhanced Twitter bot detection using ensemble machine learning," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 930-936.

[22] S. Feng, Z. Tan, H. Wan, N. Wang, Z. Chen, B. Zhang, Q. Zheng et al. "TwiBot-22: Towards graph-based Twitter bot detection." Advances in Neural Information Processing Systems 35: 35254-35269, 2022.

[23] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner. "Predicting susceptibility to social bots on twitter." In 2013 IEEE 14th international conference on information reuse & integration (IRI), pp. 6-13. IEEE, 2013.

[24] S. Johnson. "Detecting Malicious Tweet Bots using Machine Learning Algorithms." Turkish Online Journal of Qualitative Inquiry 12, no. 4, 2021.

[25] J. Schnebly, and S. Sengupta. "Random forest twitter bot classifier." In 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0506-0512. IEEE, 2019.

[26] E. Park, K. H. Park, and H. K. Kim. "Understand watchdogs: Discover how game bot get discovered." arXiv preprint arXiv:2011.13374. 2020.

[27] N. Capuano, G. Fenza, V. Loia, and C. Stanzione. "Explainable artificial intelligence in cybersecurity: A survey." IEEE Access 10: 93575-93600, 2022.