

CSE 472 Project 1
Data Crawler
Project Report
US Elections- Red vs Blue

Overview:

I have completed a project on the US elections. I have analyzed posts related to this topic as well as individuals connected with this debate.

Data Collections:

I have used Mastodon as the primary social media site. I created a developer account to fetch API keys that I used to crawl the data.

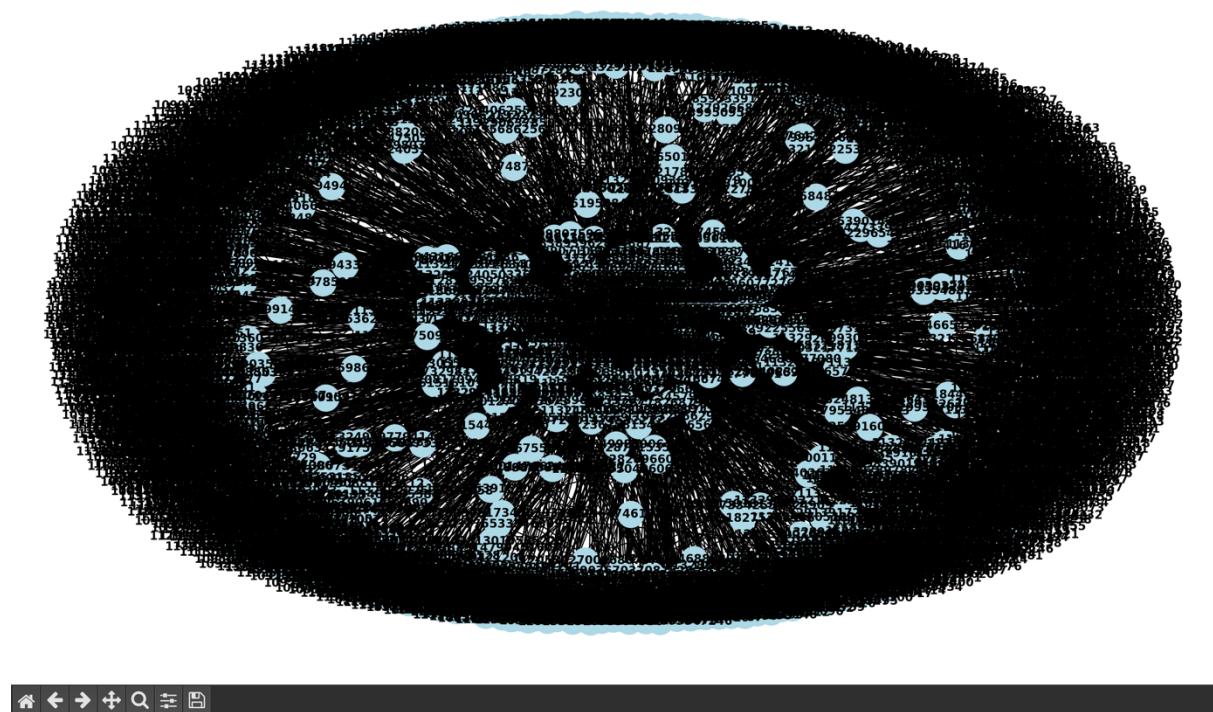
Posts: I have gathered many hashtags relating to the US elections. Almost half of them relate to the democrats and other half to the republicans. Then my crawler uses the .search_v2() function to fetch all the posts on Mastodon that used any of these hashtags. Then these posts are saved in all_statuses list. This is then written into the posts.json file. This contains dictionaries with post details along with any comments or boosts on them.

Users: I have gathered 3 users from Mastodon, Trump and Kamala Harris. Kamala Harris's account did not have much followers so I took two of her accounts that equaled the following of Trump's handle. Now I ran a loop to get top 80 followers (max) from all of them. Then I recursed this to get followers of one of those followers to get a data set of 350 users.

Graph Creation

I have used network libraries for creating two graphs.

Post Graph: this Directional graph contains posts, comments and boosts as nodes. These are then connected. One post is connected to its replies and boosts. The nodes are labeled with their post id as an identifier.



User Graph: This graph contains three seed users. These users are then used to fetch more users and they are connected to each other. Then top users are again used to fetch more users and are connected. The nodes are labeled as their id as identifier.



I made a map dictionary to store key as the id and values as its followers. Then I added all the elements as a node to the graph and recurse through the dict to log all the connections.

Classification of Posts using Llama8

I used Llamaba 3.1 8b. I made a dictionary from my posts.json file and fed it to the LLM. The instruction feeded is :

his is a post dedicated to a political leader. Read the post and classify it as toxic or non toxic.
The post may be considered toxic if there

is targeted hate speech in any form. Give only one word answer. Do not give explanation.

Then I parse the output to check if it returns toxic or non-toxic. Depending on that, I put it in toxic.txt or non_toxic.txt

Link <https://colab.research.google.com/drive/1v9h-WH94hyO2a6ul3kTVI5SQzdKyXQiU?usp=sharing>

The screenshot shows a Jupyter Notebook cell with the following Python code:

```
toxic={}
non_toxic={}
for i in tqdm(posts_dict):
    comment=posts_dict[i]
    inputs = tokenizer(
        [
            alpaca_prompt.format(
                '''This is a post dedicated to a political leader. Read the post and classify it as toxic or non toxic. The post may be considered toxic if there
                is targeted hate speech in any form. Give only one word answer. Do not give explanation.'''), # instruction
                comment, # input
                "", # output - leave this blank for generation!
            )
        ], return_tensors = "pt").to("cuda")
    response = model.generate(**inputs, max_new_tokens = 128)
    res=(tokenizer.batch_decode(response)[0])
    res=res.split()
    desc=''
    for j in res:
        if j in "Response:":
            desc = res[res.index(j)+1]
        if desc.lower() in "toxic":
            toxic[i]= posts_dict[i]
        else:
            non_toxic[i]= posts_dict[i]

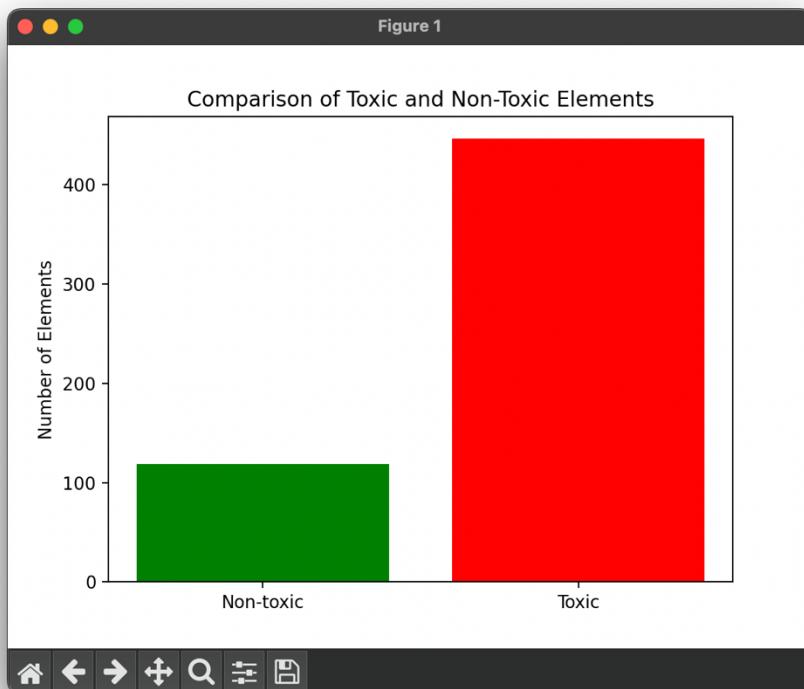
    with open('toxic.txt','w') as f:
        f.write(str(toxic))
    with open('non_toxic.txt','w') as f:
        f.write(str(non_toxic))
```

toxic.txt

```
{113292127835973709: '# Trump # MAGA # WeatherControl # weather USA # ignorance # lies # politics # political # humour # humor # cartoon # campaign # FEMA # aid # refused', 113292117229605279: '# Trump # MAGA # WeatherControl # weather USA # ignorance # lies # politics # political # humour # humor # cartoon # campaign # JoeBiden # JoePoseiden', 113292028964953767: "No wonder # Trump keeps droning on about # Hannibal Lecter, turns out he's just appealing to red-staters https://www.whas11.com/article/news/crime/mount-olivet-woman-charged-mothers-death-dismembered-body-cooked-torilena-may-fields/417-86c14547-b601-4e13-8440-0c53b4f3c721", 113291964175512188: '# US House Speaker Mike # Johnson said on Oct. 11 that he no longer has "an appetite for further # Ukraine funding" and hopes that a November electoral victory for former U.S. President Donald # Trump will bring a swift end to the war. . "I don't have an appetite for further Ukraine funding, and I hope it's not necessary," Johnson said on Oct. 11 in an interview with Punchbowl News. https://kyivindependent.com/house-speaker-has-no-appetite-for-more-us-support-for-ukraine/ # Russia # MAGA # GOP', 11329187679445747: "This piece of shit is delusional if he really believes # Colorado is going to vote for him. My state is definitely Blue, with stupid pockets of red in the sparsely populated rural areas. There's no chance # Trump carries the state. He lost here last time, and he'll lose here again. # TrumpIsUnfitForOffice # TrumpIsATraitor # wannabeDictator # TrumpIsAFascist # EndOfDemocracy # PutinOwnsTrump # TrumpIsARussianAsset # Treason # LoganAct # TrumpCannotBePresident # VoteBlue https://okmagazine.com/p/donald-trump-slammed-november-5-liberation-day-occupied-america/", 113291871137677981: 'OMG 😂 "Trump Farts on Stage, MAGA Nuts Push Insane Weather Lies, Mr. USA\\'s Products are Made in China!" - Jimmy Kimmel Live https:// youtube.com/watch?v=3gXLdl-aj-w&si=o_2EmHnBLuTvEcjl # trump # HarrisWalz2024 # HarrisForPresident # KamalaForPresident # KamalaHarris # jimmykimmellive # comedy # humor', 113291840668246758: 'THE VOTE a Palestinian perspective # Genociders # Vote # Palestine # BreakingTheDuopoly # USA # ThirdParty # AntiGenocide # SaveLives # CEASEFIRENOW # NoRed # NoBlue # Heartbreak # Anger # Fear # Compassion # Justice # Humanitarianism # Harris # Trump', 113291829013405415: "OTD in 2016, a former Miss Universe contestant verified # GOP President Donald # Trump 's claims to Howard Stern that he would enter the changing rooms to see the contestants naked https:// toilet-guru.com/trump/?s=mb # peeper", 113286292781093421: '@
```

non_toxic.txt

```
{113292121121916759: '# Trump # MAGA # WeatherControl # weather USA # ignorance # lies # politics # political # humour # humor # cartoon # campaign # JoeBiden', 113292111203133254: '# Trump # MAGA # WeatherControl # weather USA # ignorance # lies # politics # political # humour # humor # cartoon # campaign', 113292099305194475: '# MAGA # Trump # ignorance # WeatherControl # weather # politics # political # humour # humor # cartoon', 113292054686677080: '"Fascist to the core:" Former Trump official Mark Milley warns against dangerous 2nd term https://www.salon.com/2024/10/11/fascist-to-the-core-former-official-milley-warns-against-dangerous-second-term/ * Mark Milley: Trump the "most dangerous person ever" * snippet f. Bob Woodward's book "War" https://www.simonandschuster.com/books/War /Bob-Woodward/9781668052273 https:// en.wikipedia.org/wiki/Bob_Woodward Ex-Trump Official [retired Gen. Mark Milley] Scared He Will Be Out for Revenge If He Wins https:// mastodon.social/@persagen/113290632743048494 # MarkMilley # Trump2 # GOP # HeritageFoundation # Project2025 # ProtoFascism # authoritarianism # extremism # ElonMusk # AmericaPAC', 11329181087372730: "This might be an interesting time to re-read this terrific article by Oliver Sacks: https:// archive.md/ocsrD I'll put an abbreviated version in the images to give you the gist. # Trump # election2024 # KamalaHarris # psychology # Neurology", 113289224050336724: '"Those who engage in such # lies are undermining confidence in the rescue & recovery work that's ongoing," # Biden said. "These lies are also harmful to those who most need help. Lives are on the line, people are in desperate situations – have the decency to tell them the truth." And asked again as he left the room whether he planned to speak w/ # Trump , # Biden responded simply: "No." # USpol # Climate # ExtremeWeather # Hurricane # Milton # disinformation # TrumpLies # ClimateCrisis', 113289185295520461: 'Asked by reporters after his remarks at the White House whether he had spoken w/ # Trump , # Biden responded indignantly: "Are you kidding me?" He then looked at the camera w/a message directly for his predecessor: "Mr. President Trump – former President Trump – get a life, man. Help these people." # USpol # Climate # ExtremeWeather # Hurricane # Milton # disinformation # TrumpLies # ClimateCrisis', 113286019952705123: 'Look at this headline: "How # hurricane falsehoods are dividing the # Republican Party As the country digs out, # Republicans in storm-battered states appear torn between the need to curb dangerous # misinformation and fear of drawing a rebuke from # Trump just weeks before the # election ."
```



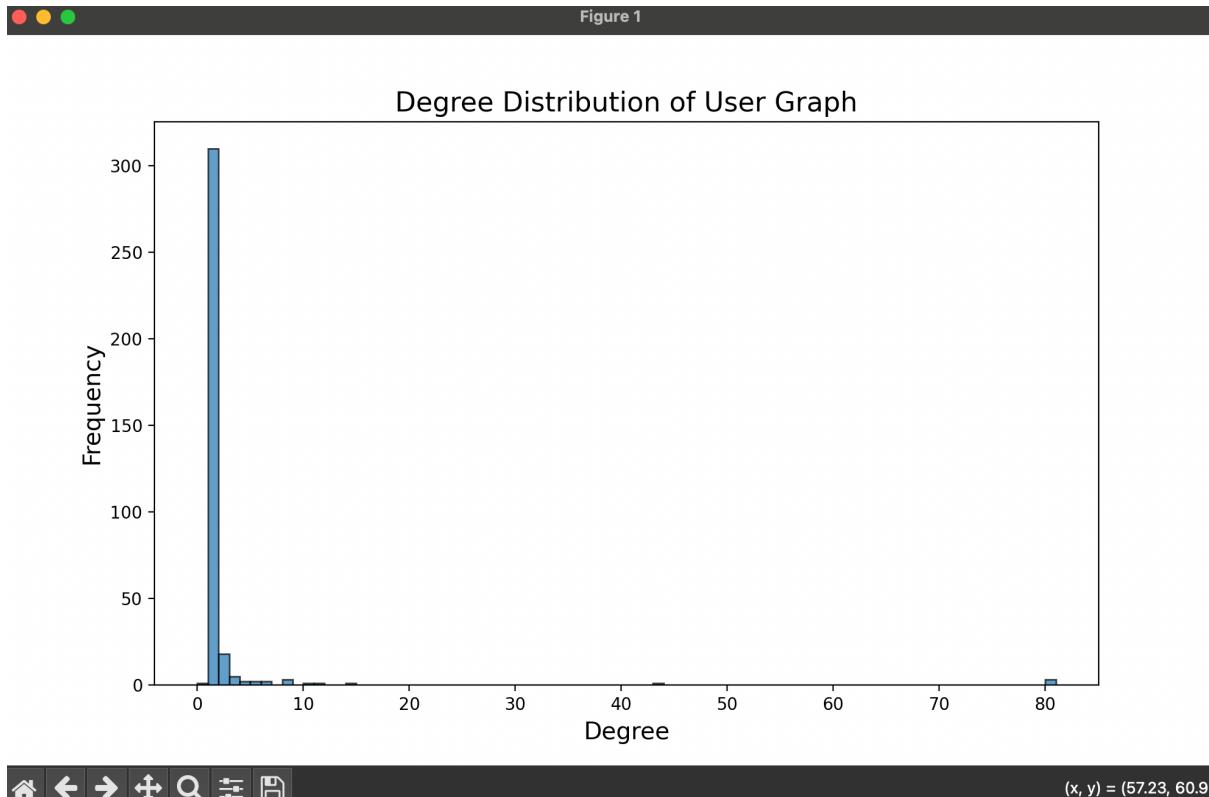
On my analysis, I found that toxic comments and posts are about 78% of the posts and non_toxic are about 22%.

This is displayed in the histogram above.

Network Measures

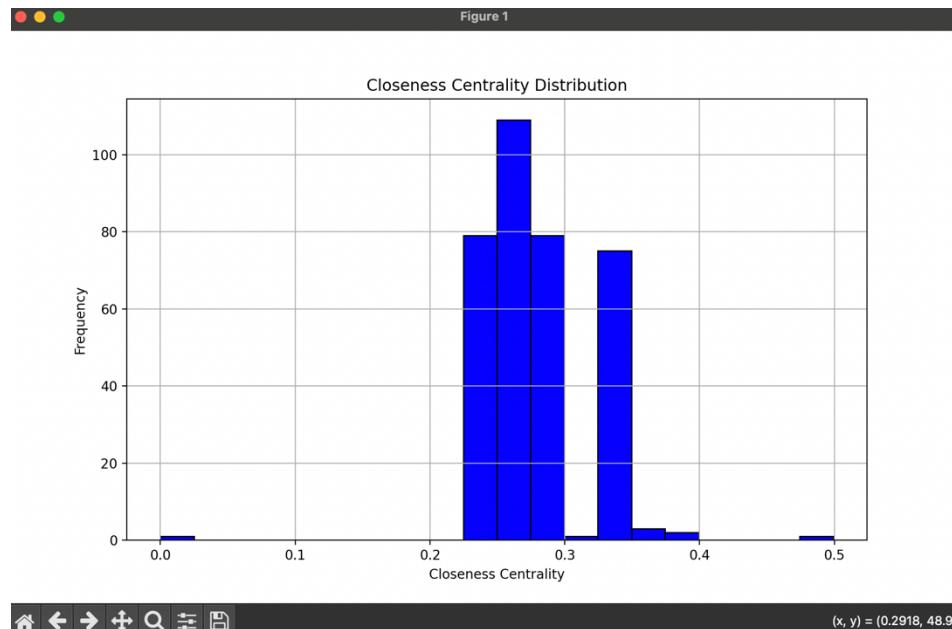
I used functions in networkx to get the degree distribution histogram, closeness centrality histogram and betweenness centrality histogram.

Degree distribution histogram

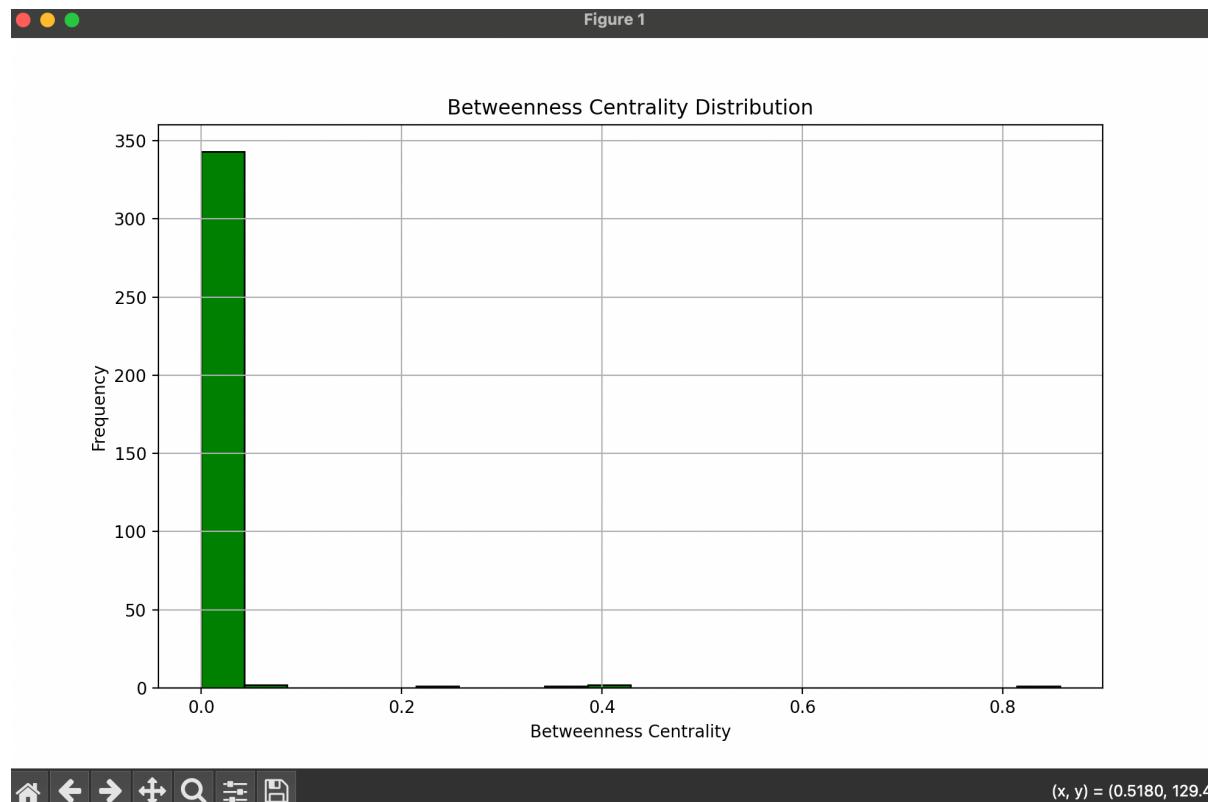


I fetched all the degrees for all the nodes of the graph. Then I plot the histogram.

Closeness centrality histogram



Betweenness centrality histogram



1 hop friend average : Too long! Look at it in the terminal while running.
Global Average Number of Friends (Across Entire Network): 2.10

Analysis:

When studying the social media followers of Trump and Harris for the 2024 elections interactive graphs display a pattern; most users have only a few connections while a small number have many connections that point to them being influential figures, with broad networks in the system. These influential individuals likely have a role in spreading content or opinions within their communities. The distribution of closeness centrality highlights that the majority of users are grouped around a value. Indicating they are neither too distant, from the network hubs nor entirely isolated. This suggests that the network allows for communication, between its sections to a certain extent. Though the distribution of centrality is unevenly skewed – most users exhibit betweenness values while only a few play crucial roles as connectors linking distinct parts of the network. These pivotal nodes wield influence in connecting separate communities and present opportunities, for strategic messaging or influence efforts.

References:

Mastodon documentation: <https://mastodonpy.readthedocs.io/en/stable/>

Networkx documentation: <https://networkx.org/>

Llama 3-8b: https://colab.research.google.com/drive/1InnyfcR4rl_qRTmwqy-isIXdutGPFciu?usp=sharing

OpenAI ChatGPT: <https://chatgpt.com/>