CS492 Winter 2024
Final Project Handin - Description
April 8, 2024
Alex Hejmej [20779600], Tailai Wang [20816387], Isaiah Witzke [20845973]
Link to GitHub Repository, Link to Live Website (Final Deliverable), Link to Video Tour (YouTube)

Interactive Storytelling Website about AI Bot Toxicity on Social Media Platforms

What questions were being addressed?

The main inspiration of our research project was to investigate how bots affect social media platforms after seeing so many on Twitter. We wanted to answer the following initial questions in our project:

1. Where do bots come from / who is making these bots?
2. What are bots' common message(s)/intention(s)? Are bots effective at drawing attention to their message(s)?
3. Are bots able to manipulate a social media user into engaging with their content?
4. What are social media platforms doing to moderate & shut down bots? Are they successful?

During Milestone 2, we were able to answer these four questions well enough through our secondary research, but new questions that arose in Milestone 2 also presented us with more reason to do our own research on bot detection tools. We concluded that the bot problem was most salient for Reddit, as their recent IPO put a spotlight on how their moderation is mostly done by unpaid volunteers. Thus, during this last stretch of the project, we focused on doing performance analysis of tools that could potentially assist moderators with fighting bots, and presenting our findings on our website.

What Form did the Project Take?

Our project since Milestone 2 is meant to be a reproducible data analysis. On our GitHub, we detail how we procured the live Reddit data, documented how we tested the open-source tools we discovered, and packaged everything into a single repository that a reader could clone and play with. We decided to present the results of this research project as a website so that viewers can understand our findings in a way that's slightly more structured and interactive than a traditional web article. The website is meant to be read like a story - it covers the background of the bot problem, what the industry landscape looks like, and then finally dives into our analysis of Reddit bot tools.

Summary of Findings

In our data analysis, we gathered over 1000 test humans and bots. We then tested 3 very different open-source tools on this test set, looking to understand the strengths and weaknesses of each tool:

1. Heuristic Model ("Digital Eye Test"): The model uses public metadata from the test users and attempts to classify bots based on attributes like verification, account age, and karma. We found that this model is rudimentary at best, but serves as a good base to see what tooling can do.
2. Random Forest Model ("Cosine Similarity Test"): The model takes recent posts & comments from the test users and does TF-IDF & Cosine Similarity against training data to assess whether the user's behavior is more similar to that of a bot's, or a human's. We found that this model does well with traditional bots, but gets easily fooled by LLM-generated content.
3. OpenAI GPT-2 Detector: OpenAI released this public tool to detect whether a piece of text was generated by its GPT-2 LLM. We found that this tool was quite successful at identifying bots powered by GPT-2, but is unsurprisingly incapable of doing anything else.

Our final suggestion was to use the Random Forest and the GPT-2 detector in tandem, effectively creating a general-purpose bot detector. All of this is described in detail on our final deliverable website!