2022

Coronavirus Misinformation on Reddit

Austin E. Lee
University of Vermont

Coronavirus Misinformation on Reddit

Austin Lee

Undergraduate Honors College Thesis

Department of Computer Science

College of Engineering and Mathematical Sciences

Thesis Advisors

Dr. Jeremiah Onaolapo

Dr. Laurent Hébert-Dufresne

University of Vermont

2022

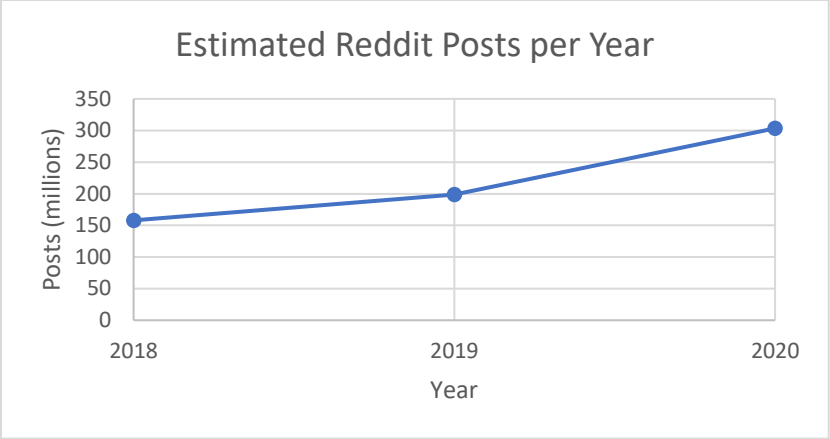
Contents

Abstract

Social media networks play a large part of our daily lives. With millions of posts and comments being processed per day, it is only natural that some of these posts present false information that could potentially harm more naïve people. Reddit, in particular, has seen a major rise in such posts and comments in past years. This paper gives an insight into how to best categorize Reddit comments as misinformation or not and determine the user's sentiments regarding the post using sentiment analysis. Finally, the findings will be presented in a meaningful and easily accessible way in the form of an online dashboard.

Introduction

Social media networks have long been a part of our daily lives. Millions of people are constantly engaging with social media- using it for entertainment, news/politics, and sharing personal content. Giant social media corporations, such as YouTube, Facebook, and Reddit process an estimated hundreds of terabytes of data from users every day, an amount that will only continue to increase [1]. Reddit, in particular, is a forum-like social media platform where users can submit posts or comment on posts in their respective "subreddits". Each of the approximate 3.1 million subreddits is like a separate forum revolving around a certain topic, for instance "r/politics" primarily focuses on the discussion of U.S politics [2]. Subreddits are also community moderated, meaning that they are governed by a volunteer group of people. Every day, Reddit attracts approximately 52 million users and in 2020 alone garnered 303.4 million posts and 2 billion comments as depicted in Figure 1 [2], [3]. As these numbers continue to grow, it becomes increasingly difficult to track and moderate uploads due to lack of resources, knowledge on the topic, or simply indifference. Because of this, misinformation can easily slip through the cracks of poor regulations and potentially negatively impact unknowing users.



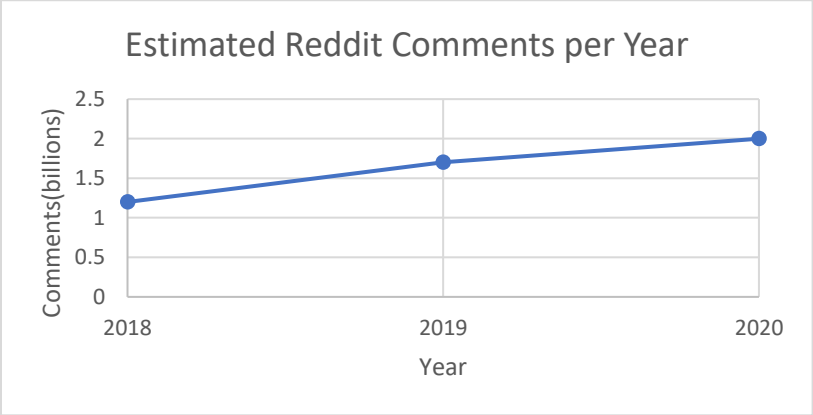


Figure 1: Estimated Reddit posts and comments uploaded between 2018 and 2020 [2].

The propagation of misinformation, whether intentional or not, is an ever-present issue that plagues social media platforms and, by extension, the greater Internet. Misinformation refers to when a person spreads a false statement as a fact. This act can be intentional or unintentional. For example, if a user were to post something along the lines of "vaccines cause autism", this would be a classic case of misinformation. Combine that comment with a few false statistics/claims and a shady link, then that post may convince another user that vaccines do, in fact, cause autism. To be clear, there is no scientific evidence that there is a relationship between vaccines and autism [4].

Misinformation can be incredibly difficult and ambiguous to define. It is very taxing to take a suspected misinformative post, prove that it is in fact false, and then take the appropriate measures to deal with it. Reddit, for instance, has no automated way to detect misinformative posts and instead relies on moderators to manually remove the content or other users to diligently prove it false. A huge issue with this is that this method relies heavily on the trust that the moderator or user makes the right call when flagging a post as misinformative. In addition, many people are already uncertain of the facts

and may end up improperly flagging these posts. Above all, the average person cannot be expected to have all the correct knowledge to accurately find misinformation 100% of the time [5]. As misinformation can be centered around any topic, the potential quantity can be quite enormous and therefore difficult to control.

Often times, misinformation can have devastating real-life consequences, misleading individual victims or communities into self-destructive behavior. For example, a review paper noted that out of about 1500 abstracts and 190 fully reviewed papers about health information, only 3 reported adverse health effects due to misinformation, one involving a poisoned dog, one involving emotional distress, and one involving a failed liver and kidney due to self-medication [6], [7]. While this number appears to be low, the paper notes that it may be much higher due to the possibility that the victim of misinformation did not consider their information as misinformation, did not notice any adverse effects due to the misinformation they followed, or did not remember where they learned the misinformation. It is also cited that 39% of people in the U.S believe in herbal remedies or vitamin supplements as potential alternative "cures" for cancer, which will likely not have any real medical benefits [6]. In addition, a study showed that YouTube videos which spread misinformation by downplaying the health risks imposed by smoking led viewers to have a more positive attitude towards smoking [8]. The study explains that younger people experienced the most positive attitudes towards smoking tobacco, specifically in e-cigarette and hookah products, and would potentially be more likely to seek out additional similar content, further solidifying their "self-justifying" beliefs [8]. There is a high possibility that these misinformed victims will then potentially propagate their beliefs further or engage in additional misinformative content, thus misinforming more people and hurting them as well. It can be inferred that there are similar adverse effects surrounding misinformation in health-related topics, such as coronavirus [9].

Recently, it seems that the spread and blind trust in misinformation has become much more prevalent. Since the beginning of the coronavirus pandemic in late 2019/early 2020, viral misinformation campaigns have gained a lot of traction. Although anti-medicine propaganda, for example anti-vaccine movements, have always been actively spread, misinformation focused on the coronavirus now seems to be the focus. Examples of this misinformation can deal with the severity of the coronavirus, effectiveness of the vaccines, alternative cures, and racially motivated sentiments [10]–[12]. Additionally, a study measured 5 different misinformation topics involving coronavirus in different media articles, being general misinformation/conspiracies, President Donald Trump mentions, Infodemic

coverage, fact checking, and finally Trump only articles. Overall, it found that among the misinformative articles, 46.6% contained general misinformation, 37.9% contained Trump mentions, 23.4% contained Infodemic ideology, 16.4% were fact-checked or did fact-checking, and finally 10.3% were only about Trump. Next, among the general misinformation, about 26% of that was about coronavirus miracle cures, and other random topics involving the Deep State, bio-weapons, anti-Semitism, and population control conspiracies comprised the rest [13].

This paper aims to explore a potential method for identifying, categorizing, and determining the user's sentiments surrounding coronavirus misinformation in Reddit comments using keyword searches and sentiment analysis. This paper outlines potential answers to these questions:

- Is there a way to quantify and categorize coronavirus misinformation on different subreddits? Then, can I determine the users' sentiments regarding the truthfulness of these comments?

The code is designed in a way such that it allows for easy replication and scalability, as well as the ability for future researchers to implement their own searches. For example, this project primarily searches for coronavirus related terms when checking for misinformation, however one may be able to use the code to search for other instances of misinformation topics. Finally, the data is presented in a user friendly and simple dashboard website. The point of the website is for everyday people to extrapolate data and make sense of it without a strong background in computer science, social media behavior, or misinformation theory. The dashboard can operate in semi-real time, meaning that it will most always be up to date with a sample of the latest Reddit comments. Due to technical knowledge limitations, this dashboard will be semi-real time unlike the real time dashboards featured in [14] and [15].

Background

In recent years, social media data scraping and statistical/empirical analyses have become more prevalent topics among computer science research communities and others. This massive stream of data is constantly changing with the emergence of new real-world events and Internet trends. In the case of coronavirus misinformation, the severity of coronavirus, anti-vaccine sentiments, and racially charged biases (among others) are all targets for malicious misinformation creators [10], [15], [16]. For example, "it's just the flu" or "vaccines cause autism" are common misbeliefs/phrases in victims of misinformation.

To collect Reddit data, researchers frequently use Pushshift API, a Reddit submission aggregator and searching tool. After using this tool to obtain a dataset, researchers would be able to parse and utilize the data as they see fit. For example, one may be able to search all recent comments for the keyword "science", and the API could return a JSON file of all instances of the word, the average amount of times it appears in a comment, and other statistics based on what was specified in your data key. [17], for instance, collects QAnon related data from a site called Voat instead of Reddit. First, a sample of various Voat "subverses" (in my case it will be Reddit "subreddits"), was collected based on their relation to QAnon sentiments. This was done empirically. Researchers implemented a depth first search algorithm to search each post's comment's replies, and all instances of search terms were cataloged. The researchers also made sure to collect all public data on users with notable contributions. This process can be seen in other papers such as [13], [18]–[21].

Sentiment analysis is the use of natural language processing and text analysis to determine the overall sentiment of a corpus of words. It is designed to figure out the opinion of the author and how he/she felt writing the text [22]. NLTK's VADER tool specifically deals with social media posts. NLTK's researchers build a gold standard list of lexical features, using quantitative and empirical methods, that mapped to different sentiment scores. This was used to fuel their machine learning algorithm that determines an author's sentiment of a social media post [23]. This tool ended up performing as well as or even better than other competing sentiment analysis tools.

The StoryWrangler project gathers a constant stream of Twitter posts, infers patterns based on specific words, phrases, or hashtags, and finally creates graphs and visuals and presents them on a user-friendly website [15]. While its focus is more on social trends or political discourse rather than misinformation, StoryWrangler provides an excellent way to analyze the popularity and propagation of these online trends in real time. The Hoaxy project deals specifically in misinformation, namely the initial creation, spread, and subsequent fact checking by other users [14]. Again, this data is presented as a website that allows you to search for frequencies of key terms, who is sending the post, and general user activity. To present this data, I want to create a dashboard with data similar to UVM's StoryWrangler

project or Hoaxy [14], [15]. The dashboard portion of this paper is inspired by those built in these projects. This paper aims to fuse together portions of each of these areas; data collection, data analysis through sentiment analysis, and data presentation via an easily accessible dashboard.

Data/Methods

Pushshift.io is a massive archive consisting of data from Reddit. The Reddit data is copied into the database in real time and allows researchers to retrieve certain data from older time periods as well. Comments and posts are stored in the database as soon as they are posted for use in the Pushshift API or website archives. This project focused on Reddit comments scraped from several different subreddits [24].

In this project's dataset, a single comment data point consists of 13 subpoints. These are "author", which refers to the comment's writer/submitter, "author_fullname", which refers to an author's unique tag, "author_premium", which determines if the author had Reddit Premium at the time of the posting, "body", which refers to the comment's text, "created_utc", which refers to when the comment was submitted, "id", which is the unique ID tag of the comment, "link_id", which refers to which comment chain the comment links to, "parent_id", which refers to which comment/post the comment was submitted under, "permalink", which links to the webpage of the comment, "score", which refers to the total upvotes the comment received, "subreddit", which refers to the subreddit the comment was posted to, "subreddit_id", which refers to the subreddit's unique ID tag, and finally "total_awards_received", which refers to the total number of Reddit awards awarded to the comment.

Overall, two separate datasets were collected, the first consisting of Reddit comments from January 2021 to September 2021 and the second consisting of comments pulled in real time from February 2022 to present. The first dataset is a random sample of at most 2000 comments per week between that time. Currently, the second dataset is constantly being updated using a Linux cronjob, which pulls in data every hour. This is an aggregation of all the comments posted to each subreddit within each hour. While there is a limit of 2000 comments per request, there is usually not that many comments to pull from the site. The second data set does not have data on the scores/upvotes for each comment, as each comment had been collected at the exact time of posting and therefore would not have accumulated any. Furthermore, Pushshift retroactively collects score/upvote data, so the first dataset contains that data. Each dataset contains data from 11 different subreddits. A breakdown of this data can be found in Figure 2.

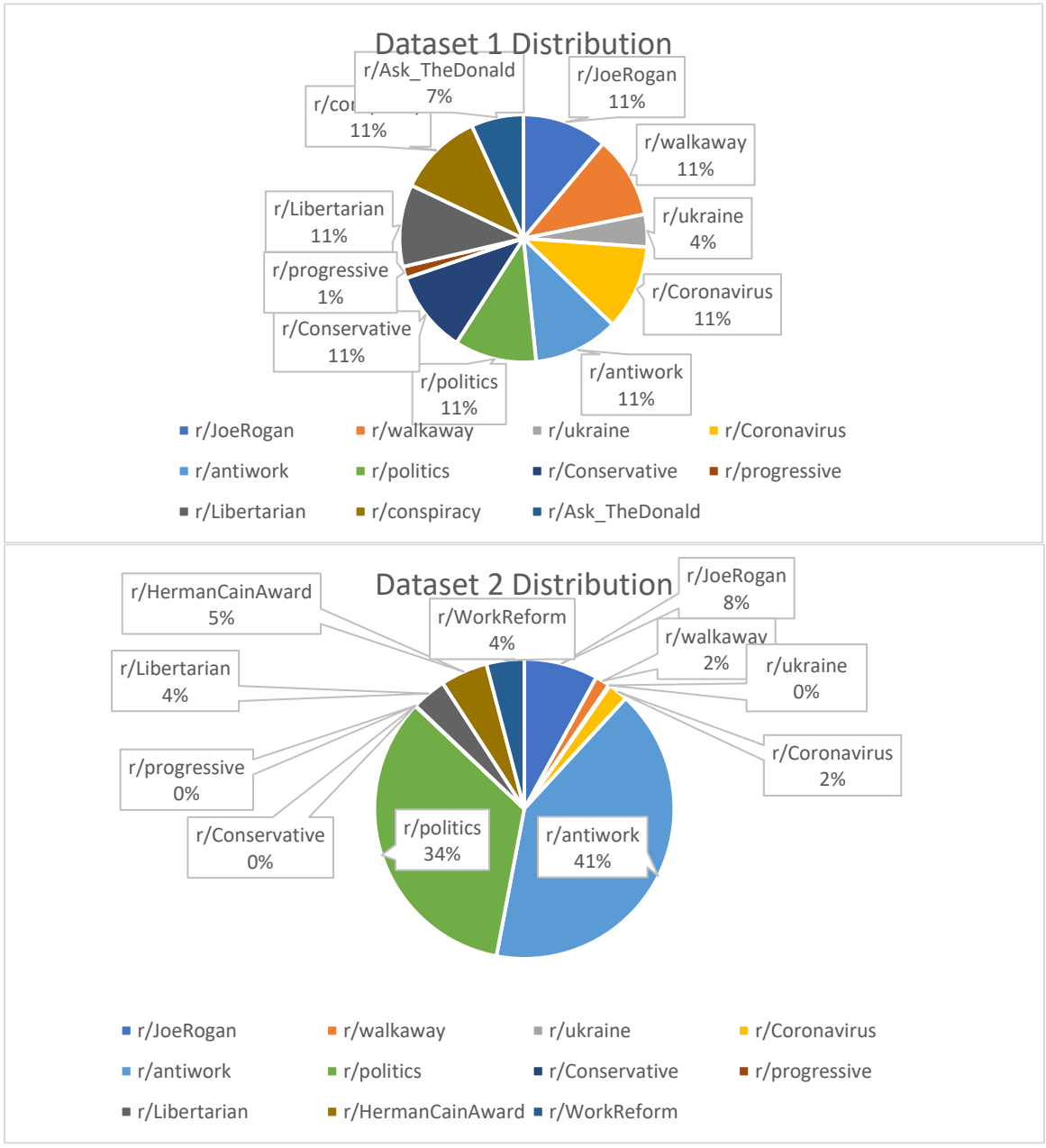


Figure 2: This represents the comments collected per subreddit distribution. Top is the first dataset and bottom is the second dataset.

Table 1: Breakdown for number of comments collected and unique authors per subreddit.

Subreddit	Number of Comments (Set 1)	Unique Authors (Set 1)	Number of Comments (Set 2)	Unique Authors (Set 2)
r/JoeRogan	56000	18719	89985	19771
r/walkaway	54048	9855	17408	5203
r/ukraine	21462	3581	1528	915
r/Coronavirus	56000	20765	24356	9457
r/antiwork	56000	21844	463880	144249
r/politics	54000	31060	383805	103474
r/Conservative	54000	16602	317	199
r/ progressive	7630	1710	174	134
r/ Libertarian	54000	12206	42316	6971
r/conspiracy	56000	18499	N/A	N/A
r/ AskThe_Donald	34412	5458	N/A	N/A
r/HermanCainAward	N/A	N/A	57325	13817
r/WorkReform	N/A	N/A	45946	13817
Total	503552	143116	1127040	283558

Each subreddit in each dataset was chosen based on empirical analysis. Subreddits such as these are relatively popular, often hit the front page (r/all) and tend to have some posts regarding coronavirus. r/JoeRogan was chosen based on the community's views regarding vaccines. Joe Rogan, a podcaster who regularly talks about current events, politics, and science, has pushed antivax and pro-ivermectin sentiments onto his viewers during his podcasts. Ivermectin is a drug that helps to cure parasitic diseases [25]. There are two types, one for humans and one for horses, and due to an onslaught of pro-ivermectin misinformation, many people ended up taking the one for horses to try and cure coronavirus. The FDA has not recommended the use of either type of ivermectin to cure coronavirus [25]. In addition, he has also featured guests who are antivax. Overall, r/JoeRogan is a prime example of a hotspot for coronavirus misinformation. r/walkaway was also chosen based on the community's views. The #walkaway movement is a campaign on people's reasons for "walking away" from liberal political beliefs. In the past, there have been frequent posts speaking out against the lockdowns or mask mandates in response to the coronavirus epidemic. r/ukraine was chosen based on

current events. The subreddit grew from around 30k subscribers to upwards of 600k following the Russian invasion of Ukraine [26]. r/russia, on the other hand, recently went private due to backlash. Because of this, the Pushshift API was unable to collect data from r/russia. r/Coronavirus was chosen based on relevance to the topic, as well as r/conspiracy. r/antiwork and r/WorkReform were each chosen based on their recently exploding userbase. Both subreddits revolve around workplace culture including practices, work-life balance, and work-related grievances. With anti-coronavirus mandates implemented at nearly every company, it would be reasonable to assume that people would talk about them in these subreddits. The political subreddits, r/politics, r/Conservative, r/progressive, r/Libertarian, and r/AskThe_Donald were each chosen to have a gauge on if misinformation was being passed around in more politically oriented subreddits. Note that due to an error in the data collection, most of the data for r/Conservative was corrupted in the second dataset, which is why the numbers are so small. Finally, r/HermainCainAward is a subreddit dedicated to making fun of people who publicly express anti-vaccine or coronavirus hoax sentiments and then later suffer consequences (usually dying) for their misguided beliefs. It is more of a joke subreddit, but still functions as popular forum for coronavirus related topics and misinformation.

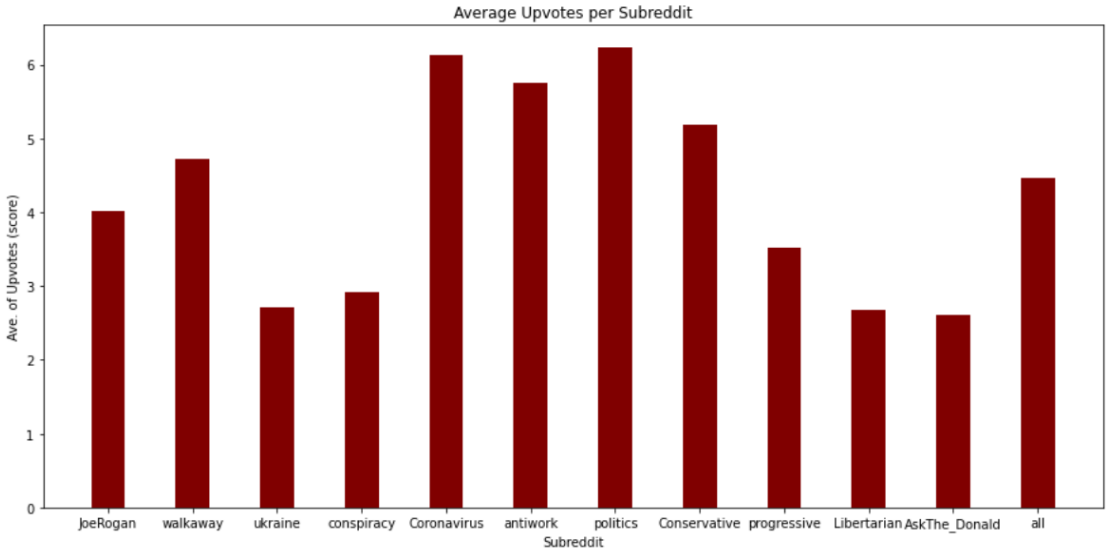


Figure 3: This represents the average number of upvotes on each comment per subreddit for the first dataset. Because the second dataset's information was collected as soon as the comments were posted, no upvote data was gathered.

Results

After collecting the data, it was then searched for instances of certain keywords. This dictionary of keywords was built using common terms relevant to different subreddits and popular coronavirus related words. If a term from the dictionary was located in the body of the comment, the code would mark that comment as a candidate for further analysis. Overall, the dictionary contained 37 terms, ranging from different names for coronavirus (such as "rona" and "covid-19") to common misinformation buzzwords (such as "jab" or "ivermectin") as noted in Table 2. In the first dataset, 69733 comments were flagged, which accounts for about 6.18% of the whole dataset. The most flagged keyword was "covid", appearing in 16490 comments. In the second dataset, 86307 comments were flagged using this dictionary, which accounts for about 17% of the whole dataset. The most flagged keyword was "vaccine", appearing in 14112 comments.

Table 2: This Table contains a sample of flagged comments for each word.

Term	Comments Present (Dataset 1)	Percent of Data (Dataset 1)	Comments Present (Dataset 2)	Percent of Data (Dataset 2)
coronavirus	1738	0.003451	636	0.000564
corona	2037	0.004043	803	0.000712
covid-19	2172	0.004313	1069	0.000948
covid	12089	0.024007	16490	0.014631
vaccine	14112	0.028024	7115	0.006312
virus	6012	0.011939	3070	0.002723
vaccinated	6037	0.011988	3985	0.003535
hydroxychloroquine	76	0.000151	74	0.000006
jab	538	0.001068	447	0.000396
ivermectin	224	0.000448	698	0.000619
hoax	356	0.000707	424	0.000376

The next step was to perform a sentiment analysis on the flagged data. The first step in this function was to tokenize the data. Essentially, the body of a single comment was broken down into an array of strings, each containing a value for the sentiment. This was done using NLTK's tokenizer functions [27]. Next, all stop words were removed from the array. Stop words are words that will not

add to the meaning of the sentence. In this case, the array was scrubbed of all words inside spaCy's (en_core_web_sm) default stop words list [28]. Words like "and" or "of" would be removed in this fashion. Afterwards, each string would be lemmatized, or in other words broken down to the base word. For example, the words "comes" would lemmatize to "come". This was done using NLTK's WordNetLemmatizer() function [27].

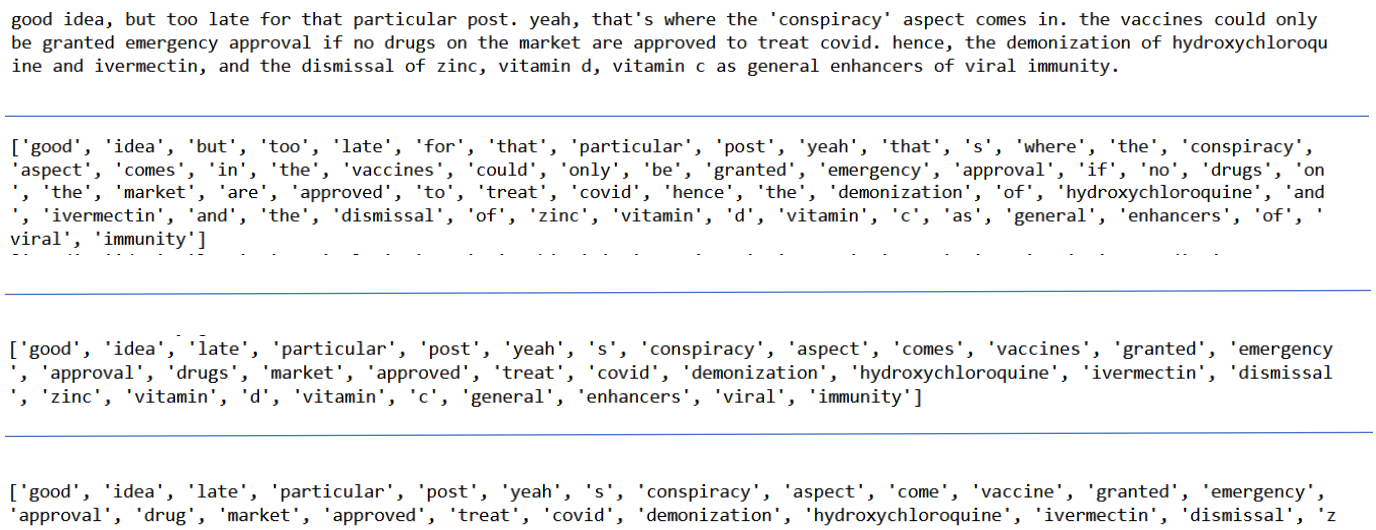


Figure 4: Top features a body of text about to undergo the sentiment analysis. Top middle features the body of text after being tokenized. Bottom middle features the array after removing all the stop words based on the list. Bottom features the lemmatized array.

After completion of cleaning, tokenizing, and lemmatizing, the comment's array was then analyzed based on NLTK's pre-trained sentiment analysis tool, VADER [23]. VADER is a sentiment analysis tool specifically made for social media posts. The VADER tool is used to compute the polarity of the post, either positive or negative. Each word in its dictionary is associated with a sentiment score, with scores closer to 1 being more positive and scores closer to -1 being more negative. In addition, a 0 denotes that the word is neutral. For example, the word "good" would receive a +1 for being positive. After computing the score for each word (either a -1, 0, or 1), an average is taken of the whole text. This is then used to determine the overall sentiment of the post. For example, if the body of the comment expressed that "vaccines are good", then it would receive a high positive score as most of the important terms are either positive or neutral. Users can then use this data to extrapolate the general sentiment

surrounding certain topics in a subreddit. If the term “ivermectin” were trending with a positive score, one may infer that pro-ivermectin misinformation is being passed around in the subreddit [23].

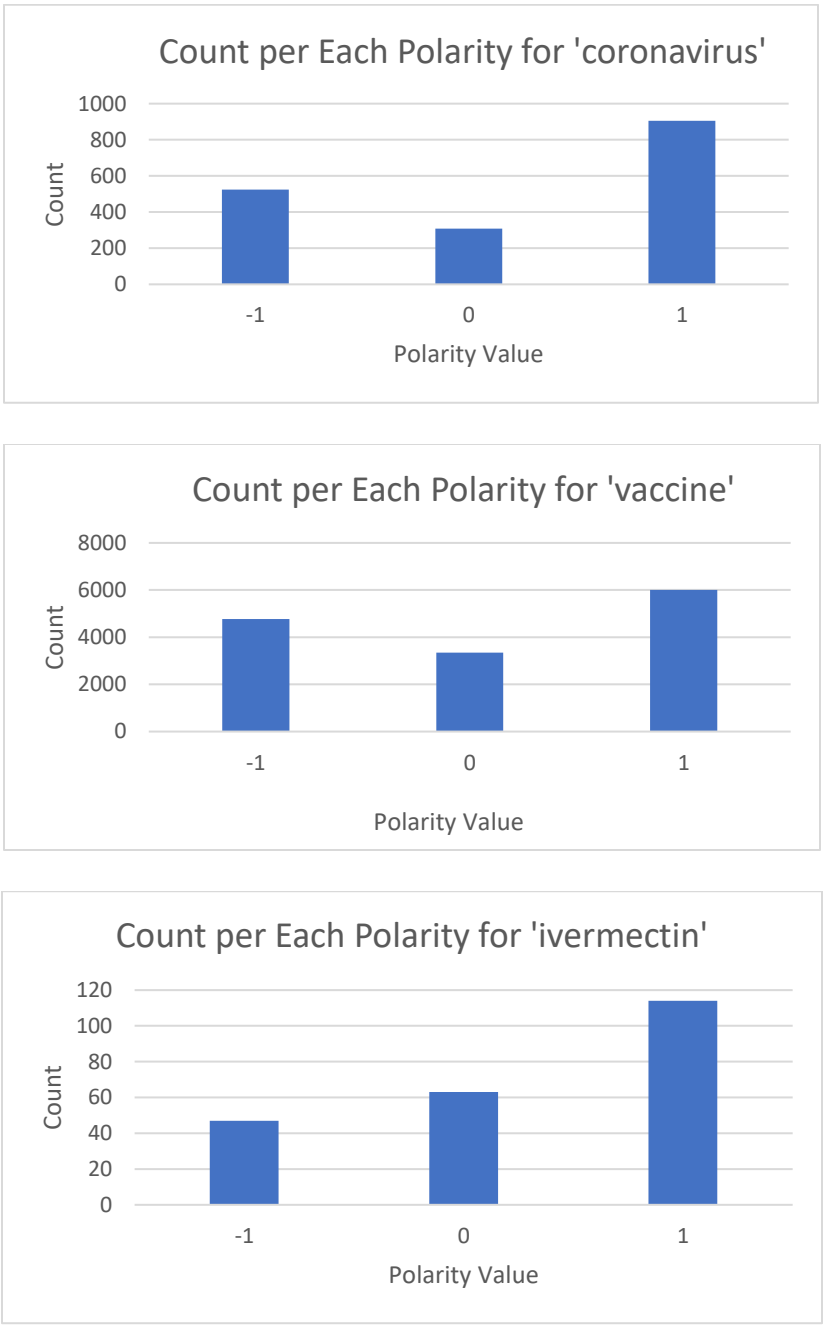


Figure 5: These graphs depict frequency counts of the polarity for each post containing the words “coronavirus”, “vaccine”, and “ivermectin”. Based on the data, the term “coronavirus” tends to have more positive posts, “vaccine” is fairly mixed, and “ivermectin” also tends to be positive.

Finally, after collection and analysis, the data is passed into a PostgreSQL database where it is then linked to the dashboard webpage. Currently, the dashboard features the information gathered in the results as easy to read graphs. It also has a search function that allows users to locate specific authors/keywords or to browse the dataset based on number of upvotes, subreddit, and misinformative posts. As of now, the dashboard does not have real time functionality. Instead, it requires manual insertion of data into the PostgreSQL database after collection, cleansing, and analysis. That being said, the graphics and analytics of the website are set up in a way such that you simply need to drag and drop a finalized .csv file of data into the PostgreSQL database, and the rest of the site will automatically update. A graphic for this process can be found in Figure 6. This allows for easy scalability and reproduction for future datasets. Inspiration for the dashboard was taken from [14] and [15].

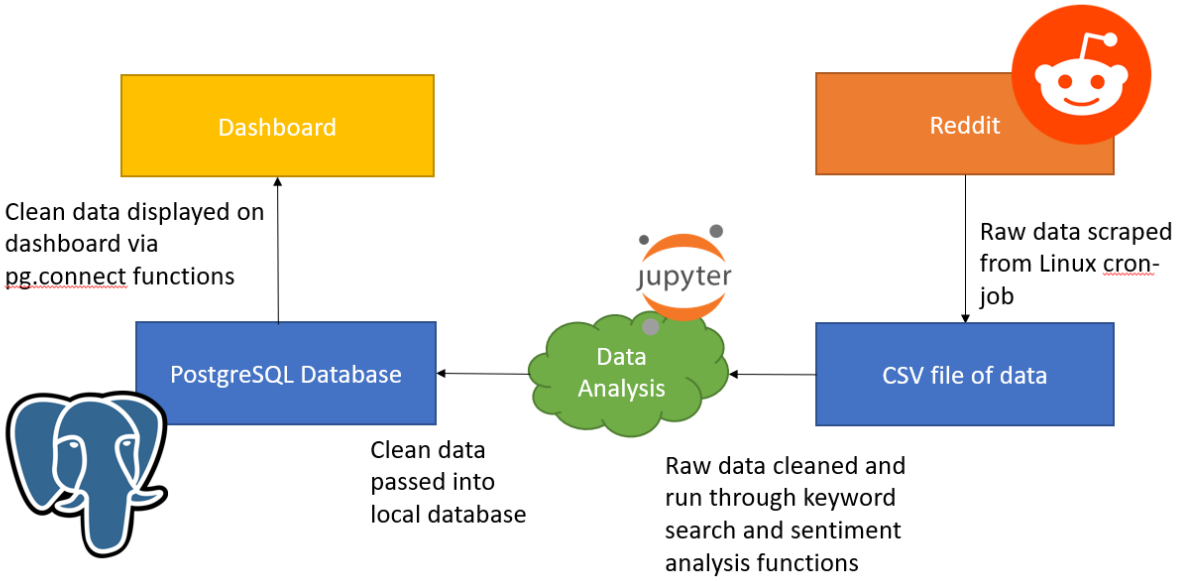


Figure 6: Flow chart depicting process for collecting data, cleaning/analyzing data, and finally displaying the data.

Discussion

I think this methodology for detection and sentiment analysis worked well on a lower level. A keyword search is by far the simplest method for detection, however, is limited by the depth of the keyword dictionary. The keyword search could easily identify posts that had misinformation buzzwords and topics. By searching for terms such as “hydroxychloroquine” or “ivermectin”, it can be surmised that the post is misinformative in nature, as both keywords tend to be associated with miracle cure type misinformation. This, of course, can be extended to any number of terms as the keyword search is an

easy process. Next, the sentiment analysis portion of the methodology worked well for determining the average sentiment of each post. By taking the sentiment scores of each post, I was able to determine the average sentiment for each keyword topic. Featured in Figure 5 are the terms coronavirus, vaccine, and ivermectin. We can see that the term "vaccine" was a very neutral term despite the assumed controversy appearing in anti-vax campaigns and misinformation. The term "ivermectin" was incredibly positive, showing that people perceive this topic in a more positive light than what may be ideal. And finally, the term "coronavirus" was also seen in a positive light. At first glance, this finding may not make so much sense, as one may assume that "coronavirus" may be more negative or even neutral. However, I suspect it may have to do with posts along the lines of "x country did well with coronavirus prevention efforts", where "well" is the positive term.

Limitations

One of the biggest limitations of this project was the number of comments collected. Due to request limitations on the Pushshift API, the program could only fetch 2000 comments per week for the first dataset. The time frame could have been lower; however, it would have greatly increased the time needed to collect the data. The second dataset was not as hurt by this limit, as 2000 comments were almost never submitted to a subreddit in an hour. The section above mentioned an unexpected finding with the term "coronavirus". Based on how this term was analyzed, the sentiment analysis seemed to struggle with more factual/non emotional type posts. A user posting "coronavirus killed x amount of people" may not, in actuality, contain a sentiment, however the algorithm may pick up the word "killed" and associate that as a negative post. I suspect that this reason played a part in determining the sentiments of each keyword. Within the detection part of the methodology, it is impossible to determine between people spreading misinformation and people talking about other people spreading misinformation. This may be solved by utilizing an analysis on the parent comments of each comment. Finally, due to limitations in web design experience, the dashboard requires an extra manual update feature to load in new data. This was implemented in the form of a Python script which updates the graphical images on the dashboard when called upon.

Future Work

At this time, the most important thing for this project is a more thorough dataset. Because of time limitations and data collection errors, one of the datasets contained data that only spanned about 7 months of 2021. Subreddits such as r/WorkReform, created January 2022, were created after this time resulting in potential disconnects between analysis of each dataset. Furthermore, there was an issue regarding the dataset where the data was scraped immediately after posing, in which no

upvote/score data was recorded. Ideally, the dashboard could feature a method in which after acquiring, analyzing, and presenting this data, it would then go back after a certain time period and recollect the upvote/score data. Next, a bigger dataset could lead to additional methods for misinformation detection and categorization. For example, a machine learning algorithm could probably analyze each comment much better than the keyword/sentiment analysis approach. In addition, the dashboard could include a feature in which users can help categorize data as well in order to build a classification dataset for the ML algorithm to use. A method such as this would require some parts of the dashboard to be rewritten using Ajax (currently written using PHP and JavaScript), a web development concept used for implementing asynchronous features [29]. An upgrade to Ajax frameworks would also be the answer for changing the data collection, analysis, and presentation from semi-real time into a fully automated, real-time structure. Finally, a real-time time series chart may be an interesting graphic to implement in the site as well, as time data is also being collected by the web scraper however it is not being used.

Conclusion

This research project established a means to collect Reddit comments through the use of Pushshift API, clean those comments, mark the comments for potential instances of misinformation and determine the overall sentiment surrounding the post. Then, the data can be stored in the PostgreSQL database which is linked to a user-friendly dashboard for casual use by researchers, academics, and everyday people alike. I hope these findings will help expose some of the patterns and generalities of coronavirus misinformation on Reddit (and eventually other social media platforms) to everyone. Misinformation will only become more widespread as long as the Internet exists, therefore, it is important that people understand what this misinformation looks like, where the misinformation is coming from, and what the misinformation is trying to accomplish, in order to better detect and avoid it.

Works Cited

[1] "Digital 2021 July Global Statshot Report," *DataReportal – Global Digital Insights*. https://datareportal.com/reports/digital-2021-july-global-statshot (accessed Oct. 28, 2021).

[2] "Reddit User and Growth Stats (Updated Oct 2021)," *Backlinko*, Feb. 25, 2021. https://backlinko.com/reddit-users (accessed Apr. 02, 2022).

[3] "Reddit Statistics For 2022: Eye-Opening Usage & Traffic Data," Jan. 07, 2021. https://foundationinc.co/lab/reddit-statistics/ (accessed Apr. 02, 2022).

[4] F. DeStefano, "Vaccines and Autism: Evidence Does Not Support a Causal Association," *Clin. Pharmacol. Ther.*, vol. 82, no. 6, pp. 756–759, 2007, doi: 10.1038/sj.clpt.6100407.

[5] N. M. Krause, I. Freiling, B. Beets, and D. Brossard, "Fact-checking as risk communication: the multi-layered risk of misinformation in times of COVID-19," *J. Risk Res.*, vol. 23, no. 7–8, pp. 1052–1059, Aug. 2020, doi: 10.1080/13669877.2020.1756385.

[6] B. Swire-Thompson and D. Lazer, "Public Health and Online Misinformation: Challenges and Recommendations," *Annu. Rev. Public Health*, vol. 41, no. 1, pp. 433–451, Apr. 2020, doi: 10.1146/annurev-publhealth-040119-094127.

[7] A. G. Crocco, M. Villasis-Keever, and A. R. Jadad, "Analysis of Cases of Harm Associated With Use of Health Information on the Internet," *JAMA*, vol. 287, no. 21, pp. 2869–2871, Jun. 2002, doi: 10.1001/jama.287.21.2869.

[8] D. Albarracin, D. Romer, C. Jones, K. Hall Jamieson, and P. Jamieson, "Misleading Claims About Tobacco Products in YouTube Videos: Experimental Effects of Misinformation on Unhealthy Attitudes," *J. Med. Internet Res.*, vol. 20, no. 6, p. e229, Jun. 2018, doi: 10.2196/jmir.9959.

[9] X. Li, X. Zhong, Y. Wang, X. Zeng, T. Luo, and Q. Liu, "Clinical determinants of the severity of COVID-19: A systematic review and meta-analysis," *PLOS ONE*, vol. 16, no. 5, p. e0250602, May 2021, doi: 10.1371/journal.pone.0250602.

[10] F. Tahmasbi *et al.*, "'Go eat a bat, Chang!': On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19," *ArXiv200404046 Cs*, Mar. 2021, Accessed: Apr. 23, 2021. [Online]. Available: http://arxiv.org/abs/2004.04046

[11] L. Bursztyn, A. Rao, C. P. Roth, and D. H. Yanagizawa-Drott, "Misinformation During a Pandemic," National Bureau of Economic Research, w27417, Jun. 2020. doi: 10.3386/w27417.

[12] Y. Wang, M. McKee, A. Torbica, and D. Stuckler, "Systematic Literature Review on the Spread of Health-related Misinformation on Social Media," *Soc. Sci. Med.*, vol. 240, p. 112552, Nov. 2019, doi: 10.1016/j.socscimed.2019.112552.

[13] S. Evanega, M. Lynas, J. Adams, and K. Smolenyak, "Coronavirus misinformation: quantifying sources and themes in the COVID-19 'infodemic,'" p. 13.

[14] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer, "Hoaxy: A Platform for Tracking Online Misinformation," in *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, Montréal, Québec, Canada, 2016, pp. 745–750. doi: 10.1145/2872518.2890098.

[15] H. H. Wu *et al.*, “Say Their Names: Resurgence in the collective attention toward Black victims of fatal police violence following the death of George Floyd,” *ArXiv210610281 Phys.*, Jun. 2021, Accessed: Sep. 28, 2021. [Online]. Available: http://arxiv.org/abs/2106.10281

[16] J. Hua and R. Shaw, “Corona Virus (COVID-19) ‘Infodemic’ and Emerging Issues through a Data Lens: The Case of China,” *Int. J. Environ. Res. Public. Health*, vol. 17, no. 7, Art. no. 7, Jan. 2020, doi: 10.3390/ijerph17072309.

[17] A. Papasavva, J. Blackburn, G. Stringhini, S. Zannettou, and E. De Cristofaro, “‘Is it a Qoincidence?’: An Exploratory Study of QAnon on Voat,” *ArXiv200904885 Cs*, Feb. 2021, Accessed: Apr. 23, 2021. [Online]. Available: http://arxiv.org/abs/2009.04885

[18] A. Patel and K. Meehan, “Fake News Detection on Reddit Utilising CountVectorizer and Term Frequency-Inverse Document Frequency with Logistic Regression, MultinominalNB and Support Vector Machine,” in *2021 32nd Irish Signals and Systems Conference (ISSC)*, Jun. 2021, pp. 1–6. doi: 10.1109/ISSC52156.2021.9467842.

[19] A. Park and M. Conway, “Tracking Health Related Discussions on Reddit for Public Health Applications,” *AMIA. Annu. Symp. Proc.*, vol. 2017, pp. 1362–1371, Apr. 2018.

[20] V. Setty and E. Rekve, “Truth be Told: Fake News Detection Using User Reactions on Reddit,” in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, New York, NY, USA, Oct. 2020, pp. 3325–3328. doi: 10.1145/3340531.3417463.

[21] M. Glenski and T. Weninger, “Predicting User-Interactions on Reddit,” in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, New York, NY, USA, Jul. 2017, pp. 609–612. doi: 10.1145/3110025.3120993.

[22] R. Feldman, “Techniques and applications for sentiment analysis,” *Commun. ACM*, vol. 56, no. 4, pp. 82–89, Apr. 2013, doi: 10.1145/2436256.2436274.

[23] C. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text,” *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 8, no. 1, Art. no. 1, May 2014.

[24] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, “The Pushshift Reddit Dataset,” p. 10.

[25] E. Starkman, “What Is Ivermectin?,” *WebMD*. https://www.webmd.com/drug-medication/what-is-ivermectin (accessed Apr. 27, 2022).

[26] “Subreddit Stats - statistics for every subreddit.” https://subredditstats.com/ (accessed Apr. 08, 2022).

[27] “NLTK :: Natural Language Toolkit.” https://www.nltk.org/index.html (accessed Apr. 08, 2022).

[28] “English · spaCy Models Documentation,” *English*. https://spacy.io/models/en (accessed Apr. 08, 2022).

[29] “Ajax (programming),” *Wikipedia*. Apr. 08, 2022. Accessed: Apr. 08, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Ajax_(programming)&oldid=1081564076