

IDENTIFICATION OF SPAMBOTS AND FAKE FOLLOWERS ON SOCIAL NETWORK VIA INTERPRETABLE AI-BASED MACHINE LEARNING

Sk.Nabeel ¹ Dr. K.Venkataramana ²

¹student, Mca 2nd Year Kmmips, Tirupati, Affiliated To S.V. University, Tirupati, A.P, India

²professor, Dept Of Mca, Kmmips, Tirupati, Affiliated To S.V. University, Tirupati, A.P, India

ABSTRACT - Social networking platforms like X (Twitter) serve as hubs for open human interaction, but they are also increasingly infiltrated by automated accounts masquerading as human users. These bots often engage in activities such as spreading fake news and manipulating public opinion during politically sensitive times like elections. Most of the current bot detection methods rely on black-box algorithms, raising concerns about their transparency and practical usability. This study aims to address these limitations by developing a novel methodology for the detection of spambots and fake followers using annotated data. To this end, we propose an interpretable machine learning (ML) framework, leveraging multiple Algorithms with hyper parameters optimized through cross-validation, to enhance the detection process. Fur the more results showcase the model's ability to identify key distinguishing attributes between bots and legitimate users which offers a transparent and effective solution for social network bot detection.

Key words : Spambots , Identification, Social Network, Ai-Based Machine Learning Fake Followers Via Interpretable

1.INTRODUCTION

Social networks have become the key source of information in the new age of mankind. X formerly known as Twit term is presently among the most prevalent and widely used nature and expanding user base. These bots can be useful as legitimate bots produce a lot of educational tweets, such as blogs and news updates. Malicious bots, however, disseminate ante spam or harmful material. The characteristics used by current Twitter bot identification algorithms are often derived from user data, including timestamps, friendship, behaviour, and network connection. Nevertheless, feature Engi nearing requires a lot of work and effort. Social bots have the potential to facilitate the dissemination of misinformation, including fake news, rumours, and hate speech, by rapidly amplifying low-credibility content on X through interactions with high-profile users and strategic mentions. Most of the aforementioned issues are controlled through the use of bots. A botnet is a collection of bots designed to execute specific tasks, while a Sybil account represents a fabric catted identity that does not correspond to or originate from a real human user. These botnets and Sybil accounts

are frequently employed to amplify disinformation and disrupt genuine discourse, contributing to the challenges of maintain Ing integrity in online platforms Machine learning (ML) has been

successfully utilized in a vast range of areas such as sports analytics, sentiment analysis, fake news detection and social bot detection. Our study focuses on interpretable machine learning (XAI) as it has been used in different areas to improve performance and to gain better comprehension of the model. Figure 1 provides the most commonly used Inter printable AI techniques among which SHAP and LIME are the most popular. Interpretable ML provides insight into how a particular data point or data point affects the prediction model using a variety of methods such as factor analysis, local interpretation model-agnostic interpretation (LIME), and Shapley additive interpretation (SHAP). The added transparency helps users understand and trust AI systems while it also allows stakeholders to identify biases in these systems thus promoting accountability and fairness in AI applications. Overall, descriptive ML plays an important part in closing the disparity between AI algorithms and human comprehension which supports informed decision-making and increasing trust in AI technology. Thus, utilizing XAI for social network

bot detection (SNBD) is an important step to gain a better understanding of its detection process. social media sites and thus it plays an important role in online conversations and helps connect millions of active users. However, its substantial social and economic influx Ence has also made it an attractive target for malicious actors seeking to manipulate and influence public opinion and decision-making. X has for some time been a prime target for automated programs, or "bots," due to its open Existing research utilizes various characteristics of the social network to differentiate between human and auto mated accounts. These features include user activity patterns (e.g., tweet frequency, timestamps), account metadata (e.g., follower/following ratios, account age), and social network structures (e.g., retweet and mention networks) etc. Supervised ML models and deep neural networks have been widely employed for this purpose. Traditional bot detection systems such as heuristic methods fail against evolving spambots, network-based approaches depend on narrow social networks, and earlier ML models employ

lime tied characteristics, disregarding linguistic, temporal, and sentiment trends. Furthermore, the majority are not explaining able, which makes it challenging to evaluate the data. Our Interpretable AI-based model addresses these gaps by intel grating diverse feature sets. We enhance transparency with XAI which ensures improved accuracy, robustness, and inter probability.

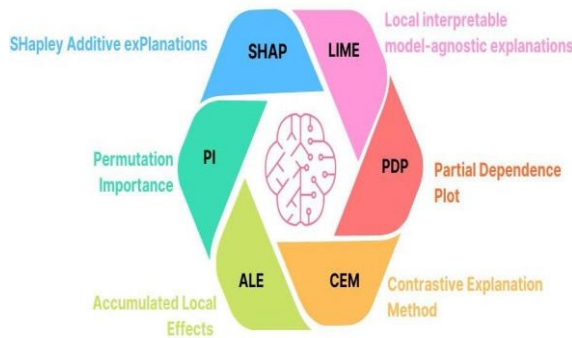


FIGURE 1. Interpretable AI techniques

Furthermore, clustering and anomaly detection methods have been explored for unsupervised detection of anomalous behaviours linked to bots. While these methods have shown promising results, they often lack scalability and adaptability due to their dependency on handcrafted feature engineering and static datasets. Moreover, the heavy reliance on black-box ML models limits their interpretability and creates barriers to understanding how decisions are made. Several challenges reduce the effectiveness of current bot detection methodologies. One of these challenges is feature engineering which is a labour-intensive process that requires domain expertise and manual effort to adapt the models to newer datasets and bots. Furthermore, bots exhibit dynamic and adaptive behaviour through the evolution of their strategies to mimic human users more effectively and evade detection algorithms. As a result, black-box detection models struggle to adapt to the constantly evolving nature of bot activities. Additionally, the lack of model interpretability in these methods undermines trust and transparency. Evaluation without interpretability is a challenge as we don't know if the model is identifying bots based on meaningful patterns or merely overfitting to noise in the data. Additionally, most methods are designed to optimize detection accuracy without considering the broader goals of generalizability and adaptability which are critical for real-world deployment on social networks. These gaps highlight the need for more transparent and interpretable detection framework.

2. LITERATURE REVIEW

The burgeoning interest in bot-detection challenges has precipitated a proliferation of academic inquiry, yielding a plethora of articles that proffer diverse methodologies. That notwithstanding, a gap exists in the extant literature, as the overwhelming majority of these approaches fail to provide transparent and interpretable results. In the subsequent sections, a concise review of prevailing bot detection strategies will be presented, accompanied by an examination of the challenges that necessitate further investigation. The preponderance of bot detection methodologies relies on supervised ML paradigms, which necessitate the utilization of one or multiple annotated datasets to train ML classifiers and develop an efficacious framework. These annotated datasets are frequently generated through human annotation, although alternative approaches such as leveraging pre-existing established models, crowdsourcing, or automated annotation techniques have also been employed to construct datasets for both detection purposes. Table 1 highlights the key literature for interpretable AI-based bot detection. This literature discusses the purpose of each study and our findings on each research.

3. SNBD METHODOLOGIES

In this study, the authors devised a novel approach by creating a corpus of honeypot accounts, specifically designed to attract spammer interactions, and subsequently logged the corresponding profile information. This dataset was then augmented with a collection of regular user profiles, thereby enabling the development of a comprehensive classification algorithm that incorporates both user-centric and content-centric features. In another research, authors took a similar method, attempting to detect botnets that were run by the same person. Reference employs crowdsourcing techniques for both recognition on Facebook, and while it appeared to provide decent results; however, the inherent limitations of this method became apparent when the perpetual evolution and proliferation of bots rendered the approach increasingly unsalable, thereby underscoring the need for more adaptive and dynamic bot detection strategies.

4. CHALLENGES OF SNBD

Despite the plethora of scientific endeavours that have yielded various methods for detecting online social bots, as indicated in the aforementioned studies, there are still many outstanding difficulties. Even though many SNBD approaches employ more than 1,000 attributes to train their method, it remains unclear whether increasing the number of features necessarily enhances model efficiency. Moreover, the authors of the highlighted significant impact of utilizing an extensive feature set on the scalability of bot detection systems. Interestingly, they also note that employing various subsets of publicly available

labelled datasets can enhance model generalizability, as observed in the same study. Notably, the performance of machine learning-based bot detection models varies across different datasets. Cones quaintly, the accumulation of additional datasets is essential to ensure that our training data encompasses a comprehensive range of bot behavioural features. The same conclusion is parties' tweets from before and throughout the 2017 election cycle, demonstrating an increase in the use of social bots. It is clear that Twitter bot identification is a difficult process that frequently needs thorough and robust treatment. Several ML-based methods, such as the Butternut have been offered with a total of 1200 distinct characteristics combined with an ML classifier. An enhanced version of this system, Bathometer, is detailed in , which needs X API keys to obtain user data during real-time calculations, making it inefficient to utilize real-time labelling tools in the case of large datasets. There is an increasing number of Twitter bot identification programs that use machine learning and data (statistical) analysis such as the Steeler, the Debut, and the Retweet-Buster (Robust) obtained from the ad, which geostationary-grained categorization of bots, giving distinct datasets for each sort of bot. As a result, one major difficulty in online social bot identification is determining what qualities genuinely constitute social bot. X bots are often used for malevolent objectives ranging from distributing fake news to propaganda and astroturfing. The writers of examined 245,000 profiles on X between the 2016 US presidential election and the 2018 midterm elections, detecting around 31,000 bots. The authors of conducted an exhaustive analysis of 43 million election-related tweets pertinent to the U.S. Congress investigation into Russian interference during the 2016 U.S. election campaigns.

5. CONCLUSION

This research presents a unique way to differentiate between bots and real users on X by using an interpretable ML frame work that extracts and analyses attributes for the the extraction of a diverse set of features derived from the datasets discussed in Section III-A. The model was trained on various features that were finalized through explainable AI techniques to improve the detection of social and spam bots as well as fake followers. This approach increased the accuracy and reliability of our model and gave important insights into potential patterns which enhanced transparency for social security. This is done through the incorporation of the XAI techniques SHAP and LIME into the model which allows the researchers to understand the impact of the features on the model. This information allowed us to reduce the size of the feature set to include the most important features which reduced the workload for the ML model. The significance of this study lies in its ability to bridge the gap between model accuracy and transparency thus addressing the key challenges in bot detection by offering an interpretable methodology.

6. REFERENCES

- [1] E. Canoe-Marin, M. Mora-Cantal lops, and S. Sánchez-Alonso, "Twitter as a predictive system: A systematic literature review," *J. Bus. Res.*, vol. 157, Mar. 2023, Art. no. 113561, doi: 10.1016/j.jbusres.2022.113561.
- [2] F. Tabassum, S. Mubarak, L. Liu, and J. T. Du, "How many features do we need to identify Bots on Twitter?" in *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity*, I. Ssemwanga, A. Goulding, H. Modulation-Sandy, J. T. Du, A. L. Soares, V. Hisami, and R. D. Frank, Eds., Cham, Switzerland: Springer, 2023, pp. 312–327.
- [3] R. Al-Azawi and S. O. AL-Memory, "Feature extractions and selection of bot detection on Twitter a systematic literature review," *Intelligences Arif.*, vol. 25, no. 69, pp. 57–86, Apr. 2022, due: 10.4114/intertie. vol25iss69pp57-86.
- [4] Zhigang A.A. Ghorbani, "An overview of online fake news: Chirac erization, detection, and discussion," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102025, due: 10.1016/j.ipm.2019.03.004.
- [5] Y. Bosham, I. Malakhov, K. Beznosov, and M. Ripeanu, "Design and analysis of a social botnet," *Compu. Newt.*, vol. 57, no. 2, pp. 556–578, Feb. 2013, Doi: 10.1016/j.comnet.2012.06.006.
- [6] Z. Yang, C. Wilson, X. Want. Gao, B. Y. Zhao, and Y. Dai, "Uncovering social network Sybils in the wild," *ACM Trans. Know. Discovery from Data*, vol. 8, no. 1, pp. 1–29, Feb. 2014, dui: 10.1145/2556609.
- [7] D. Javed, N. Z. Janghi, and N. A. Khan, "Football analytics for goal pre diction to assess player performance," in *Proc. Int. Conf. Innov. Technol. Sports (Reveal DNA ICITS)*, Apr. 2023, pp. 245–257, Doi: 10.1007/978 981-99-0297-2_20.
- [8] M. Humayun, Daved]hanjh i, M. F. Almufareh, and S. N. Almuayqil, "Deep learning based sentiment analysis of COVID-19 tweets via resam pling and label analysis," *Comput. Syst. Sci. Eng.*, vol. 47, no. 1, pp. 575–591, 2023.
- [9] S. N. Almuayqil, M. Humayun, N. Z. Jhanjhi, M. F. Almufareh, and D. Javed, "Framework for improved sentiment analysis via random minor ity oversampling for user tweet review classification," *Electronics*, vol. 11, no. 19, p. 3058, Sep. 2022, doi: 10.3390/electronics11193058.
- [10] F. Al-Quayed, D. Javed, N. Z. Jhanjhi, M. Humayun, and T. S. Alnusairi, "A hybrid transformer-based model for optimizing fake news detection," *IEEE Access*, vol. 12, pp. 160822160834, 2024, doi:10.1109/ACCESS.2024.3476432.

[11] D. Javed, N. Jhanjhi, N. A. Khan, S. K. Ray, A. A. Mazroa, F. Ashfaq, and S. R. Das, "Towards the future of bot detection: A comprehensive tax onomical review and challenges on Twitter/X," *Comput. Netw.*, vol. 254, Dec. 2024, Art. no. 110808, doi: 10.1016/j.comnet.2024.110808.

[12] S.Lundberg and S. Lee, "A unified approach to interpreting model pre dictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.(NIPS)*. Red Hook, NY, USA: Curran Associates Inc., Jan. 2017, pp. 4768–4777.

[13] M. Aljabri, R. Zagrouba, A. Shaahid, F. Alnasser, A. Saleh, and D. M. Alomari, "Machine learning-based social media bot detection: A comprehensive literature review," *Social Netw. Anal. Mining*, vol.13,no.1, pp. 1–40, Jan. 2023, doi: 10.1007/s13278-022-01020-5.

[14] K. Hayawi, S. Saha, M. M. Masud, S. S. Mathew, and M. Kausar, "Social media bot detection with deep learning methods: A systematic review," *Neural Compute. Appl.*, vol. 35, no. 12, pp. 8903–8918, Mar. 2023, doi: 10.1007/s00521-023-08352-z.

[15] S. Kaguta and E. Ferrara, "Deep neural networks for bot detect tion," *Inf. Sci.*, vol. 467, pp. 312–322, Oct. 2018, doi: 10.1016/j.ins.2018. 08.