**Progress_Report_1**

**Project Description:**

- Research Question:
    - How do social bots engage in public discussions about topics in the Los Angeles wildfire?
- Social Science Relevance:
    - Active participation of online social bots in political debates and discussions has been extensively researched and confirmed by previous studies. For instance, Hagen et al. (2020) examined the activity patterns of social bots in discussions surrounding the 2020 U.S. presidential election on Twitter through network analysis. However, despite the high politicization and widespread generation and dissemination of online misinformation about environmental issues, few studies have explored the influence of bots. Building on this body of work, we aim to explore the prevalence of social bots in public discussions about topics beyond U.S. politics. We hypothesize that the prevalence of bots in discussions about elections is largely due to the involvement of various stakeholders. Similarly, while the L.A. wildfires are not explicitly political, they naturally evoke discussions about state policies and governmental responses, making them a relevant event to study social bot behavior.

**Research Questions & Hypotheses:**

- Q1: How important are bots in wildfire discussions?
- Q2.1: What sub-topics do social bots prefer in wildfire discussion?
- Q2.2: How do social bots and humans differ in sub-topic preference?
- Q3: How does the popularity of bots differ across platforms?

- H1: On reddit, we expect to observe a small proportion of bots have high importance while the majority of bots have little importance.
- H2: On reddit, social bots prefer political related sub-topics in the wildfire discussion
- H3: On the pro-Democrats platform(Bluesky), we expect to observe a lower rate of use of bots than on Reddit.

**Key Concepts and Operationalization**

- Social Bot:

  Social bots are accounts operated wholly or partially by software (Ferrara et al., 2016), typically created in bulk and programmed to continuously generate large volumes of content (Howard et al., 2016). They can mimic human behavior by automatically sending friend requests, sharing posts, and liking content (Boshmaf et al., 2011; Echeverria et al., 2017). By artificially amplifying support (Howard & Kollanyi, 2016) or spreading misinformation on social media (Shao et al., 2018), social bots effectively manipulate public discourse.

- Network Nodes and Ties:

  We represent user interactions as a weighted directed graph G, where each node corresponds to a user. Each directed edge (u,v) signifies that user u has commented on one or more posts created by user v, with the weight of the edge reflecting the total number of comments made by u on v's posts.

  For a given user u in G, the in-degree of u is defined as the number of distinct users who have commented on uuu's posts, i.e., the number of unique users interacting with u. Conversely, the out-degree of u represents the extent of u''s interactions within the community, calculated as the number of posts on which u has commented.

  In addition, we also assess user importance by examining the structure of discussion trees (separate posts within each subreddit). In a discussion tree, the root node represents a post, while the other nodes represent comments that are either direct responses to the post or replies to other comments within the same thread.

  By analyzing the size (depth) and shape (width) of these trees, we can gain insights into which topics attract the most user attention, as well as identify key members within the community. For example, a particularly emotional or controversial comment may spark more reactions than the original post, causing the majority of the discussion to be concentrated within a specific branch of the tree.

- User importance:
  We use the degree centrality in each network to assess the importance of users. Higher degree centrality indicates higher importance users have in facilitating the discussion.

  In addition, we also analyze the size (depth) and shape (width) of decision trees to gain insights into which topics attract the most user attention, as well as identify key members within the community.

**Data Sources:**

- Reddit:
  - Collection Method: Web Scraping with PRAW
  - Justification: Reddit is one of the largest platforms where people engage in idea exchanges about a wide variety of topics. It is easy to keep track of discussions by focusing on several selected Reddit communities directly related to the event of interest. Reddit data can be scraped using PRAW, the Python Reddit API Wrapper.

    - Time Frame: January 5, 2025 onward
    - Data Size: 248 Posts and 18,788 Comments
    - Validity: The validity of the data may be challenged as the reddit API has a strict limit of requests allowed. The close to maximum number of posts and comments we can collect using PRAW is 250 posts and their associated comments. As the process involved expanding CommentForest which sends additional requests, we are uncertain of the true maximum capacity. Attempts to request more posts may lead to account suspension. The data collected is not a complete picture of the online discussion. This problem may be resolved using a collection method which collects "top" posts containing the "LA wildfire" keyword within a set time frame of 20 days from January 5, 2025 to January 25, 2025 on old.reddit.com. In this case, the post contents and top-level comments(max 200) can be scraped.
    - Alternative solution1: We are currently attempting to build a scraper that does not use PRAW to resolve the limited data issue. Ideally, this new scraper should be able to access 25 posts on 40 pages(max 1,000 posts) and 200 top comments associated with each post(max 200,000 comments).
    - Alternative solution1. We first identify several most frequent subreddits in the initial posts and then using keyword like"big fire" "wildfire" search under each subreddit. After reaching limit for one subreddit, we use another account to scrape under another subreddit. In this way, we can increase data size to 250*n (the total number of subreddits we pick)
- Bluesky:
    - Collection Method: Bluesky API (https://docs.bsky.app/)
    - Justification: The Bluesky API is free and publicly available without the need for applications and approval. The platform itself is emerging.
    - Time Frame: January 17, 2025 to January 31, 2025 (working on scraping older data)
    - Data Size: 4896 threads with their comments
    - Validity: Data scraping currently encounters a 'literal_error'. I will continue working on it to scrape older data. I have not yet identified an API or method to detect bot users/bot-generated threads within the dataset. This issue can be resolved by conducting further research on relevant papers or GitHub projects. Alternatively, we can modify a bot detection project designed for a similar platform, such as X, or BotBoster-Universe to meet our needs.
    - Additional Info: Ideally, I will Bluesky scrape all threads dated after January 7, 2025, the date the Southern California wildfires began.

**Data Cleaning and Wrangling:**

- Reddit:
    - Bot/human labeling
      The bot probability is used to identify whether a user is a bot or not. The bot probabilities of all comments and posts associated with that user is averaged to

create a bot probability for the specific user. A threshold of 70% will be used initially to qualify users as bots. The percentage of bots will be stored as a node attribute for subreddits. We plan to alter the threshold to create multiple visualizations.

- Sub-topic detection
  We plan to detect sub-topics by combining existing labels (subreddit titles) with the unsupervised LDA model. We can utilize the existing subreddit titles, as some of them have clear thematic tendencies. For example, r/california_politics is related to politics, and r/environment is related to environmental issues. Furthermore, posts within the same subreddit may have overlapping themes, and different themes may also appear within a single subreddit. Based on this, we will apply the LDA topic model to perform unsupervised topic analysis on the content of the posts. We will then combine these two methods and manually check to determine our sub-topic list and categorize the posts.
- Network matrix constructing
  We plan to construct user*user matrix for further network analysis.
  Also, We will identify users using the author_fullname attribute which is a unique code associated with each user. If the user is found to be active(posting/commenting) in two subreddits, a tie is formed between the two subreddits. The weight of the tie will be reflecting the number of users active in both subreddit. We also plan to conduct community detection on our network.

**Data Analysis and Visualization:**

- Dependent variables: account type (bot/human)
- Independent variables: sub-topic type; degree centrality
- Data Analysis & Visualization Method:
  - Reddit: We plan to visualize network structures of subreddits given different bot qualification thresholds. This should reveal engagement of social bots in online discussions of Los Angeles wildfire. This should provide insights on whether social bots cluster on specific subreddits and hence topics. We hypothesize that there will be a cluster of subreddits related to politics with a high percentage of bot activity.

**Responsibilities:**

Hugo He: Presentation slides, video, Bluesky data scraping, data processing

Moe Wu: Bot detection, Data cleaning, wrangling, analysis, and visualization

Yilin Xu: Reddit Scraping, README, data cleaning, wrangling, analysis, and visualization

**Additional Information:**

**BotBuster:** [https://github.com/quarbby/BotBuster-Universe](https://github.com/quarbby/BotBuster-Universe)