Francisco Arenas
Shalim Montes Hernandez

Project Proposal

Problem:

Social media allows for digital communication where users can create online communities to share information, ideas, personal updates, and other content. However, despite being created for human use, social media platforms everywhere are being plagued by higher rates of non-human users. Social media bots are automated software that simulate human behavior to perform a range of tasks like boosting stats and engagements on posts via likes, views, and comments. While these relatively benign goals may appear harmless, automated bots are far more concerning when used for malicious purposes. Recent bots have been tasked with goals such as pushing extremist propaganda, impersonating real people, or attempting to steal personal information from human users by appealing to carnal desires. This combination of malicious-oriented automated programs and increased deployment of these bots calls for a critical analysis of public online activity on social media platforms. For this reason, we propose a model that aims to predict whether an account or online activity originates from a human user or an automated bot. We hope our study and model may provide an accurate detection of bots, thereby reducing exposure to malicious activity and misinformation.

Context:

As we lay out the structure of our ideas and approach to creating an effective model, our project proposal will revolve around these three research questions:

- **RQ1: What are social media bots, and why are they dangerous?**
- **RQ2: What percentage of "users" on popular social media platforms are bots?**
- **RQ3: How can you distinguish between a real user and a bot?**

The core analysis and approach to answering these questions will result from a literature review of current methods and definitions, and other reputable news sources, for a public understanding.

Ahmad Wani, Mudasir, and Suraiya Jabin. "A Sneak into the Devil's Colony-Fake Profiles in Online Social Networks." ***arXiv*****, 2017, doi:arXiv:1705.09929.**

Summary: Presents the online social network (OSN) as a grouping of nodes connecting global participants, which are vulnerable to cyber criminals, especially through the use of bots. Introduces categories of features used to train classifiers in order to detect fake profiles. "Mostly the aim of these cyber criminals is to steal the user's personal, professional, political, social or financial information by exposing the users with unwanted information on the web like pornography, etc. in order to deceive them (pg 1-2)."

Thoughts: Provides an extensive list of definitions and profile/bot types cyber criminals use on online social networks (social media). Helps ground us further on the tactics used in creating bots. The features collected are noteworthy in helping us ensure our dataset is capturing the right features (e.g., pages liked by the user, rate of like activity, number of replies, etc.).

Ballard, Shawn. "Are Bots Winning the War to Control Social Media?" *WashU Arts & Sciences*, 1 Nov. 2022, https://artsci.washu.edu/ampersand/are-bots-winning-war-control-social-media.

Summary: Records an interview with developers of an algorithm to distinguish human from non-human users on Twitter with up to a 98% accuracy. They found that bots may account for 25-68% of Twitter users' engagement during certain times and topics discussed.

Thoughts: Reports findings on previous approaches to identify bots on social media, and really points the spotlight on the current infiltration of bots in popular websites. Also introduces the never-ending battle between identification algorithms and bot generation.

Ng, Lynnette Hui Xian, and Kathleen M. Carley. "A Global Comparison of Social Media Bot and Human Characteristics." *Scientific Reports*, vol. 15, no. 1, Mar. 2025, doi:10.1038/s41598-025-96372-1.

Summary: Delves deep into the linguistic differences between bots and humans (e.g., increased hashtags and positive terms) and compares these characteristics across global events.

Thoughts: Offers a short but helpful overview of bot activity and influence on social media across scholarship with specific examples. Directly answers the question of what a social media bot is and provides a definition of sorts. Also provides concrete numbers of how many bots are prevalent in social media platforms.

Orabi, Mariam, et al. "Detection of Bots in Social Media: A Systematic Review." *Information Processing &amp; Management*, vol. 57, no. 4, Jul. 2020, p. 102250, doi:10.1016/j.ipm.2020.102250.

Summary: Provides a meta-analysis of social media bot detection methods and identifies gaps in the research area.

Thoughts: Helpful in grounding our approach to align with current methods. We can potentially fill in the gaps where needed; however, the gaps introduced seem out of reach for the time-frame of the project (e.g., real-time detection).

Shane, Scott. "The Fake Americans Russia Created to Influence the Election." ***The New York Times*****, 7 Sep. 2017, https://www.nytimes.com/2017/09/07/us/politics/russia-facebook-twitter-election.html.**

Summary: Reports a historic propaganda attack during the 2016 Presidential election through the use of social media bots. "On Election Day, for instance, they found that one group of Twitter bots sent out the hashtag #WarAgainstDemocrats more than 1,700 times."

Thoughts: A concrete example of a successful malicious attack using bots on popular social media platforms to influence real users or projections. Highlights the danger posed by social media bots and the need for research in the area.

Data:

The dataset we plan to mainly use is: "**twitter-human-bots**"

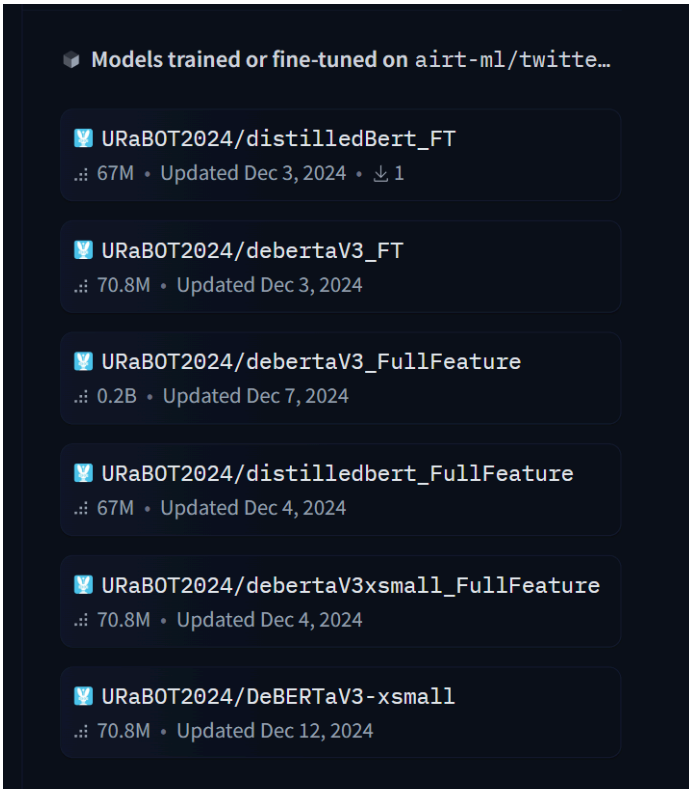
- **Metadata info**:
  - Last Updated: 2023
  - Author: AIRT-ML
  - Website: Hugging Face
  - Dimension: 37,438 x 20
  - Labelled dataset
- **Access**:
  - Hugging Face datasets have an API we can call to to easily download datasets using the following lines:

```
from datasets import load_dataset

ds = load_dataset("airt-ml/twitter-human-bots")
```

- **Processing**:
  - There may be some features that we have to get rid of. There is a feature that shows the account creation date but the latest date is the year 2017 which isn't entirely useful in the year 2026
  - Select tweets only done in English

  - Clean up missing values in examples
  - Go over unique feature values to see what we should keep or delete
- **Plan**: Why did we choose this dataset? What features does it have?
  - Features:
    - Account creation date, default profile, default profile picture, description of tweet, how many posts are saved to favorites, how many followers, how many friends, whether geo is enabled, unique account id, language, location, background image url, profile picture url, user name, statuses count, whether the account is verified, average tweets per day, account age in days, and whether the account is a human or a bot
  - We chose this dataset because it's public, and it contains a label column showing whether an account is a human or a bot. We can proceed with supervised learning methods. The features align with what we looked at in our literature review, and other projects have trained models using this dataset.



Models trained or fine-tuned on airt-ml/twitte...

URaBOT2024/distilledBert_FT
67M • Updated Dec 3, 2024 • ↓ 1

URaBOT2024/debertaV3_FT
70.8M • Updated Dec 3, 2024

URaBOT2024/debertaV3_FullFeature
0.2B • Updated Dec 7, 2024

URaBOT2024/distilledbert_FullFeature
67M • Updated Dec 4, 2024

URaBOT2024/debertaV3xsmall_FullFeature
70.8M • Updated Dec 4, 2024

URaBOT2024/DeBERTaV3-xsmall
70.8M • Updated Dec 12, 2024

Additional Datasets we may look into if the original one doesn't work:

- Tweets and Social Network Data for Twitter Bot Analysis
  - Article listing 8 datasets they found:
    - – **botometer-feedback-2019**: (529): Botometer feedback accounts manually labeled by K.C. Yang [8]
    - – **botwiki-2019**: (704) Self-identified bots from https://botwiki.org [8]
    - – **gilani-2017**: (2,652) Manually annotated human and bot accounts [3]

- – **midterm-2018**: (50,538) Manually labeled human and bot accounts from 2018 US midterm elections [8]
- – **pronbots-2019**: (21,964) Pronbots shared by Andy Patel (github.com/r0zetta/pronbot2)[8]
- – **varol-2017**: (2,572) This dataset contains annotations of 2573 Twitter accounts. Annotation and data crawl is completed in April 2016 [6]
- – **verified-2019**: (2,000) Verified human accounts. [8]
- – **cresci-stock-2018**: (25,987) A dataset of (i) genuine, (ii) traditional, and(iii) social spambot Twitter accounts, annotated by CrowdFlower contributors [1]

Approach:

- **Problem Type**: Supervised Learning
- **ML Algorithms**: k-NN
    - Other potential possibilities: Logistic Regression w/ Gradient Descent, Naive Bayes, Neural Networks
- **Guides**: We aim to take advantage of the collected features highlighted in the articles, and modify our dataset accordingly to provide as best a model as possible. Thankfully, at a quick glance, our selected dataset seems to contain the right features to train on.
- **Difference:** We haven't attempted to classify an entry based on text yet (NLP).
- **Why:** k-NN is useful for text classification problems. Although the tweet text isn't the only feature we are working with, it will play a large role in predicting whether an account is a human or a bot. Text, in itself, is highly dimensional and nonlinear which means other approaches will try to establish a linear decision boundary which will not accurately capture how text actually behaves.

Computational Needs:

We plan to run our experiments on our own local machines.

Questions/concerns:

- How would we go about learning about NLP, are there modules/resources you recommend?
- Do you think k-NN is an appropriate approach for this problem, or is there an approach you would recommend given the timeframe of this project?