A Review of Social Bot Detection Methods

Weijie Mo
Xihua University
Chengdu, Sichuan, China
Email: 212024085404015@stu.xhu.edu.cn

Abstract

Social media bot (automated accounts) attacks represent a form of organized malicious activity that can pose significant threats to public opinion, democratic processes, public health, financial markets, and other areas. While researchers have been actively developing various models to detect these bots, attackers continue to evolve their techniques to evade detection. This keeps the field both challenging and fast-evolving, continuously pushing the development of more advanced detection technologies. To support future research and practical solutions, this paper presents a systematic review of social media bot attacks, existing detection methods, and the key challenges in this domain. The review includes a refined taxonomy of detection techniques, an overview of the main technologies used to identify bots on social media, and a comparative analysis of current approaches. Several important gaps in the literature are identified: most existing studies primarily focus on the Twitter platform and rely heavily on supervised machine learning methods; publicly available datasets are often limited in size or lack sufficient accuracy; there is a clear need for integrated systems capable of real-time detection. Furthermore, raising public awareness is essential to equip regular users with the knowledge to recognize and defend against bot activity.

Index Terms

Social Media, Bot, Socialbot, Malicious attack, Attack detection, Detection methods taxonomy, Systematic literature review

I. Introduction

The emergence of social media platforms such as Twitter, Facebook, and Instagram brought a revolutionary transformation in communication tools. These internet-based applications allow users to create and exchange user-generated content [1]. Social media rapidly penetrated into the daily lives of many people, possessing the potential not only to transform communication methods but also to shape opinions and influence lives [2]. It has become a fundamental component of communication infrastructure for both individuals and organizations. For instance, Facebook had 3.065 billion users in the fourth quarter of 2023, representing nearly two-fifths of the global population [3].

As the influence of social media increased, it became a tool accessible to everyone. While legitimate uses exist, numerous influence seekers and malicious actors exploit social media for hidden agendas. To harness the power of social media, individuals and organizations seek to gain influence, thus the emergence of Social Media Bots (SMB). Social Media Bots are computer programs capable of generating content and interacting with users [4].

Bots account for a significant portion of online activity. According to Twitter, approximately 8.5% of its users are bots [5]. A study on social bots reveals that 9% to 15% of English-speaking active users on Twitter exhibit bot-like behavior [6]. SMB may serve benign, neutral, or malicious purposes [7]. Examples of benign bots include those that automatically post earthquake alerts, chatbots that interact with users and fulfill their needs [8], and news bots that disseminate news content for media outlets. Neutral bots may post or repost jokes [9] or share nonsensical content [10].

Malicious bots, the most extensively studied category in social media, appear in various forms, and new types continue to emerge [11]. Malicious SMB are typically controlled by a human operator, or botmaster, who manages their actions and orchestrates attacks. Common types include spam bots that spread unsolicited messages, political bots that engage in political discourse, and Sybils—fake accounts used to build deceptive influence.

Fake profiles (such as SMB) are among the most critical security threats to Online Social Networks (OSNs) [12]. Moreover, studies suggest that SMB pose risks to public health, as people increasingly seek medical advice online while bots are employed to promote products and manipulate discussions [13]. These threats are real and the consequences can be severe. In 2010, bots infiltrated U.S. midterm election discussions on Twitter, supporting certain candidates while discrediting others [14]. A similar incident occurred in the 2016 U.S. presidential elections, where nearly one-fifth of the conversation on Twitter was generated by bots and over 5,000 accounts were subsequently suspended by Twitter [15]. In 2019, another 5,000 bots were involved in pushing protests against the “Russiagate hoax,” and were later suspended as well.

Bots have also interfered with debates about vaccinations, spreading biased content to sway public opinion [16], [4]. In one instance, a bot-driven campaign generated discussion around a dormant company, inflating its market value to $5 billion before it was suspended [4].

These incidents are only the exposed examples; more attacks may remain undiscovered and new ones are likely to surface. If successful, bots have the potential to manipulate

public opinion, jeopardizing democracy, public health, and financial markets. Additionally, they erode trust in social media, as bots are tools for disseminating fake reviews, fake news, fake sentiments, fake followers, and fake likes.

As a result, researchers have shown great interest in identifying SMB and equipping platforms and users with tools for defense. Some focus on classifying automatically generated content rather than detecting bot accounts [17], while others work on predicting vulnerable populations [18]. However, the most widely used strategy remains detecting bot accounts, whether individually or in botnets, and at various stages, from creation to integration into social media.

Researchers have responded to this need. A taxonomy of detection techniques was proposed by Ferrara et al. [4] and adopted by later reviews [19], [20]. But now a systematic review is warranted—one that adheres to a scientific search strategy to ensure comprehensive coverage of detection models.

REFERENCES

[1] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59–68.

[2] Luxton, D. D., June, J. D., & Fairall, J. M. (2012). Social Media and suicide: A public health perspective. *American Journal of Public Health*, 102(S2), S195–S200.

[3] Population: the numbers (2023). Retrieved from: https://ourworldindata.org/world-population-growth

[4] Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7), 96–104.

[5] Subrahmanian, V. S., et al. (2016). The DARPA Twitter bot challenge. *Computer*, 49(6), 38–46.

[6] Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *ICWSM*.

[7] Stieglitz, S., Brachten, F., Ross, B., & Jung, A. K. (2017). Do social bots dream of electric sheep? *ECIS*.

[8] Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). A new chatbot for customer service on social media. *CHI*.

[9] Veale, T., Valitutti, A., & Li, G. (2015). Twitterbots that tell jokes. *ACL Workshop on Computational Humor*.

[10] Wilkie, A., Michael, M., & Plummer-Fernandez, M. (2015). Speculative method and Twitter bots. *Disruptive Social Science?*.

[11] Cresci, S., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2019). Cashtag piggybacking: spam and bot activity in stock microblogs. *ACM Transactions on the Web*, 13(2), 1–27.

[12] Viswanath, B., Post, A., Gummadi, K. P., & Mislove, A. (2011). An analysis of social network-based sybil defenses. *ACM SIGCOMM Computer Communication Review*, 41(4), 363–374.

[13] Allem, J. P., & Ferrara, E. (2018). Could social bots pose a threat to public health? *American Journal of Public Health*, 108(8), 1005–1006.

[14] Ratkiewicz, J., et al. (2011). Detecting and tracking political abuse in social media. *ICWSM*.

[15] Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 US Presidential election online discussion. *First Monday*, 21(11).

[16] Broniatowski, D. A., et al. (2018). Weaponized health communication: Twitter bots and trolls in the vaccine debate. *AJPH*, 108(10), 1378–1384.

[17] Almerekhi, H., & Elsayed, A. (2015). Detecting automatically generated tweets using classifiers. *International Journal of Computer Applications*, 120(16).

[18] Halawa, S., Greene, D., & Cunningham, P. (2016). Who will retweet this? Detecting strangers who retweet your messages. *ICWSM*.

[19] Alothali, E., Zaki, N., Mohamed, E. A., & Alashwal, H. (2018). Detecting social bots on Twitter: A literature review. *2018 IEEE Intl Conf on Innovations in Information Technology (IIT)*.

[20] Karataş, G., & Şahin, F. (2017). A review on social bot detection techniques. *2017 International Conference on Computer Science and Engineering (UBMK)*.