DSC 180B Report Checkpoint

> 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.

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

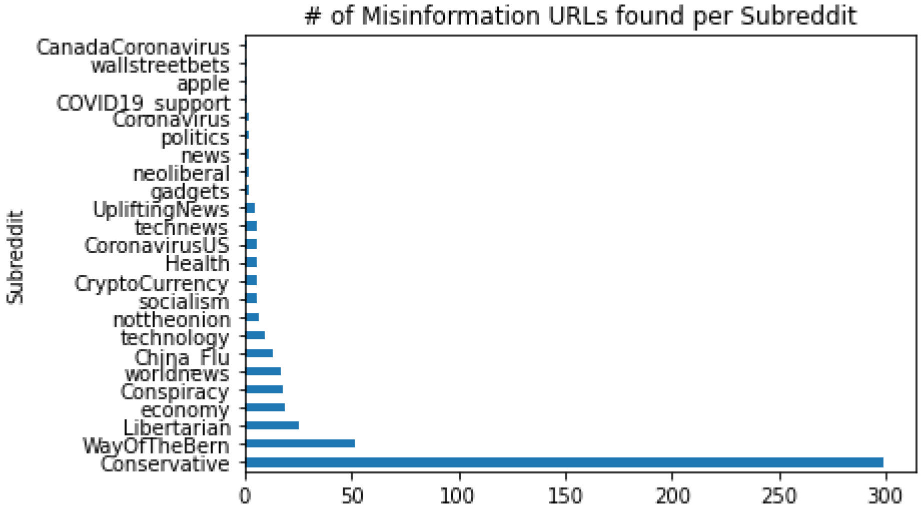
> Once the data is all saved in a dataframe, it'll be uploaded to a google sheet using the gspread package. This is done as a bulk upload, so it is fast and efficient.

Results

Exploratory Data Analysis

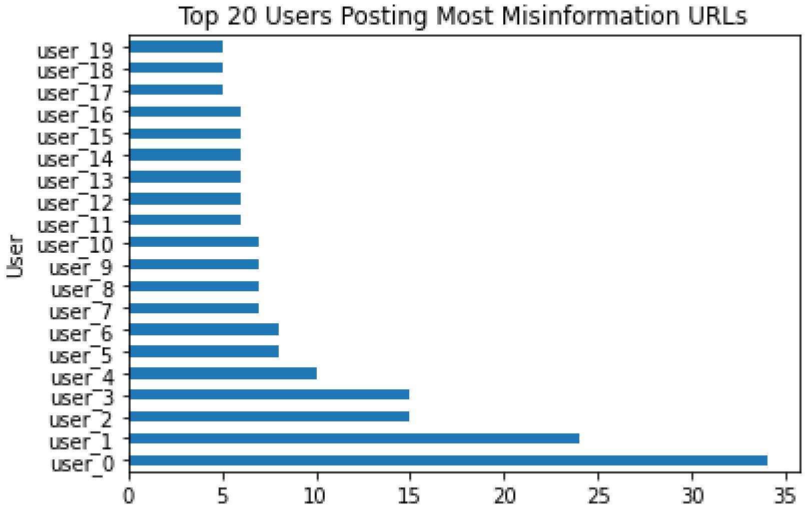
> The question we wanted to address is "How does the spread of misinformation vary from subreddit to subreddit, and what is the relationship between these subreddits?"

> 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.

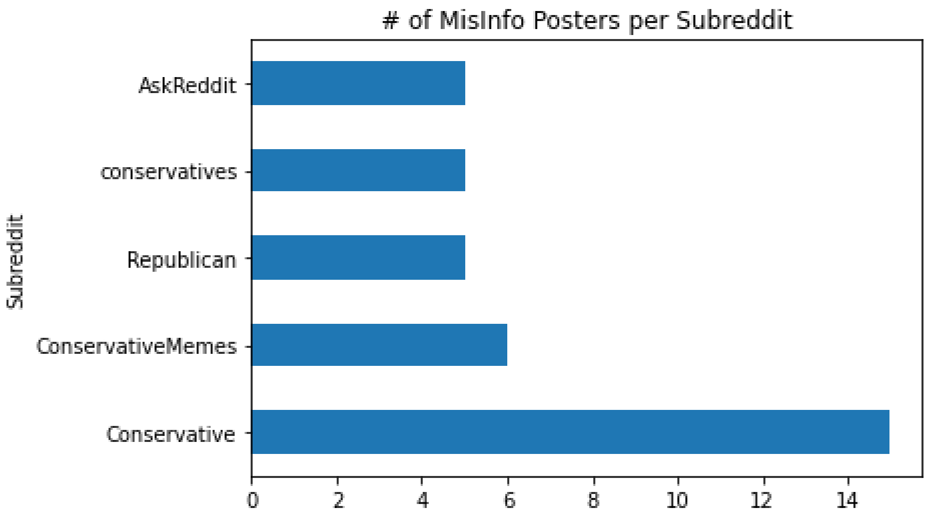


> Based on the misinformation submissions we have, we identify the users and rank them based on how much misinformation they individually post. This way, we can identify the most prolific users in regards to the spread of misinformation.

> 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.



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/

David Aminifard, Samuel Huang

DSC 180A, Section A10

12/05/21

Project Proposal Statement

As social media has grown in popularity, namely Reddit, its use for rapidly sharing information based on categories or topics (subreddits) has had massive implications for how people are usually exposed to information and the quality of the information they interact with. While Reddit has its benefits, e.g. providing instant access to - nearly - real time, categorized information, it has possibly played a role in worsening divisions and the spread of misinformation.

We are aiming to analyze the prevalence of misinformation in certain subreddits and the overlap of users between different subreddits. This is similar to our current project in that we are still analyzing the nature of misinformation in online social media platforms. However, this is different because Reddit fundamentally has a different structure than Twitter. Reddit exists in subgroups called “subreddits” which are typically built around a group of liked-minded people who are interested in a specific topic. Therefore, when studying the misinformation of a topic like COVID-19, it would be more about the subgroups that spread misinformation instead of influential individuals.

We decided to analyze Reddit because we thought it would be interesting to see similarities and differences between the spread of misinformation on Twitter compared to Reddit. While, many times, Twitter is compared to the public square, we believe

Reddit is more comparable to how social interactions occur in real life. Reddit exists in groups of like minded people similar to how humans typically group with people that they consider similar to themselves. Furthermore, on Reddit we can analyze how misinformation relates to certain topics. For example, if we are on a subreddit about Unix, we assume that it would be unlikely that we would find many posts that contain misinformation. However, if we are on a subreddit about politics, it would be more likely that we would find posts that contain misinformation.

So, in short, our project would be to download data using the Python Reddit API Wrapper or PRAW and the data will be used to assess the influence of misinformation given certain Subreddits and to showcase relationships among users who comment in those Subreddits. The PRAW documentation can be found here, where already publicly available subreddit and user data is made accessible, which we will use for our analysis. PRAW's source code (GitHub repo) is available here, it is evidently a popular library used for accessing data on the Reddit platform

We plan to create a website that shows the relationships between subreddits and the spread of misinformation. Also, the website will contain a ranking of subreddits based on the frequency of misinformation per subreddit.