DSC 180B Report

David Aminifard, Samuel Huang
Justin Eldridge, Section A10

Project Github
Project Website

Introduction

Last quarter's project involved downloading twitter data based on pre downloaded tweet IDs related to COVID-19. We peeled back the layers of a misinformation network using k-core analysis in addition to performing exploratory data analysis. While this project yielded interesting findings, it seemed limited to COVID-19-related misinformation only, and finding similar pre-categorized twitter data such as this can be very difficult. As a result, we decided to continue our investigation of misinformation in social media on Reddit's platform.

We decided to base this project off of Reddit data because Reddit offers pre-categorized information by virtue of having many Subreddit communities, which are, generally, communities leaning towards a specific topic.

Similar projects such as ours have been done in the past. For example, a paper title, "The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources" , written by Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn, also looks into the spread of misinformation on the Reddit Platform. The paper compares the spread of mainstream and alternative news URLs on Reddit, Twitter, and 4chan, and it finds that 4chan and six selected Subreddits have a very high percentage of alternative news URLs compared to other communities [1].

Our study will also involve identifying the frequency of misinformation on selected Subreddits based off of the presence of misinformation URLs. Our data will primarily consist of Reddit submissions and their respective attributes such as whether there is misinformation, subreddit name, the author, upvotes, downvotes etc. This will allow us to gauge the presence of misinformation on Subreddits individually and analyze the relationships between Subreddits based on their shared users.

In order to identify URLs as misinformation or not, we will use a third party list of misinformation domains made publicly available by iffy.news [2].

Methods

System specifications used for our scripts:

- At least 2 gigabytes of RAM
- Stable Internet

First, data will be acquired using the Python Reddit API Wrapper, PRAW, which provides access to public data made available by Reddit and simply requires creating an account and an associated application. Getting data in this case involves two steps: Downloading the data and uploading the data to Google Sheets.

Choosing the subreddits

> We decided to choose subreddits in a way that provides us with a holistic view of the spread of misinformation. We identified 5 different types of subreddits to analyze: Top 5 most popular, Highest raw growth in the last year, Political, Covid related, and News. This way we have a diverse group of subreddits to see where a disproportionate amount of misinformation is compared to others.
>
> In addition, there is a dynamic topic callled "Expanded MisInfo Network Subreddits", where we can programmatically add new subreddits based off of analysis of users who are known to post misinformation. So if a subreddit is seen as popular among the misinfo-posters we analyze (based off of their comments), we can add that subreddit to our list for future analysis.

Downloading the data

> In order to download the data, we simply download the last 1000 submissions sorted by the upvote count per Subreddit (equivalent of sorting by "Top"). We store the submission data as entries in memory, and then upload it into a google sheet using Python's gspread package.
>
> While the data is in memory, we analyze the URLs for every submission. If there is a URL, it will be identified as misinformation based on whether its domain name is in our list of misinformation domains. If the URL is identified as misinformation, it will be flagged as so.

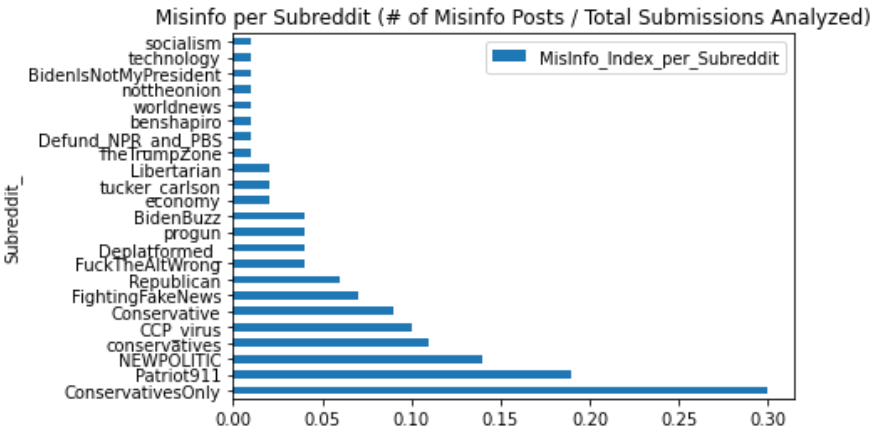
Uploading the Data

> Once the data is all saved in a dataframe, it'll be uploaded to a google sheet using the gspread package. This is done as a bulk upload, so it is fast and efficient.

Results

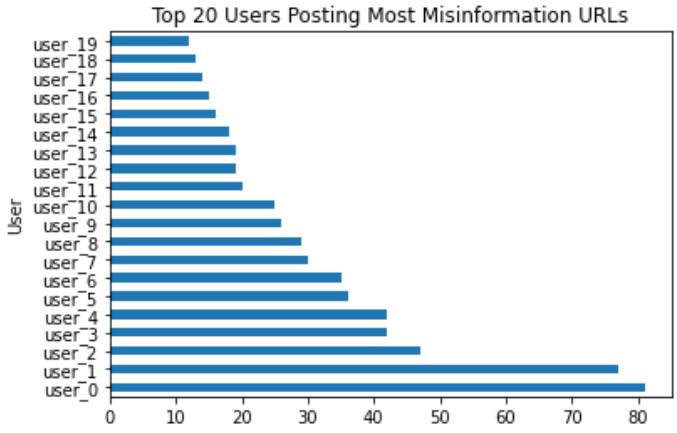
Exploratory Data Analysis

> The question we wanted to address is "How does the spread of misinformation vary from subreddit to subreddit, and what is the relationship between these subreddits?"
>
> We first identified all of the posts with misinformation URLs and created a separate dataframe. Afterwards, we created a bar chart that showed the count of misinformation posts based on each individual subreddit, giving us a good overview on the prevalence of misinformation in each community.

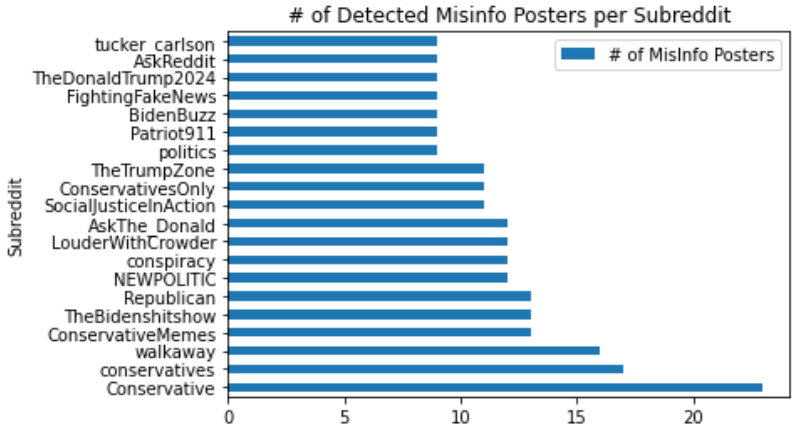


> Based on the misinformation submissions we have, we identify the users and rank them based on how much misinformation they individually post. This way, we can identify the most prolific users in regards to the spread of misinformation.

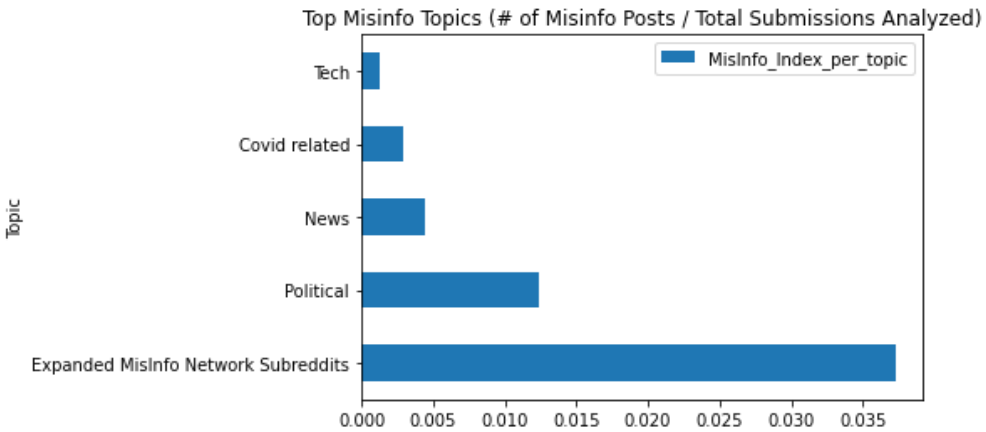
> Next, we find the subreddits these users are most active in. This is done by analyzing their comments.



> Finally, the subreddits that share the most users who have posted high amounts of misinformation are shown below.

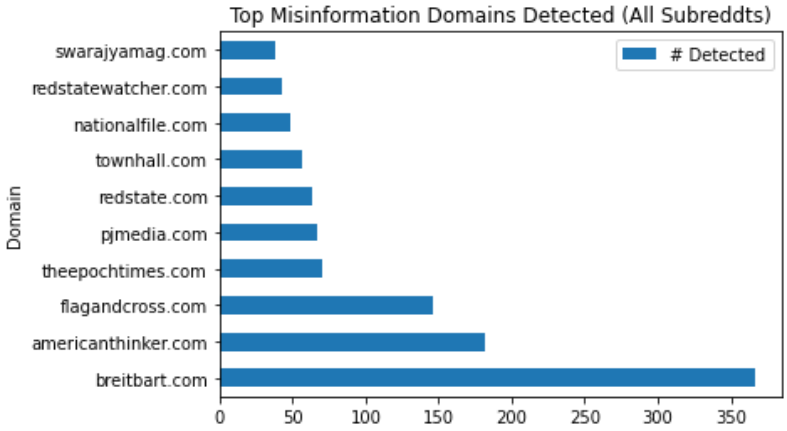


> This provides some intuition of the misinformation network on Reddit, because the subreddits with the most misinformation-posters are probably more integral to the misinformation network.
>
> In addition, we ranked the top categories based on their proportions of misinformation submissions, and we found that political and news-related subreddits tend to have a higher proportion of misinformation posts, which was inline with our initial assumption.

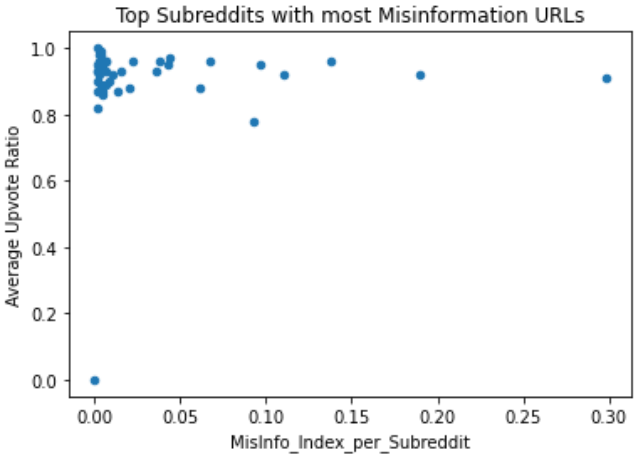


> Interstingly, we also ranked the domains that we detected the most, across all subreddits, and we found

> that "breitbart", "americanthinker", and "flagandcross" are posted the most often compared to other domains that we encountered.



> Lastly, we plotted subreddits on a scatter plot based off of their average upvote ratio versus the proportion of misinformation posts. We found that subreddits tend to have a similar average upvote ratio despite the proportion of misinformation they have. This indicates that subreddits are generally idealogically homogenous.



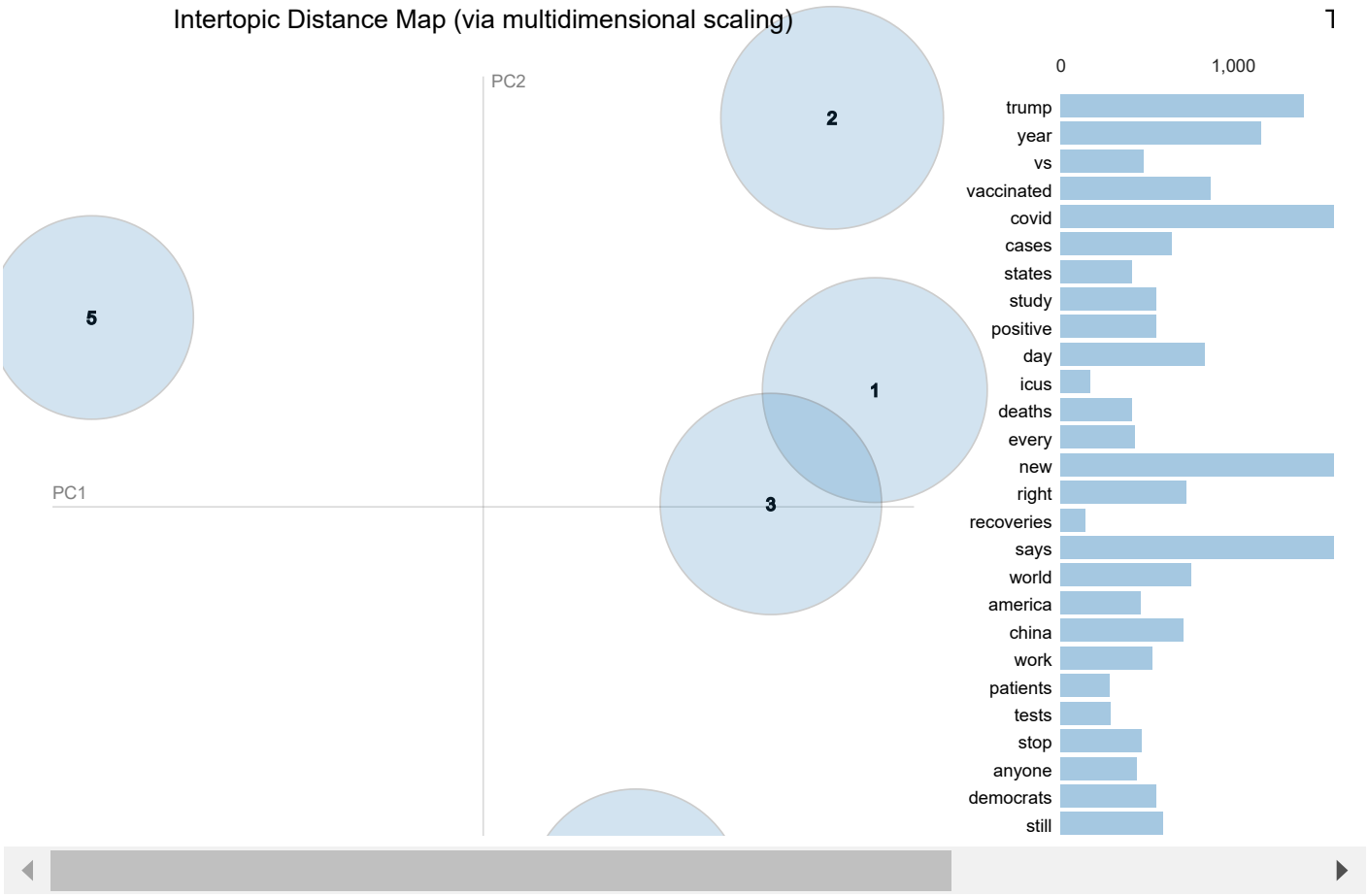
Further Analysis

> We also created an interactive topic model and undirected graph, which has been made publically available on our website. The topic model provides an interesting outlook on what topics can be derived from the word frequencies found when analyzing the titles of the misinformation submissions we detect. The distance between topics is displayed on a two dimensional graph, and a user can click on the topics to see what words they're made up of. As for the interactive undirected graph, the user can see which subreddits are at the periphery of the misinformation network and interact with the nodes to gain a better understanding.

Selected Topic: 0 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metri (2)

λ = 1



Discussion & Conclusion

> We began this study asking how the spread of misinformation may vary from subreddit to subreddit and what the relationship between these subreddits may entail. We found that subreddits tend to be idealogically homogenous despite the proportion of misinformation submissions. Also, we found that certain subreddits seem to play a larger role in the misinformation network than others, i.e. "Conservative", "Conservatives", "walkaway". We found that the top misinformation domains shared among the subreddit we analyzed are "breitbart", "americanthinker", and "flagandcross". Lastly, our hypothesis was confirmed to be correct when we found that political and news-related subreddits had a higher proportion of misinformation submissions that other subreddit categories. While our study is limited by the small sample size of submissions we analyze from each subreddit, our analysis seems to provide a thorough outlook on the misinformation network on Reddit because we analyze subreddits of different categories and expand our view of the misinformation network through utilizing user-comment activity.

Works Cited

1. Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In Proceedings of the 2017 Internet Measurement Conference (IMC '17). Association for Computing Machinery, New York, NY, USA, 405–417. DOI:https://doi.org/10.1145/3131365.3131390

2. Iffy+ MIS/disinfo sites. Iffy.news. (2021, June 20). Retrieved February 3, 2022, from https://iffy.news/iffy-plus/