State of the Art V3

March 10, 2025

1 Introduction to the Problem

The growth of social networks has revolutionized digital communication, enabling instant information dissemination and facilitating global interactions. However, this advancement has also brought significant challenges, such as the use of bots to manipulate interactions, amplify biased discourse, and influence social and political processes. Studies estimate that between 9% and 15% of active Twitter/X users exhibit automated behaviors, while data from the platform itself indicates that approximately 8.5% of accounts are bots, highlighting the magnitude of the problem.

Malicious bots are widely used to spread misinformation, artificially inflate engagement metrics, and compromise digital security. With the advancement of artificial intelligence, these automated accounts have become more sophisticated, capable of convincingly simulating human interactions, making their detection increasingly challenging.

The impact of bots extends beyond public opinion manipulation, affecting critical sectors such as healthcare, economics, and political stability. Studies show that bots played a role in spreading misinformation about vaccines, reducing immunization adherence, and undermining trust in health institutions. In the financial sector, bots have been responsible for manipulating company valuations, causing atypical market movements. Additionally, in politics, notable cases include interference in the 2016 U.S. elections, where about 20% of Twitter interactions were generated by bots, and the coordinated spread of fake news to influence political protests.

With the increasing sophistication of bots, it is essential to develop more advanced and adaptable detection methods. Traditional rule-based approaches have proven insufficient against increasingly realistic and dynamic bots. Thus, this study proposes an innovative solution that combines artificial intelligence, machine learning, and community participation, integrating different techniques to enhance the identification and blocking of malicious bots, making social networks safer and more authentic.

2 Types of Bots in Social Networks

Bots play diverse roles in social networks and can be classified based on their intentions and ability to mimic human behavior. Bots can be benign, neutral, or malicious. Benign bots, such as those that automatically publish news articles or virtual assistants used in customer service, do not pose threats and are not targeted by detection systems. On the other hand, malicious bots are responsible for spreading misinformation, manipulating metrics, and exploiting digital vulnerabilities. Since this study focuses on detecting harmful bots, the following classification considers only malicious bots.

The categorization of malicious bots is not entirely rigid, as some types share similar objectives and strategies. However, it is possible to group these bots into different classes according to their functions and impact on social networks.

2.1 Main Types of Malicious Bots

Malicious bots can be classified into different categories based on their digital manipulation strategies:

- **Cashtag Piggybacking Bots**: Create coordinated campaigns to promote low-value stocks by exploiting the popularity of high-value stocks.

- **Spam Bots**: Distribute malicious links, send unsolicited messages, and exploit trending topics to maximize the spread of unwanted content.
- **Astroturfing Bots**: Create the illusion of massive support for a candidate or idea, artificially manipulating public debates.
- **Sybils**: Use pseudonymous identities to exert disproportionate influence in online discussions, artificially generating a large volume of interactions.
- **Fake Accounts for Botnet Command & Control**: Use social networks as command channels for botnet attacks, embedding encrypted instructions in seemingly harmless posts.
- **Pay Bots**: Illicitly generate profits by copying reputable content and incorporating micro-URLs that redirect users to traffic-paying sites.
- **Social Botnets in Political Conflicts**: Interfere in political debates, propagating biased or irrelevant content to divert attention from legitimate discussions.
- **Infiltrators in Organizations**: Pose as friends or colleagues on social networks to obtain sensitive information about users or organizations.
- **Influence Bots**: Create realistic automated identities to artificially shape discussions and persuade users to adopt specific opinions.
- **Doppelgänger Bots**: Clone real user profiles to create authentic-looking fake identities used in malicious activities such as phishing and fraud.

2.2 Impact of Malicious Bots

Different types of bots perform specific functions that, when combined, can have devastating consequences in the digital society. Spam bots and astroturfing bots amplify biased discourse, creating a false impression of consensus. Influence bots and sybils distort public debate by generating artificial interactions. Doppelgänger bots and infiltrators exploit individual vulnerabilities, increasing the risks of cyberattacks and identity theft. Political social botnets directly interfere in elections and governmental decisions, while pay bots and cashtag piggybacking bots manipulate financial markets to obtain illicit profits.

As malicious bots become more sophisticated, their tactics become harder to detect. Many use advanced artificial intelligence models to simulate authentic human interactions, rendering traditional detection methods ineffective.

3 Review of Existing Work

In recent years, various initiatives have been developed to identify and combat the proliferation of bots on social networks. These tools employ different approaches, ranging from rule-based heuristics to advanced artificial intelligence and machine learning algorithms. Below, three of the main existing solutions are analyzed: Pegabot, LiveDune, and Bot Sentinel, highlighting the methodologies adopted, their advantages, and limitations.

3.1 Pegabot

PEGABOT was launched in 2018 by the Institute of Technology and Society of Rio de Janeiro (ITS Rio) and the Technology & Equity Institute, with funding from the European Union. Its objective is to allow any user to analyze Twitter/X accounts to determine the probability of being operated by bots, based on public information.

Pegabot calculates an automation probability index based on three main components. The user profile analysis considers data such as name, description, number of followers and followings, number of tweets, and favorites, assigning a higher suspicion score to recent profiles with short descriptions, no profile picture, or unusual name patterns. The interaction network analysis examines hashtags and mentions in the user's posts to identify typical patterns of automated behavior, allowing the detection of activities related to spam dissemination or monothematic content. Finally, the sentiment analysis evaluates the last 100 published tweets, classifying them as positive, negative, or neutral. Bots tend to exhibit a more pronounced bias in their content, with lower emotional diversity, which facilitates their identification.

Pegabot has some limitations that may affect its effectiveness in bot detection. The analysis model can generate false positives, especially for legitimate profiles that frequently post about a single topic, causing authentic accounts to be mistakenly classified as bots. Additionally, the tool is currently unavailable (”Internal Server Error” in 2025), suggesting a lack of continuous maintenance and possibly indicating that its detection system is not up to date to handle the new strategies adopted by modern bots.

3.2 LiveDune

LiveDune is a social media management platform widely used by businesses to monitor statistics, schedule posts, and analyze competitors. One of its most relevant features is the bot verification on Instagram, which helps prevent advertising campaigns from being targeted at profiles with artificially inflated metrics.

Its algorithm identifies fake followers and artificial engagement through a detailed profile analysis, performed in a fast process that takes approximately 30 seconds. To determine the authenticity of an account, the platform evaluates three main factors. The general profile information analysis considers the total number of followers, the engagement rate (ER), and the average number of likes and comments, flagging profiles with abnormally low engagement or irregular growth as suspicious. The audience engagement assessment compares the relationship between likes and comments, identifying discrepancies that may indicate automated interactions, and uses visual indicators to facilitate result interpretation. Finally, the monthly account evolution is analyzed to detect sudden spikes and drops in the number of followers, which may suggest the purchase of followers or other forms of artificial profile growth manipulation.

LiveDune offers significant advantages for businesses and influencers looking to ensure the authenticity of their audience. The tool allows the filtering of legitimate followers, preventing investments in inauthentic profiles and making advertising campaigns more effective. Additionally, it generates detailed reports on account authenticity quickly, enabling users to make informed decisions in a short time. However, the platform has some limitations, as its primary focus is on analyzing influencers and advertising campaigns, which reduces its effectiveness in detecting bots used for misinformation dissemination. Moreover, account verification is a paid service, costing $15.30 per month for 5 accounts, which limits access for users who are unwilling or unable to invest in the tool.

3.3 Bot Sentinel

Bot Sentinel, created in 2018 by Christopher Bouzy, is a platform designed to combat disinformation and targeted harassment on Twitter/X. Unlike other tools, its focus is not only on bot identification but also on analyzing accounts that promote toxic trolls, hate speech, and manipulation campaigns.

Bot Sentinel uses a classification system based on artificial intelligence and machine learning, trained with thousands of accounts and millions of tweets to identify suspicious behaviors and content manipulation. The analysis process occurs in three stages. The first phase is data collection, where a user’s tweets are automatically analyzed, considering speech patterns, posting frequency, and interactions to detect anomalous behaviors. Next, there is classification based on Twitter’s rules, where the model recognizes accounts that repeatedly violate the platform’s guidelines, without considering factors such as ideology, religion, or geographical location. Finally, a trustworthiness score is assigned, ranging from 0% to 100%, indicating the likelihood of an account being involved in harassment, trolling, or manipulation campaigns, with higher scores corresponding to a greater degree of suspicion. In addition to the web version, the platform includes a browser extension, allowing users to evaluate profiles directly while browsing Twitter, making the identification of suspicious accounts more accessible and practical.

Bot Sentinel offers significant advantages in identifying manipulative accounts and fighting disinformation. With an accuracy of 95%, the tool effectively classifies profiles involved in harassment, trolling, and manipulation campaigns. Unlike other solutions that merely detect bots, Bot Sentinel also focuses on toxicity and disinformation spread, making it a useful tool for improving the quality of online discussions. Furthermore, it is a free and accessible platform, allowing any user to utilize it without financial restrictions. However, the system has some limitations. Since it relies on platform rules, it may fail to detect automated accounts that operate outside typical toxicity patterns, allowing some sophisticated bots to evade detection. Additionally, its analysis is exclusive to Twitter/X, not covering other social networks, which limits its applicability in a landscape where disinformation spreads across multiple digital platforms.

3.4 Comparison of Tools

Below is a comparison of the main bot detection tools, highlighting their differences in detection methods, social network focus, and limitations.

- **PEGABOT**
  - **Main Social Network:** Twitter/X
  - **Detection Method:** Heuristic Rules + Network Analysis
  - **Analysis Focus:** Probability of a profile being a bot
  - **Limitations:** May generate false positives; inactive as of 2025
- **LiveDune**
  - **Main Social Network:** Instagram
  - **Detection Method:** Statistical Metric Analysis
  - **Analysis Focus:** Verification of fake followers and artificial engagement
  - **Limitations:** Focused on influencers; paid service
- **Bot Sentinel**
  - **Main Social Network:** Twitter/X
  - **Detection Method:** AI + Machine Learning
  - **Analysis Focus:** Identification of manipulative behavior
  - **Limitations:** Limited to Twitter; does not identify ”neutral” bots

Each of these tools employs a distinct method for identifying and analyzing bots, ranging from heuristic, statistical, and AI-based approaches. However, no solution is universally effective, as the continuous evolution of bots necessitates constant adaptation of detection strategies.

3.5 Synthesis of the Review

The review of existing tools reveals that, while each has specific advantages, there are significant gaps in bot detection, such as:

1. Lack of integration across social networks – Most solutions analyze only one platform, making it difficult to identify bots that operate in a coordinated manner across multiple networks.
2. Limited focus on disinformation – Some tools, like LiveDune, prioritize engagement authenticity but do not detect bots that spread fake news and manipulation.
3. Need for more sophisticated techniques – The evolution of Large Language Models (LLMs) has made bots harder to detect, requiring more advanced approaches.

Based on these limitations, our project proposes an innovative approach that combines machine learning, semantic analysis, and behavioral pattern detection, providing a more effective and comprehensive system in the fight against bots on social networks.

4 Our Solution: Bot Blocker

Building upon the analysis of existing tools, which have demonstrated both advancements and limitations in bot detection, we propose an innovative solution that combines the strengths of these approaches while addressing their main weaknesses. Bot Blocker emerges as an alternative that integrates artificial intelligence, community participation, and expert supervision, ensuring more precise and adaptable detection.

While tools like Pegabot, LiveDune, and Bot Sentinel rely on heuristic rules, statistical methods, or machine learning to identify bots, Bot Blocker expands this approach by allowing direct user interaction, providing a transparent and intuitive voting system. The platform reinforces online transparency and accountability, enabling users to view account classification histories, analyze voting patterns, and identify biases, creating a safer and more trustworthy environment.

Moreover, bot detection in Bot Blocker is not limited to automated assessments. The platform offers a simple voting process, in which any user can evaluate suspicious accounts and justify their choices. This system, combined with an intuitive and responsive design, allows even non-technical users to actively participate in bot identification, making the process more inclusive and efficient.

Another key feature of the platform is the personalized and community-managed blocklist, allowing each user to control which accounts they wish to avoid while also having the option to follow the community blacklist, which is maintained and updated based on collective reports. This functionality surpasses the limitations of previous tools, which do not offer users the ability to manage blocks dynamically and personally.

To ensure fair evaluations and prevent abuse, Bot Blocker implements a verification hierarchy with verifiers and administrators, who monitor suspicious activities, prevent false positives, and refine algorithms to keep up with evolving bot strategies. This model balances community participation and human supervision, ensuring that bot detection remains fast, accurate, and resistant to manipulation attempts.

Thus, Bot Blocker not only addresses the issues identified in previous solutions but also introduces a new paradigm in bot detection, combining advanced technology, collective intelligence, and continuous adaptation to combat manipulation and disinformation on social networks.

5 Differences and Improvements Over Existing Solutions

Compared to tools such as Pegabot, LiveDune, and Bot Sentinel, Bot Blocker innovates by combining artificial intelligence with community participation and allowing personalized block management. Additionally, its application is not restricted to a single social network, making it more flexible and effective in identifying bots and online manipulation. The main improvements include:

5.1 Feature Comparison

The following list highlights how Bot Blocker improves upon existing solutions:

- **Identification Method**
  - Pegabot: Heuristics + Rules
  - LiveDune: Statistical Metric Analysis
  - Bot Sentinel: Machine Learning + AI
  - **Bot Blocker (Proposed):** Verifier Analysis + Community Participation
- **Main Social Network**
  - Pegabot: Twitter/X
  - LiveDune: Instagram
  - Bot Sentinel: Twitter/X
  - **Bot Blocker (Proposed):** Applicable to multiple social networks
- **Focus on Disinformation**

  - Pegabot: Partial
  - LiveDune: No
  - Bot Sentinel: Yes
  - **Bot Blocker (Proposed):** Yes (Enhanced Detection)

- **User Interaction**
  - Pegabot: None
  - LiveDune: None
  - Bot Sentinel: Score visualization only
  - **Bot Blocker (Proposed):** Voting System and Community Feedback

- **Block Management**
  - Pegabot: Not available
  - LiveDune: Not available
  - Bot Sentinel: Trustworthiness score
  - **Bot Blocker (Proposed):** Personal and Community Blocklists

5.2 Key Differentiators

- Bot detection across multiple social networks, whereas existing tools are limited to specific platforms.
- Community voting system, allowing users themselves to participate in the classification process.
- Human verification mechanisms, reducing false positives and unfair classifications.
- Flexible block management, offering users greater personalization.

Thus, Bot Blocker proposes a more comprehensive, interactive, and adaptable system, significantly improving existing methods in the fight against bots on social networks.

6 Conclusion

Bot Blocker emerges as an innovative solution for identifying and blocking bots, combining community participation with a human verification system. Its key advantages include transparency, personalization, and effectiveness, giving users the power to manage their online experience more securely and informatively. With a collaborative approach and advanced technology, this tool represents a significant step toward making social networks more authentic and protected against automated manipulation.

References

M. Orabi, D. Mouheb, Z. Al Aghbari, and I. Kamel, "Detection of Bots in Social Media: A Systematic Review," *Information Processing & Management*, vol. 57, no. 4, p. 102250, 2020. Available at: ScienceDirect. DOI: 10.1016/j.ipm.2020.102250.

Instituto de Tecnologia e Sociedade do Rio de Janeiro (ITS Rio) and Instituto Tecnologia & Equidade, *Let's talk about transparency in the use of bots on social media?*, 2025. Available at: PegaBot. Accessed on: March 10, 2025.

Livedune, *How to check if there are bots in your Instagram account*, June 4, 2024. Available at: Livedune. Accessed on: March 10, 2025.

Bot Sentinel Inc., *More than just bots...*, 2025. Available at: Bot Sentinel. Accessed on: March 10, 2025.