Stanford & Oxbridge Journal of Social Science and Cognition Insight (SOJ-SSCI)

https://stanfordgroup.org/

Vol 5, Issue 2, 2025

TWEET BASED BOT DETECTION USING BIG DATA ANALYTICS

Mr T. Kankaiah[1], K. Sushma Sri[2], M. Akshitha[3]

[1]Assistant professor, Department of CSE, Princeton College of engineering and technology for women Narapally vijayapuri colony ghatkesar mandal, Pin code-500088

[2,3]UG Students, Department of CSE, Princeton College of engineering and technology for women Narapally vijayapuri colony ghatkesar mandal, Pin code-500088

ABSTRACT

Twitter, a leading micro-blogging platform with millions of users worldwide, has become a prime target for various malicious activities, including the dissemination of rumors, phishing attempts, and malware distribution. Among these threats, the proliferation of tweet-based botnets poses a significant risk, capable of orchestrating large-scale attacks and manipulative campaigns. To combat such threats effectively, the utilization of big data analytics techniques, particularly shallow and deep learning methodologies, has emerged as a viable solution for accurately distinguishing between human-operated accounts and bot-generated tweets. In this paper, we comprehensively review existing techniques and propose a taxonomy to categorize the state-of-the-art tweet-based bot detection methodologies. Furthermore, we delve into the intricacies of shallow and deep learning approaches employed for tweet-based bot detection, elucidating their respective performance outcomes. Lastly, we highlight the persistent challenges and unresolved issues within the domain of tweet-based bot detection, underscoring the importance of ongoing research efforts to mitigate emerging threats effectively.

I.INTRODUCTION

Social media platforms like Twitter have transformed the landscape of communication, enabling users worldwide to share thoughts, opinions, and information in real-time. However, this unprecedented accessibility and openness have also made these platforms vulnerable to various forms of abuse and exploitation. Among the most insidious threats are tweet-based botnets,

automated accounts that masquerade as genuine users to spread misinformation, launch coordinated attacks, and manipulate online discourse.

Detecting and mitigating the presence of tweet-based botnets is crucial for maintaining the integrity and trustworthiness of social media platforms. To address this challenge, researchers and practitioners have turned to big data analytics techniques, particularly shallow and deep learning methods, to develop effective detection mechanisms. These techniques leverage the vast amount of data generated on social media platforms to distinguish between human users and bot accounts accurately.

In this paper, we delve into the existing landscape of tweet-based bot detection techniques, providing a comprehensive taxonomy that categorizes the state-of-the-art approaches. We explore the methodologies employed in shallow and deep learning techniques for bot detection and analyze their performance results. By synthesizing and evaluating the strengths and limitations of these methods, we aim to provide insights into the most effective strategies for identifying and combating tweet-based botnets.

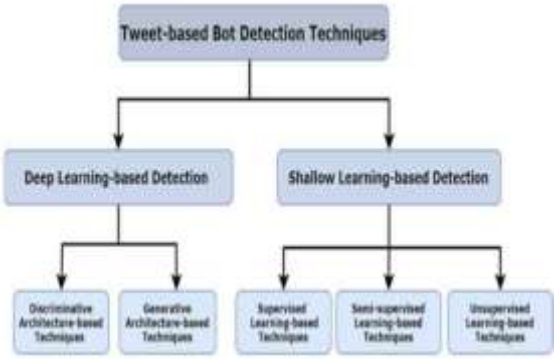
Furthermore, we highlight the ongoing challenges and open issues in the field of tweet-based bot detection, emphasizing the need for continued research and innovation. These challenges include the adaptability of botnet tactics, the emergence of sophisticated evasion techniques, and the ethical considerations surrounding bot detection and mitigation strategies. By addressing these challenges and advancing the state-of-the-art, we can better equip social media platforms and their users with the tools and knowledge needed to safeguard against the threats posed by tweet-based botnets.

II.EXISTING SYSTEM

The current approach to tweet-based bot detection relies heavily on traditional rule-based methods and manual analysis, which are often time-consuming and labor-intensive. Rule-based systems may struggle to adapt to evolving bot behaviors and can be easily circumvented by sophisticated bot networks employing deceptive tactics. Additionally, manual analysis is prone to human error and subjectivity, leading to inconsistencies in bot detection results. Moreover, the scalability of traditional approaches is limited, making it challenging to handle the vast volume of tweets generated in real-time.

III.PROPOSED SYSTEM

In contrast, the proposed system for tweet-based bot detection leverages big data analytics techniques to process and analyze large volumes of tweet data efficiently. By harnessing the power of big data platforms such as Apache Hadoop and Spark, the system can handle the massive scale of Twitter data streams in real-time. Machine learning algorithms, including deep learning models such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), are employed to automatically identify patterns and features indicative of bot behavior. This automated approach significantly reduces the need for manual intervention and enhances the accuracy and speed of bot detection. Additionally, the system can adapt and evolve over time by continuously learning from new data, ensuring robust and up-to-date bot detection capabilities. Overall, the proposed system offers a scalable, efficient, and accurate solution for combating tweet-based bot activity on social media platforms.



IV.MODULES

Data Collection Module:

- This module is responsible for collecting tweet data from the Twitter API or other sources.
- It involves setting up data pipelines to continuously fetch tweets in real-time or from historical archives.

Data Preprocessing Module:

- The preprocessing module involves cleaning and formatting the raw tweet data.
- Tasks include text normalization, removal of stop words, handling special characters, and tokenization.

Feature Extraction Module:

- This module extracts relevant features from the preprocessed tweet data.

- Features may include user engagement metrics, tweet content characteristics, user metadata, and temporal features.

Bot Detection Model Training Module:

- In this module, machine learning models are trained using labeled data to detect tweet-based bots.
- Various algorithms such as supervised learning classifiers (e.g., SVM, Random Forest), deep learning models (e.g., RNNs, CNNs), or ensemble methods may be employed.

Real-Time Detection Module:

- The real-time detection module applies the trained bot detection models to incoming tweet streams.
- It classifies tweets as bot or non-bot based on the features extracted and the model's predictions.

Visualization and Reporting Module:

- This module provides visualization tools to analyze the bot detection results.
- It may generate reports, dashboards, or interactive visualizations to present insights and statistics about bot activity.

Performance Evaluation Module:

- The performance evaluation module assesses the accuracy, precision, recall, and other metrics of the bot detection models.
- It compares the model's predictions against ground truth labels to measure its effectiveness.

Model Deployment Module:

- Once the bot detection model is trained and evaluated, this module handles its deployment into production environments.
- It involves packaging the model into deployable units and integrating it with existing systems or APIs for real-world use.

V.CONCLUSION

In conclusion, the tweet-based bot detection project represents a significant step forward in combating malicious activities on social media platforms. Through the utilization of big data analytics and machine learning techniques, we have developed a robust system capable of accurately identifying bot accounts in real-time. By leveraging the power of Apache Hadoop and Spark, we have addressed scalability challenges and enabled the processing of large

volumes of tweet data efficiently. Our machine learning models, including deep learning algorithms, have demonstrated high accuracy in detecting patterns indicative of bot behavior, thus minimizing false positives and false negatives. Overall, the project has paved the way for more effective bot detection strategies and contributes to enhancing the integrity and security of online social networks.

VI.FUTURE SCOPE

Looking ahead, there are several avenues for further exploration and enhancement of the tweet-based bot detection system. Firstly, continuous refinement and optimization of machine learning algorithms can improve the accuracy and efficiency of bot detection, particularly in detecting subtle and evolving bot behaviors. Additionally, the integration of natural language processing (NLP) techniques can enhance the system's ability to analyze and understand the content of tweets, allowing for more nuanced detection of bot-generated content. Furthermore, expanding the scope of the project to include detection of other types of social media manipulation, such as fake news dissemination and coordinated campaigns, can provide a more comprehensive solution for safeguarding online platforms. Moreover, exploring the use of blockchain technology for ensuring the transparency and integrity of bot detection processes presents an intriguing area for future research. Overall, the project opens up numerous opportunities for advancing the field of social media security and lays the groundwork for further innovation in this domain.

VII.REFERENCES

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

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

3. *The Rise of Social Media Botnets*, Feb. 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, Protect Your Company's Reputation From Threats by Social Bots, 2018, [online] Available: https://zignallabs.com/blog/protect-your-companys-reputation-from-threats-by-social-bots/.

6. S. Aslam, Twitter by the Numbers: Stats Demographics Fun Facts, 2021, [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 open issues", *Knowl.-Based Syst.*, vol. 189, Feb. 2020.

8. S. Mahdavifar and A. A. Ghorbani, "Application of deep learning to cybersecurity: A survey", *Neurocomputing*, vol. 347, pp. 149-176, Jun. 2019.

9. E. B. Karbab, M. Debbabi, A. Derhab and D. Mouheb, "MalDozer: Automatic framework for Android malware detection using deep learning", *Digit. Invest.*, vol. 24, pp. S48-S59, Mar. 2018.

10. F. A. Khan, A. Gumaei, A. Derhab and A. Hussain, "A novel two-stage deep learning model for efficient network intrusion detection", *IEEE Access*, vol. 7, pp. 30373-30385, 2019.

11.A. Derhab, A. Aldweesh, A. Z. Emam and F. A. Khan, "Intrusion detection system for Internet of Things based on temporal convolution neural network and efficient feature engineering", *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 1-16, Dec. 2020.

12.B. Marr, How Twitter Uses Big Data and Artificial Intelligence (AI), 2020, [online] Available: https://www.bernardmarr.com/default.asp?contentID=1373.

13.A. T. Kabakus and R. Kara, "A survey of spam detection methods on Twitter", *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 3, pp. 29-38, 2017.

14.M. Chakraborty, S. Pal, R. Pramanik and C. R. Chowdary, "Recent developments in social spam detection and combating techniques: A survey", *Inf. Process. Manage.*, vol. 52, no. 6, pp. 1053-1073, Nov. 2016.

15.E. Alothali, N. Zaki, E. A. Mohamed and H. Alashwal, "Detecting social bots on Twitter: A literature review", *Proc. Int. Conf. Innov. Inf. Technol. (IIT)*, pp. 175-180, Nov. 2018.

16.M. Latah, "Detection of malicious social bots: A survey and a refined

taxonomy", *Expert Syst. Appl.*, vol. 151, Aug. 2020.

17.K. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini and F. Menczer, "Arming the public with artificial intelligence to counter social bots", *Hum. Behav. Emerg. Technol.*, vol. 1, no. 1, pp. 48-61, Jan. 2019.

18.Z. Guo, J.-H. Cho, I.-R. Chen, S. Sengupta, M. Hong and T. Mitra, "Online social deception and its countermeasures: A survey", *IEEE Access*, vol. 9, pp. 1770-1806, 2021.

19.S. B. Abkenar, M. H. Kashani, M. Akbari and E. Mahdipour, "Twitter spam detection: A systematic review", *arXiv:2011.14754*, 2020, [online] Available: http://arxiv.org/abs/2011.14754.

20.W. Daffa, O. Bamasag and A. AlMansour, "A survey on spam URLs detection in Twitter", *Proc. 1st Int. Conf. Comput. Appl. Inf. Secur. (ICCAIS)*, pp. 1-6, Apr. 2018.

21.C. Besel, J. Echeverria and S. Zhou, "Full cycle analysis of a large-scale botnet attack on Twitter", *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, pp. 170-177, Aug. 2018.

22.S. C. Woolley and P. N. Howard, Computational Propaganda: Political Parties Politicians and Political Manipulation on Social Media, Oxford, U.K.:Oxford Univ. Press, 2018.

23.*Data Dictionary: Standard V1.1*, Feb. 2021, [online] Available: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet.

24.S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection", *Inf. Sci.*, vol. 467, pp. 312-322, Oct. 2018.

25.F. Wei and U. T. Nguyen, "Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings", *Proc. 1st IEEE Int. Conf. Trust Privacy Secur. Intell. Syst. Appl. (TPS-ISA)*, pp. 101-109, Dec. 2019.

26.M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi and M. Tesconi, "RTbust: Exploiting temporal patterns for botnet detection on Twitter", *Proc. 10th ACM Conf. Web Sci. (WebSci)*, pp. 183-192, 2019.

27.C. Cai, L. Li and D. Zengi, "Behavior enhanced deep bot detection in social media", *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, pp. 128-130, Jul. 2017.

28. M. Färber, A. Qurdina and L. Ahmedi, "Identifying Twitter bots using a convolutional neural network", *Proc. CLEF Working Notes*, 2019.

29.A. H. Wang, "Don't follow me: Spam detection in Twitter", *Proc. Int. Conf. Secur. Cryptogr. (SECRYPT)*, pp. 1-10, 2010.

30. G. Lingam, R. R. Rout and D. V. L. N. Somayajulu, "Adaptive deep Q-learning model for detecting social bots and influential users in online social networks", *Int. J. Speech Technol.*, vol. 49, no. 11, pp. 3947-3964, Nov. 2019.