Behind the Chat: Agent-Based Modeling of Twitch Mod Bots vs. Spambots

GitHub Repository URL**: https://github.com/Peterayade/EECS-4461-Project.git**
Team Information - Team Number**: Team 5**
- Team Members**:
- Peter AYADE
- Mingran CHEN
- Shenice THOMAS

Section 1: Phenomenon of Interest

On live streaming platforms like Twitch, chat rooms are the main space for viewers to interact with streamers. But as live streaming becomes more popular, human and bot accounts that send spam or ads have begun to flood in.Human accounts that send ads and spam messages try to "steal" the popularity of the current live-streaming room to achieve their own publicity purposes. Compared with bot accounts, malicious human accounts are fewer in number, but more intelligent and more difficult to be detected and managed by mod bots.

Bot accounts use automated scripts to send spam ads, post malicious links, and disrupt the chat experience. To deal with these interruptions, Twitch relies on mod bots to automatically detect and block spam. However, this is not just a simple "ban and move on" situation, it's an ongoing battle. Spam bots are constantly evolving to evade detection, and mod bots must constantly adapt to keep up.

One solution to understanding the ongoing battle between mod bots and spambots is Agent-Based Modeling (ABM), which treats both as intelligent agents following decision rules and interacting dynamically. By simulating different strategies, we can observe how spambots evolve—modifying message formats, reducing posting frequency, or mimicking human conversation, while mod bots refine their algorithms to improve accuracy and minimize false bans. This phenomenon highlights a critical issue in digital media ecosystems, where AI-driven moderation must continuously adapt to evolving threats. Twitch serves as a real-world case study of AI-to-AI interaction, where automated systems compete in real time, while AI-to-human interaction plays a key role as mod bots engage with users by issuing warnings, filtering content, and shaping the chat experience. Beyond Twitch, the ability of spambots to manipulate metrics, influence conversations, and evade detection extends to broader online spaces, including social media and comment sections. Studying this dynamic provides insights into bot moderation effectiveness, human-chat interactions, and ethical content governance, helping to improve detection systems, reduce false positives, and create a safer, more engaging user experience.

Section 2: Research Sources & Summary

Summary 1: Exploring Agent-Based Chatbots

(Source: Exploring Agent-Based Chatbots: A Systematic Literature Review)

The paper Exploring Agent-Based Chatbots: A Systematic Literature Review examines how multi-agent systems (MAS) enhance chatbot capabilities beyond traditional rule-based models. (Calvaresi et al., 2023). Unlike conventional chatbots, which follow predefined rules, agent-based chatbots can learn, adapt, and collaborate using AI and MAS, making them more effective across domains like education, healthcare, finance, and e-commerce. These chatbots can handle complex conversations, personalize interactions, and make real-time decisions (Calvaresi et al., 2023). Key stakeholders include businesses, researchers, and consumers. However, challenges such as scalability, cross-platform knowledge sharing, and personalization remain (Calvaresi et al., 2023). Future research should focus on developing adaptive, efficient, and seamlessly integrated chatbots that evolve based on user behaviour.

Summary 2: Moderation Tools Used by Twitch Moderators

(Source: Categorizing Live Streaming Moderation Tools: An Analysis of Twitch)

Twitch moderation presents unique challenges, as moderators must manage live chat interactions in real-time (Wohn et al., 2023). Unlike traditional social media, Twitch moderation combines automated tools and human intervention to regulate chatrooms. The study examines Twitch AutoMod, Nightbot, Moobot, and BTTV, categorizing their functions into chat control, viewer control, content control, and settings control (Wohn et al., 2023). Moderators rely on both AI-driven and manual moderation techniques to combat spam, harassment, and policy violations. The paper discusses current tool limitations and suggests improvements such as better filtering, AI advancements, and adaptive moderation techniques (Wohn et al., 2023).

Section 3: The Core Components of the Simulation

§3.1 Entities: In live streaming platforms like Twitch, the key entities are streamers, human viewers, spam bots, malicious human spammers and mod bots.

- **Streamers:** Content creators who host live streams, interact with viewers, and rely on moderation tools to manage their chat. They may set chat rules, assign human moderators, or enable automated mods to handle disruptions. Channel owners do not directly control bot behaviour. Some streamers may also take an active role in moderation.
- **Human Viewers:** Engage in conversation, react to content, and support streamers through interactions like chat messages, follows, donations and subscriptions.
- **Spam Bots:** Automated accounts that flood chat rooms with ads, links, and disruptive messages reduce stream quality and overwhelm moderation tools.

- **Malicious Human Spammers:** Real users who post ads or misleading content but can evade detection better than bots by mimicking and adjusting human behaviour.
- **Mod Bots:** Automated systems that detect, issue warnings and block spam bots but struggle against adaptive spam strategies.

Mesa Analogy: The Virus on a Network

Spam bots behave like a computer virus, spreading rapidly through chatrooms unless stopped by mod bots, which function like antivirus software. Just as antivirus programs receive security patches to recognize and neutralize new threats, mod bots update their filtering techniques to detect and block evolving spam patterns. Meanwhile, human spammers act like virus mutations, constantly adapting their tactics to evade detection, such as modifying message formats, reducing frequency, or mimicking human behaviour to bypass automated filters. Additionally, streamers act as hubs in the network, influencing how moderation is applied. On Twitch, viewers, moderators, and automated tools are all connected to the streamer, who determines the moderation style. Some streamers enforce strict rules, while others allow more relaxed chat behaviour. Because moderation rules vary by the streamer, they select human moderators, enable mod bots, and set custom chat policies to shape the chat environment. Therefore, just as viruses continuously evolve, requiring constant antivirus updates, the ongoing battle between spam bots and mod bots perfectly aligns with the Virus on a Network analogy.

§3.6 Affordances: Live chat affordances include message posting, tagging, liking/reacting using emotes, reporting, and moderator actions such as muting banning and filtering.

- **Spam Bots:** Takes advantage of affordances like high-frequency posting and lack of verification to flood chat with unwanted content.
- **Mod Bots:** Use affordances such as auto-filtering, sentiment analysis, and flagging to identify and remove harmful messages.

Mesa Analogy: Epstein Civil Violence Model

This model simulates the relationship between law enforcement and rebels in a civil unrest scenario. Spammers act as rebels, attempting to spread content (spam messages), disrupt order, and evade suppression. Meanwhile, mod bots function like law enforcement, working to control, suppress, and remove spam to maintain stability. Just as rebels in a civil unrest situation change tactics to resist authority, spammers continuously adapt by altering message formats or bypassing detection. In response, mod bots must also evolve, improving their filtering techniques to keep up with new spamming strategies. This constant struggle between spammers and mod bots mirrors the dynamic battle between rebels and law enforcement in the Epstein Civil Violence Model.

§3.3 Algorithms: Twitch uses automated moderation algorithms to detect and manage spam live chat. These algorithms help mod bots filter messages, prevent abuse, and maintain a good chat experience.

- **Spam Detection Algorithms:** This algorithm scans chat messages and looks for patterns of repeated content, links, or spam behaviour. If a message looks suspicious, it is flagged or removed automatically.

- **Reputation-Based Filtering:** This system prioritizes messages from trusted users (e.g., long-time followers, subscribers, verified accounts). New accounts or suspicious users are flagged, meaning their messages might be hidden or reviewed before appearing in chat. The reason for this is that new accounts are more likely to be spam bots, while older accounts have a better reputation.
- **Adaptive Learning (AI-Based Detection):** Mod bots evolve by learning from past spam patterns. If spammers change tactics (e.g., using slightly different words or posting less frequently to avoid detection), adaptive learning helps mod bots detect new variations.

Mesa Analogy: Boid Flocking Model

Boid Flocking Model This model simulates how birds (boids) adjust their flight paths by avoiding obstacles and regrouping dynamically in response to their surroundings. It represents adaptive behaviour, similar to how spam bots and mod bots continuously evolve their strategies in response to each other. Spam bots adjust their behaviour to evade detection, mimicking human interactions or changing message patterns. Mod bots update their filtering and detection techniques, adapting to new spam tactics just as boids adjust their flight patterns. This constant back-and-forth adjustment between spam bots and mod bots mirrors the way boids navigate and adapt as a group, making the Boid Flocking Model a strong analogy for this interaction.

Section 4: Simulation Anticipated Outcomes

We use the Boid model to simulate the process of Mod bot fighting spam bot. The total number of boids represents the total number of accounts in the live broadcast room, green represents managed accounts, and red represents unmanaged accounts. Through the changes in the number of accounts of different colours, we can observe the process of confrontation.

Speed can be regarded as the response speed of a mod bot. Vision of Bird can be regarded as the management ability of the mod bot. Minimum Separation can be regarded as the cost of management. Because Minimum Separation has a certain confrontational relationship with Vision range, the larger the separation distance, the smaller the possibility of being included in Vision range, and vice versa.

References

Calvaresi, D., Dubovitskaya, A., Taveter, K., Schumacher, M., & Främling, K. (2023). Exploring agent-based chatbots: A systematic literature review. Journal of Ambient Intelligence and Humanized Computing, 14(8), 11207–11226. https://doi.org/10.1007/s12652-023-04626-5

Cai, J., & Wohn, D. Y. (2019). Categorizing live streaming moderation tools: An analysis of Twitch. International Journal of Interactive Communication Systems and Technologies, 9(2), 36–50. https://doi.org/10.4018/IJICST.2019070103

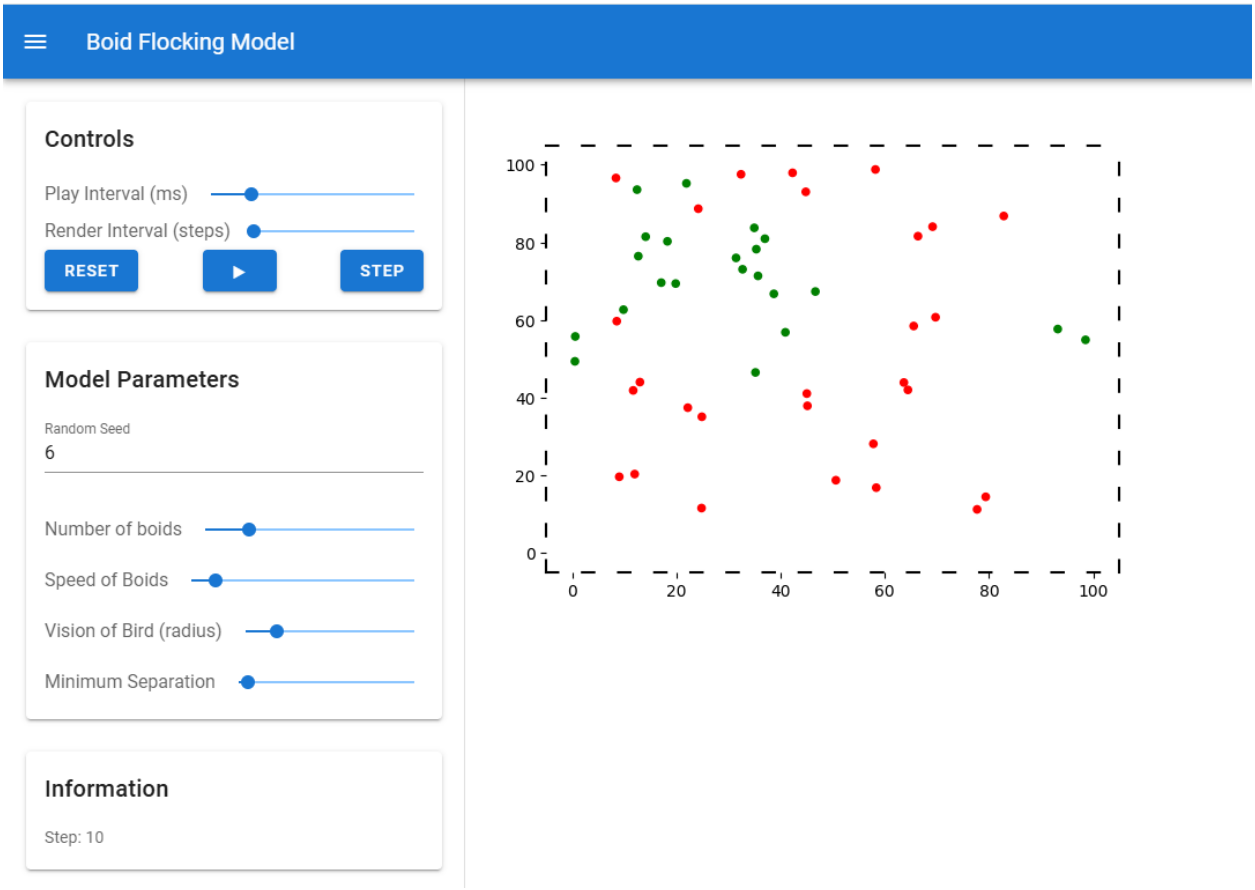

Figure 1: Twitch Simulation