Introduction: A Pandemic of Fear

There is a dual pandemic raging, a pandemic of fear and disease. In the digital age, information is more accessible than ever before. Much like the Corona virus itself, the reproductive ratio of information today is extremely high. With the rapid and far stretched availability of information comes the issue of reliability. Historically, news and media outlets like print newspapers and radio have been the exclusive proprietor of news for the public. With the invention of the internet and social media platforms, this is no longer a reality. The consumption of information has shifted away from journalism and welcomed opinion and unverified sources from strangers on social media. In this new reality, certainty of information is threatened and rumours outpace scientific fact. As subjective opinions and bias muddle the truth then reliability becomes a question and fear is bred.

Scapegoating has been a common thread throughout history especially in relation to infectious disease. In tracking disease over time, you will find that history does not in fact repeat, it rhymes. While fear and panic is the common thread that holds these events together, the expression of such fear has reached greater heights in the digital age. Today, scapegoating and xenophobia has taken a new shape as it hides behind a screen.

Long before germ theory existed, minority groups were often blamed for disease. In 1348, during the peak of the Bubonic Plague in medieval Europe, it was estimated that the plague had killed 25 million people.[1] Dr. Liise-anne Pirofski, chief of infectious diseases at Albert Einstein College of Medicine, has suggested that "When disease strikes and humans suffer… the need to understand why is very powerful. And, unfortunately, identification of a

[1] "Jewish History Sourcebook: The Black Death and the Jews 1348-1349 CE," Internet History Sourcebooks Project, accessed May 12, 2020, https://sourcebooks.fordham.edu/jewish/1348-jewsblackdeath.asp.

scapegoat is sometimes inevitable."[2] By the fall of 1348 a rumour had circulated widely that the deaths were due to an international conspiracy of Jewry to poison Christendom. The consequences of this rumour lead to a massacre of Jewish communities across Medieval Europe. With Jewish communities accused of spreading contagion and poisoning wells, between the years of 1348 and 1351, more than 200 Jewish communities were targeted and wiped out completely.[3] In moments like these it becomes evident that as humans search for answers to unexplainable and life threatening events, pre existing biases can fuel irrational fears and suspicions. In this case, many lost their lives to acts of fear in addition to the plague itself. Although the bubonic plague received many names throughout history, it was never referred to as the "Jewish Death". Beyond this, many viruses and pandemics throughout history have been named in conjunction with blame.

Many news sources today have likened COVID-19 to the 1918 flu. Even today the 1918 flu is widely referred to as the Spanish flu. It is a misnomer that the name is a result of the flu's origins. In fact, the flu actually originated in Kansas. John M. Barry's "The Great Influenza: The Epic Story of the Deadliest Plague in History," provides a history of the epidemic. Barry's research suggests that the first identifiable cases of the flu came about in Haskell County, in Kansas. Eventually they spread to Fort Riley and began infecting other military bases, eventually moving into Europe by way of troop ships. As this was happening, France, Germany, and Britain had war censors in effect controlling news reports while Spain did not. As a result, the blame was fixed on Spain and many still refer to the flu today as the Spanish flu.[4]

[2] Donald McNeil, "Finding a Scapegoat When Epidemics Strike," *Internet History Sourcebooks Project*, August 31, 2009, https://sourcebooks.fordham.edu/jewish/1348-jewsblackdeath.asp.
[3] The Black Death and the Jews.
[4] McNeil.

Fast forwarding to the 1980's, although scientists now believe that HIV originally emerged in the 1930's in West Africa, it wasn't until the early 1980's that doctors became aware of HIV and AIDS.[5] The history of HIV/AIDS is intimately tied to stigma. In 1981, before the virus was isolated, researchers did not know what was causing the unusually high rates of rare pneumonia and cancer in their patients. In June of 1982, a group of cases among gay men in Southern California caused scientists to draw the conclusion that the cause of the immune deficiency only pertained to gay men and the syndrome was initially called gay-related immune deficiency (or GRID).[6] Much like the 1918 flu, the naming of the virus identified a specific identity or population. The danger here was that the name was perpetuated initially by the medical community. Eventually, later that same year, the CDC used the term 'AIDS' (acquired immune deficiency syndrome) for the first time to identify the illness. Although, it wasn't until the disease was reported in haemophiliacs and Haitians that the distinction was made.[7] Despite the discovery that the virus was not limited to gay men, the stigma persisted. The history of stigma related to AIDS/HIV is deep and still present. In fact, a Columbia University alum was published earlier this month in the guardian for being denied the opportunity to give blood as a result of his sexual orientation. Because of COVID-19, the FDA fast-tracked new guidance to ease restrictions against gay men donating blood, but most blood banks haven't yet applied the

[5] "History of HIV/AIDS," CANFAR, accessed May 12, 2020, https://canfar.com/awareness/about-hiv-aids/history-of-hiv-aids/.
[6] Ibid.
[7] Ibid.

revision and many willing individuals are still barred from donating blood. This is just one example of the lasting stigma of AIDS/HIV and its implications for gay men.[8]

Public Health crises in the age of social media face the issue of misinformation. Dissemination of scientific and medical knowledge is clouded by opinion and speculation. In the case of the 2014 Ebola outbreak in West Africa, the key role in saving the lives of infected people was played by the management and dissemination of accurate medical information. The early stages of the Ebola outbreak mirror the beginning of the COVID-19 outbreak with regard to information and facts. In the early stages of both outbreaks there was a great deal of uncertainty about the morbidity, reproductive ratio, effects of the virus, and more. While the novel public health situation needs to be assessed by health professionals, technology can't help but spread news and rumours at a rapid pace. Fear is both fueled and calmed by information. Confirmation of facts is not always immediately possible seeing as peer-reviewed literature and scientific information are lagging behind the behemoth of information that is social media. For better or for worse, social networks and mobile phones are crucial sources of information in the early stages of epidemics. With the Ebola outbreak, there were many serious complications caused by the fear and misinformation spread through social media platforms like Facebook and Twitter. When the virus was still confined to West Africa, the Iowa Department of Public Health was forced to issue a statement dispelling false rumors that the Ebola Virus had been confirmed in Iowa and had spread among its citizens. Social media posts on various platforms virulently claimed that Ebola can spread through air, water, and food— none of which correspond with the

[8] Alexandra Villarreal, "Gay Men in US Still Unable to Donate Blood despite New Coronavirus Rules," The Guardian (Guardian News and Media, April 18, 2020), https://www.theguardian.com/us-news/2020/apr/18/us-blood-donation-gay-men-coronavirus.

scientific information on the subject.[9] The spread of misinformation on social media has been likened to a computer virus, where the misinformation continues to spread and infect other people who then infect others. Social media platforms like Twitter, Facebook, and Reddit are the vectors and the hosts of such viruses. The 2014 Ebola outbreak in West Africa highlighted the dangers of social media in an increasingly globalized and interconnected world. The spread of rumours and misinformation permeated borders and social media spread such information beyond the concerned African countries. The resulting consequences of the spread of misinformation sparked local mistrust and even outward hostility towards foreign health workers and NGOs.[10] Furthermore, one of the largest consequences of social media during this outbreak was the stigmatization that occured. Ebola survivors, health workers, and medical professionals were all subject to stigmatization as a result of rumours and false information spread through social media platforms.[11] We drew on the similarities of the Ebola outbreak in 2014 to help us make sense of the current issues of stigma and xenophobia surrounding the current COVID-19 pandemic. Studies of the Ebola outbreak taught us a very important lesson regarding social media. Outbreaks of biological viruses are not unlike the outbreaks of misinformation, rumours, and stigmatization.

Upon observing these similarities between public perceptions and responses to the Ebola epidemic in West Africa, we decided to model our project, in part, after a Ebola study using the MIT Media Cloud big data platform.

[9] Joachim Allgaier and Anna Lydia Svalastog, “The Communication Aspects of the Ebola Virus Disease Outbreak in Western Africa – Do We Need to Counter One, Two, or Many Epidemics?,” *Croatian Medical Journal* 56, no. 5 (2015): pp. 496-499, https://doi.org/10.3325/cmj.2015.56.496.
[10] Ibid.
[11] Ibid.

The study, *Digital Health Communication and Global Public Influence: A Study of the Ebola Epidemic,* responded to the frustration among scientists and health care professionals over the correlation between the massive circulation of misinformation on the internet and the global public perceptions of and responses to the 2014 Ebola epidemic. The project itself used the MIT Media Cloud, the big data platform, to analyze all news stories in English using the keyword "Ebola". Their data set pulled from all articles from the U.S. Mainstream Media, U.S. Regional Media, U.S. Political Blogs, U.S. Popular Blogs, Europe Media Monitor, and Global Voices published between July 1st, 2014 to November 17th 2014.[12] Examining these outlets allowed the researchers to understand how social network theory and models of the networked global public contributed to the communication of health and information. Using the keyword "Ebola" they were able to source 109,400 stories that met their inclusion criteria. Through their research they were able to determine that the Center for Disease Control and World Health Organization were the two media sources with the most hyperlinks directed to their respective sites.[13] On social media, public engagement globally was directed toward stories covering risks of United States domestic Ebola infections. More so than stories focused on Ebola infections in West Africa or on verified science based information.[14] They also were able to determine that the sentiments about Ebola in the media sources were reflected in the policy responses to Ebola internationally. On the whole, they determined that social media platforms of the global public have influenced the discourse, sentiment, and even the response to the Ebola epidemic.[15]

[12] Hal Roberts et al., "Digital Health Communication and Global Public Influence: A Study of the Ebola Epidemic," *Journal of Health Communication* 22, no. sup1 (2017): pp. 51-58, https://doi.org/10.1080/10810730.2016.1209598.
[13] Ibid.
[14] Ibid.
[15] Ibid.

Background:

In our project, we have chosen to focus on popular discourse and news coverage over the past few months surrounding the outbreak of the novel coronavirus (SARS-CoV-2). In particular, we have chosen to examine how the news media can amplify racist or xenophobic rhetoric among the populace at large through selective, sensationalized, and misleading reporting. The social media component of our project focuses on Reddit, a site where several different communities devoted to discussion about the coronavirus have emerged. We have chosen to limit our analysis to Reddit for several reasons. Firstly, all Reddit content is filtered by an upvote-downvote system: each registered user can vote once on a post or comment to upvote or downvote it. A post that is heavily upvoted becomes visible on the front page, where more users can see and comment on it. On the other hand, a downvoted post will quickly disappear and spark little discussion. Because of this system, it is easy to tell which posts and comments have reached the most people on the website. Secondly, Reddit features communities that represent a broad spectrum of political opinions. Reddit remains home to one of the world's largest and most vocal pro-Donald Trump forums, /r/The_Donald, which, as of April 28th, 2020, has 793,141 registered users. On the other hand, pro-Bernie Sanders subreddits like /r/SandersForPresident are also fairly large, having 504,774 subscribed users. Non-partisan political subreddits, like /r/politics, are much larger than subreddits devoted to specific ideological positions or candidates, with /r/politics having 6,021,799 subscribers. Thirdly, according to Alexa's traffic rank, Reddit is the 21st most visited website in the entire world, as well as the most visited English-language discussion based website.[16] Finally, Reddit's structure is also helpful for the

[16] "Alexa - Top Sites," accessed April 28, 2020, https://www.alexa.com/topsites.

purposes of our analysis, as the website is structured into several different communities known as subreddits. Each subreddit is devoted to the discussion of a single topic, ranging from video games to sports and cooking. Because of Reddit's separation into different subreddits, we were able to limit our analysis to communities that focused on the discussion of various aspects of the current pandemic.

We see this project as an experiment in writing the history of the present. As other historians, like Krista Sigler have noted, social media websites like Twitter "[offer] students of the past (and of the present) a tool to peek into popular mentality during large events, in particular, societal disturbances." The use of social media as a primary source in historical research is a relatively new development. Though some historians, like Adam Tooze, have recently written works of history that discuss developments from the past two decades, most historians still focus on eras where social media was not prevalent. Even historians of the present often seem reluctant to directly cite or analyze social media posts as primary sources. For instance, Tooze's *Crashed*, a book focusing on the 2008 financial crisis, discusses the trending 2015 hashtag "thisisacoup."[17] Tooze, however, does not cite any social media posts themselves, but rather provides a footnote referencing a Guardian article discussing the use and promulgation of the hashtag. Our approach to social media does not focus on individual posts, but rather attempts to make use of computational techniques to analyze the large amounts of text contained in Reddit's comment sections. In particular, we were interested in determining whether different Reddit communities focused on different aspects of the coronavirus outbreak. We chose to adopt this approach both to demonstrate the potential of computational techniques in historical analysis

[17] J. Adam Tooze, *Crashed : How a Decade of Financial Crises Changed the World* (UK: Allen Lane, 2018). Chapter 22, section IV.

and to provide a more comprehensive, data-driven analysis of popular discourse on the internet surrounding the coronavirus outbreak.

Social media can offer us an unprecedented level of understanding when it comes to social attitudes and popular discourse surrounding major historical events, as previous analyses focusing on the use of social media during the Arab Spring have proven.[18] Yet, as with any primary source, social media posts should be used with caution and evaluated in light of the fact that they represent commercial platforms that are driven by profit and designed to promote certain kinds of content over others. Social media platforms themselves shape the terms of engagement with their platform and define what kinds of content and cannot be posted, making it difficult to fully see them as accurate proxies for American public discourse. In the case of Reddit, users can post links, text posts, videos, and images, but are subject both to the rules of the subreddit that they post content on and Reddit's site-wide rules. Subreddits whose content is deemed offensive or in violation of Reddit's rules are often banned or "quarantined," meaning that they generate no ad revenue for the website and are not indexed on the website's site-wide search or visible on Reddit's front page.[19] Reddit's demographics are also far from representative of the American populace at large: according to a Pew survey conducted in 2016, Reddit users tend to be disproportionately young people with college degrees, who identify as liberal at a much higher rate (43%) than the American populace at large (24%).[20]

[18] Habibul Haque Khondker, "Role of the New Media in the Arab Spring," *Globalizations* 8, no. 5 (October 2011): 675–79, https://doi.org/10.1080/14747731.2011.621287.

[19] Landoflobsters, "Revamping the Quarantine Function," reddit, accessed April 29, 2020, https://www.reddit.com/r/announcements/comments/9jf8nh/revamping_the_quarantine_function/.

[20] Michael Bartel et al. "Reddit News Users More Likely to Be Male, Young and Digital in Their News Preferences," *Pew Research Center's Journalism Project* (blog), February 25, 2016, https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/.

Despite these issues, we still believe that analyzing comments on Reddit can allow us to understand how different web communities contribute towards the development of different, ideologically shaped discourses on a major historical event. Though the site might be dominated by "liberal" users, the fact that Reddit contains a wide variety of subreddit promoting different ideological viewpoints ensures that left-wing perspectives are not dominant across the entire site, a fact that we hope to leverage in our analysis. In particular, we are interested in looking at and comparing four subreddits with different ideological perspectives on the coronavirus. The first is /r/Coronavirus, the subreddit that has been promoted by Reddit on its front page as the main hub of coronavirus-related discussion. /r/Coronavirus does not explicitly promote any ideological perspective, and because of its promotion as the "default" coronavirus subreddit by the website, it is likely that its demographics are reflective of the subreddit's as a whole, tilting towards the left, but encompassing a userbase with a wide range of ideological views. The second is /r/China_Flu, a subreddit that tends to attract a right-wing userbase because of its name. The third is /r/Lockdown_Skepticism, another subreddit that attracts a broadly right-wing userbase because of its declared purpose to provide a discussion forum to lockdown "skeptics". The final subreddit we focus on is /r/Covid19, where users are only allowed to share and discuss articles and other scholarship surrounding the coronavirus outbreak, and political discussion is banned. We decided to focus on these four subreddits because, with the exception of /r/LockdownSkepticism, they represent some of the largest non-regional subreddits devoted to discussion of the coronavirus outbreak. We decided to incorporate /r/LockdownSkepticism over larger subreddits because we were curious to see what kinds of rhetoric lockdown skeptics used to justify their positions, particularly in light of recent media attention given to anti-lockdown protests in several states.

Methodology:

To analyze recent Reddit comments, I (Daniel) used my Reddit account to access the Reddit API by creating a new Reddit "app." After obtaining a client ID from Reddit by doing so, I created a Python program and imported the Praw library, a Python wrapper for the Reddit API that allowed me to access posts, comments, user information, and other data publicly provided by Reddit. Though I had initially hoped to perform sentiment analysis on large bodies of text drawn from a number of different comment sections, I discovered that the techniques necessary for robust sentiment analysis were unfortunately beyond my abilities to implement, and opted simply to track two other measures: subreddit similarity, as measured by cosine score, and the most frequently used words on a subreddit over time. To measure both, I mined a large number of comment sections to various posts: an individual comment section varies greatly in length depending on how many comments are made, but the posts I looked at generally contained at least hundreds, if not thousands of comments. In terms of total character count, one comment section I found (with hidden, low-scoring comments not included) contained 53,960 characters, with 2425 comments in total.[21]

I looked at different numbers of posts depending on the parameters that I used for each query: for example, certain subreddits were not as active in early March, and had fewer top posts from that time than from later April. My general method was to look through the 200 "top-rated" (most upvoted, and thus most popular) posts of all time on each subreddit, and then include ones

[21] I refer to the following post: "All 11 Coronavirus Patients in Greenland Have Now Recovered. It Is Now the Only Affected Country/Territory with No Active Cases.," reddit, accessed May 4, 2020, https://www.reddit.com/r/Coronavirus/comments/fxdwoa/all_11_coronavirus_patients_in_greenland_have_now/.

with timestamps in the relevant range in my analysis. Thus, I might have looked at 20 posts in total for early March, and 30 for late April in my analysis. However, in my most comprehensive analysis, spanning top posts for all of last year, I looked at the 50 top-rated posts of all-time from four different subreddits. 50 posts and their complete comment sections seemed to represent a fairly representative cross-section of general discussion on any particular subreddit: though analyses I ran on the top 20 posts tended to be somewhat skewed by the specific content of the top posts (for example, one analysis I ran, on /r/China_Flu included several references to "tedros" because there were a large number of highly upvoted anti-WHO posts on the subreddit), looking at the top 50 posts seemed to produce results that were similar to analyses run on larger sample sizes. Expanding the sample size beyond this point seemed to generate diminishing returns: an average analysis of 50 comments already took roughly 20 minutes, and obtained fairly similar results to an analysis run on 75 posts, which took about half an hour to run on all four subreddits. To provide a point of comparison for my analysis of the top posts of all time, I also conducted two analyses of the 20 "hot" (trending) posts for a particular day, on April 29th and May 3rd. I chose to only look at 20 posts for this final analysis because that is roughly the number of posts displayed on the "front page" of any subreddit.

To find the thirty most common words or bigrams on a subreddit, I took my string, made up of the collected comments from a single subreddit, and removed punctuation and converted all letters to lowercase to ensure that capitalization and apostrophes, dashes, periods, or other characters would not confuse my analysis. I then removed stopwords, using the list of stopwords provided by the NLTK library[22] and adding "http," "https", and Reddit-specific terms like "post,"

[22] See this Github link for the complete list of NLTK stopwords. 262588213843476, "NLTK's List of English Stopwords," Gist, accessed May 4, 2020, https://gist.github.com/sebleier/554280.

"ban," or "moderator" to it. I also used the NLTK method to identify the most common words in a string (nltk.FreqDist()). The most commonly used terms on the four subreddits I looked at tended to be very similar, as they were all broadly focused on discussing the Coronavirus, and often included internet slang and other phrases common on Reddit that did not tell me much about the unique state of discourse on each subreddit. Thus, I decided to compare each list of 30 unique words to four others: the three other coronavirus subreddits, and /r/AskReddit (a general discussion forum where Reddit users can ask and obtain answers to any questions) to filter out general Reddit-specific terms on all subreddits.[23] I then deliberately excluded common words that the subreddit in question shared with the other subreddits from my final analysis. The final result was to reveal the unique most commonly used words on each subreddit, allowing me a better understanding of how discourse on each community differed.

My cosine similarity analysis was somewhat simpler: I simply took the top-50 analysis that I discussed before and compared the strings created from the comment sections of each subreddit against each other. Cosine similarity works by evaluating different strings as vectors of various features, which are then compared against each other by measuring the "angle" between the two. It is a relatively low-cost and simple way to judge difference between large bodies of text without relying on more advanced techniques like the use of neural networks (word2vec is particularly popular for similarity calculations), which is why I opted for its use in my project. Because I was only interested in judging relative similarity, rather than obtaining a very accurate measure of similarity for commercial or professional purposes, I believed that the use of cosine similarity was sufficient for my analysis.

[23] Such terms included "karma," the total balance of positive and negative votes that a user has received on their posts, "OP" (an abbreviation for original poster) and "repost," referring to content that has been posted already by another user.

Over the course of the project, I ran up against several limitations of the Reddit API. Most notably, Reddit prohibits third-party users from scraping data from “quarantined” subreddits, which generally contain controversial or offensive content. Though I was initially hoping to include the “uncensored” far-right subreddit /r/Wuhan_Flu in my analysis, I was not able to, as it is now quarantined. In addition, the API only allows users to sort posts using Reddit’s predefined criteria. I could use the API to find “top” posts of the past day, week, month, year, or of all time; look at “hot” posts, new posts, or “controversial” posts (posts that have a high downvote to upvote ratio). However, I could not find trending or top-voted posts from a particular time range through the Reddit API. Thus, it was difficult to do a representative analysis of posts from a particular period of time in the past on Reddit. However, because each post still has a timestamp attached to it, I was able to use the “top of all time” sorting to find highly voted posts that came from a particular timespan.

In the end, I opted to use timespans of roughly 15 days from March to April for my analysis. I had initially planned to do a week by week analysis, but my method of looking through the top 200 posts turned up too few entries for the earlier weeks, when the subreddits I was examining had been far less active. Thus, I extended my timespan to a period of roughly two weeks. I also limited my change over time analysis to /r/Coronavirus and /r/china_flu, as /r/LockdownSkepticism was only created a few weeks ago, making it impossible to conduct any kind of long-term analysis of discussion on the subreddit. In addition, /r/Covid19 explicitly bans any discussion of politics, making it difficult to see any kind of noteworthy development in discourse on the subreddit, which is focused purely on posting and discussing scientific research.

Though many of my methodological choices were shaped by the limitations of the Reddit API, I still found some interesting trends and developments, which I will share below.

The code used to analyze news media is included in the file COVIDNewsMedia.ipynb, executable using Jupyter Notebook. A static version, COVIDNewsMedia.html, can be viewed in a web browser. This research made use of the Media Cloud project, developed by the MIT Center for Civic Media and the Berkman Klein Center for Internet & Society at Harvard University. I interfaced with the Python client library built on top of Media Cloud's core engine and web service.

Media Cloud's engine crawls the web to collect millions of news stories published online. The tools also offer many functionalities for aggregating, analyzing, and delivering this information, of which I primarily relied upon the categorization of stories into political collections and the random sampling service to obtain most frequently used terms. The political categorization is performed based upon the "relative number of tweets of election story urls by [Hilary] Clinton vs. [Donald] Trump retweeters" during the 2016 election period.[24] Sources are delegated into static quintiles – Left, Center Left, Center, Center Right, and Right – based on this metric. This categorization method is suboptimal for several reasons. Media Cloud notes that "the set [of sources] has low certainty for individual members and so should be used only for aggregate questions rather than the labeling of individual members." Further investigation should also take into account the limited nature of the Clinton-Trump political spectrum, ideally determining a news source's political leaning based on the political values, not the political candidate, they endorse. As an interim fix, the current categorizations should be considered

[24] Media Cloud. "Source Manager: Left, Collection #9360520." Accessed May 11, 2020. https://sources.mediacloud.org/#/collections/9360520.

according to their limitations: “Pro-Clinton” rather than “Left” and “Pro-Trump” rather than “Right,” etc. The other key feature used in this analysis was the word counting service, which randomly samples a set of sentences in a given collection and returns the frequency counts of the most common words.[25] All requests assumed 5000 sentences to constitute a representative sample. The service removes a standard list of 176 stopwords from its results[26] and stems words prior to counting them. Words with the same stem are counted together, and the full term associated with the count is the word with the highest individual count.[27]

To mirror Daniel’s Reddit analysis as closely as possible, this component of our study employed similar techniques to identify the most frequently used words across news media. The top ten words in stories which include the word “coronavirus” were obtained over the same segments of time (early March, late March, etc.), using political collections instead of subreddits as sub-categories. Because news stories are not intrinsically divided by topic, like posts in subreddits are, mandating the mention of “coronavirus” was a makeshift way to attain only stories about the pandemic. However, considering that even the name of the virus is quite politicized, this approach could skew results. Future research should also search by other words that indicate topical relevance (such as “covid,” “pandemic,” “outbreak,” and “flu”). For our purposes, these topical words tended to cloud more interesting findings, so they were limited by removing from each list of top words the terms found among the top fifteen words in all the

[25] Berkman Klein Center for Internet & Society, “API 2.0 Spec” in *Media Cloud*, (2020), GitHub repository. Accessed May 11, 2020, https://github.com/berkmancenter/mediacloud/blob/master/doc/api_2_0_spec/api_2_0_spec.md#word-counting

[26] Roberts, Hal. “Overview of Media Cloud Methods.” *Mediacloud Blog* (blog). Accessed May 11, 2020. https://blogs.harvard.edu/mediacloud2/2011/05/01/overview-of-media-cloud-methods/.

[27] Berkman Klein Center for Internet & Society, “API 2.0 Spec” in *Media Cloud*, (2020), GitHub repository. Accessed May 11, 2020, https://github.com/berkmancenter/mediacloud/blob/master/doc/api_2_0_spec/api_2_0_spec.md#word-counting

political categories. The same approach was applied to determine the most frequently used bigrams by adding “ngram_count=2” as a parameter in the word counting service request (the default value accommodates single words).

Results:

One of the main areas of interest in our project was to examine news and social media discussions about China. We found some surprising results in our subreddit analysis of terms and bigrams related to China. In particular, China never seemed to be an especially dominant topic of discussion on /r/Coronavirus at any time. Though several posts critical of the Chinese government’s handling of the pandemic were heavily upvoted on the subreddit, many other posts that discussed other states’ responses to the coronavirus were also among the top voted posts of all time. In general, the most-used bigrams on the subreddit had to do with relatively mundane concerns, like toilet paper shortages in late March or the difficulties of grocery shopping during the pandemic. Our findings from /r/Coronavirus seemed to support the idea that the subreddit serves mostly as a nonpartisan, general source of discussion about the coronavirus pandemic for Anglophone internet users across the world.

On the other hand, r/China_Flu’s narrow focus on China’s response to the pandemic became more intense over time. In early to mid-March, China was only one topic of discussion among many, with “masks” being the most popular term on the subreddit and “US” being the third-most used after “China.” By mid-to-late April, most of the ten most-used terms on the subreddit had to do with China, including “Wuhan”, “ccp”, “china, and “chinese.” The subreddit took a more openly bigoted turn in early April, with terms like “civilized world” coming into common usage on the subreddit, usually in reference to China’s “uncivilized” status. We believe that a major turning point for the subreddit was when Trump used the “Chinese virus” label to describe the coronavirus in mid-March, making the “China flu” name (already a stigmatizing label) explicitly linked with Trump’s brand of politics. Another major incident that shaped the development of the subreddit was when reddit.com started to promote /r/Coronavirus as the “official” source of news on the Coronavirus outbreak as opposed to /r/China_Flu.[28] These two factors eventually resulted in the transformation of /r/China_Flu from a general news subreddit on the coronavirus pandemic to a community whose users frequently made use of bigoted rhetoric and became singlemindedly focused on discussing China’s culpability in spreading the virus.

None of the subreddits we researched featured racial slurs or other offensive terms among the most-used terms. However, this is likely not because of a lack of racial prejudice on Reddit or the subreddits we studied, but rather the fact that the rules of all the subreddits we studied, including even /r/China_Flu, explicitly ban bigotry or offensive content.[29] Openly racist posts are likely downvoted, reported, or deleted before they can gain much traction. Thus, it seems more likely that racist or xenophobic sentiment is disguised with “dog whistles,” including remarks that paint China or other East Asian states as dirty, uncivilized, or obsessed with the consumption of exotic animals. The evidence we uncovered supports this conclusion: terms like “civilized world” and “Wuhan virus” started to appear frequently in /r/China_Flu throughout April, and references to China’s “wet markets” also became increasingly frequent around that time. More openly racist rhetoric surrounding the coronavirus has been analyzed in greater detail

[28] Steven Asarch, “Reddit Has Started Promoting r/CoronaVirus to Combat the Spread of Misinformation on the Platform,” Newsweek, March 3, 2020, https://www.newsweek.com/reddit-fight-coronavirus-misinformation-subreddit-best-news-source-1490292.

[29] “COVID-19 Discussion,” accessed May 4, 2020, https://www.reddit.com/r/China_Flu/.

in publications focusing on other social media forums, like 4chan's infamous far-right /pol/ board.[30] However, despite some incidents like the one documented above, our findings seem to indicate Reddit users generally appear to shy away from open bigotry in favor of coded Sinophobic rhetoric that associates Chinese people with disease, barbarism, and uncleanliness.

Interestingly, /r/Coronavirus seemed to be the only subreddit that was interested in discussing the economic or work-related impacts of the virus: the most-used word by far of all time on /r/Coronavirus was "work," while frequently used bigrams included "grocery stores" and "work [at] home" (because of our stopword exclusion, we cut out the "at" from our bigram). Discussion of employment data, work, or any terms that would have had to do with the economic recession were absent among the thirty-most used words and bigrams on /r/LockdownSkepticism, which we initially believed would have brought up the economic toll of the pandemic more frequently in discussion. Though the most-used words of all time are not particularly illustrative for /r/LockdownSkepticism, we can find some clues as to popular discussion topics in the trending posts from May 4th: a popular term (used 74 times in total among the comment sections for the 20 top posts) was "social," perhaps indicating that many on the subreddit seemed more concerned about the potentially detrimental effects of lockdowns for social interaction than economic or public health consequences. Another popular term from the trending posts on April 29th was "kids," possibly referencing a popular point among lockdown skeptics that school shutdowns were unnecessary because of the low death rate among children, or the broader desire of lockdown skeptics to immediately reopen schools and other educational institutions. We can also look at the bigrams for /r/LockdownSkepticism to find some interesting

[30] Leonard Schild et al., "'Go Eat a Bat, Chang!': An Early Look on the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19," *ArXiv:2004.04046 [Cs]*, April 8, 2020, http://arxiv.org/abs/2004.04046.

trends: there are several references to “virtue signaling” and “social media,” indicating the belief of users on the subreddit that the impacts of the coronavirus have been sensationalized on the Internet by people seeking to “virtue signal” by demonstrating their compliance with lockdown orders. Finally, the call to “go outside” appeared on the list of most commonly used bigrams for /r/LockdownSkepticism, demonstrating the open disregard of the users on the subreddit for stay-at-home orders.

More predictably, both /r/LockdownSkepticism and /r/Coronavirus seemed to be interested in discussing the implications of stay-at-home orders: “home” was the fourth most-used work on /r/LockdownSkepticism, with more than 4000 unique uses, and “home” was also consistently in the list of most-used words on /r/Coronavirus after late March (when the first lockdown orders were issued in nine American states).[31] However, “home” appears along with “stay” in the late March list of most commonly used words on /r/Coronavirus, while it did not with /r/LockdownSkepticism at any point. By looking at commonly used bigrams, we determined that “stay home” was one of the most-used terms on /r/Coronavirus in early April, but almost never used on /r/LockdownSkepticism, a result in line with the general stances of users on both subreddits towards stay-at-home orders. Finally, the results we gathered from /r/covid19 also showed little beyond what we expected: references to “mask,” “vitamin,” “study,” “test,” and other terms related to scientific or medical research on the coronavirus were frequent throughout the subreddit’s history.

Our study of cosine similarity yielded somewhat predictable results: in general, /r/covid19 had the least similarity to other subreddits, likely because of its ban on political

[31] Sarah Mervosh, Denise Lu, and Vanessa Swales, “See Which States and Cities Have Told Residents to Stay at Home,” *The New York Times*, March 31, 2020, sec. U.S., https://www.nytimes.com/interactive/2020/us/coronavirus-stay-at-home-order.html.

discussion and its focus on discussing scientific research over other aspects of the pandemic. The two subreddits that were most similar according to our measures were /r/coronavirus and /r/lockdownskepticism. This could be because, as noted above, /r/coronavirus and /r/lockdownskepticism users were both generally interested in discussing topics like the impact of stay-at-home orders and lockdowns over users on other subreddits. There was also a high degree of similarity between /r/coronavirus and /r/china_flu and between /r/lockdownskepticism and /r/china_flu. The former can be explained by the fact that news stories about China's response to the pandemic tended to be popular on the /r/Coronavirus subreddit, but not to the same degree that they were on /r/china_flu. The latter seems more difficult to explain, as I could not find a single post from the top 50 on /r/lockdownskepticism that discussed China's response to the virus. However, it might be explained by the fact that the two subreddits share a similar, generally right-wing userbase, given the nature of the causes or news stories that they tend to promote.

Perhaps most surprisingly, our similarity metric deemed the language used on /r/covid19 most similar to that used on /r/lockdownskepticism. This can be explained by the fact that /r/LockdownSkepticism counts several misinterpreted, dubious, or controversial studies among its top posts of all time. For example, the 24th top-rated post of all time on the subreddit (as of May 6th, 2020) is titled "Researchers in Austria concluded that more people died there in March from untreated heart attacks than from Covid19."[32] The report it links to from the *European Heart Journal* notes a decline in hospitalizations for ACS because of strict lockdown measures,

[32] "Researchers in Austria Concluded That More People Died There in March from Untreated Heart Attacks than from Covid19," reddit, accessed May 6, 2020, https://www.reddit.com/r/LockdownSkepticism/comments/ga8gjo/researchers_in_austria_concluded_that_more_people/.

which the authors nonetheless argue are “undoubtedly critical for controlling the COVID-19 pandemic.” It does not claim that more deaths resulted from ACS than COVID-19, though it claims that such an outcome might have been possible based on data from the 1950s.[33] In general, tendentious claims based on selective interpretation of scientific studies are extremely popular on the /r/LockdownSkepticism subreddit, a fact that accounts for its apparently high degree of lexical similarity to /r/covid19.

These findings reveal how cosine similarity is an imperfect, albeit sometimes useful metric. Though cosine similarity and other computational techniques that analyze the similarity of two texts might represent a good starting point for further research, they should not be taken as an objective measure of the degree of “similarity” (however we define the term) between different subreddits. Even a perfect implementation of cosine implementation would be a superficial measure of the actual degree of similarity between two bodies of text. For instance, if we measure the “similarity” of texts purely through their degree of lexical similarity, we might find that two texts that discuss lockdowns, with one totally opposed and the other supportive, would be highly similar. On one level, this is a valid conclusion: both texts would discuss the same topic, even if they took an opposed stance to it. However, measures of lexical similarity cannot identify the crucial fact that they represent opposing viewpoints on the same issue.

In contrast to social media per Daniel’s Reddit analysis, news media was intensely occupied with the topic of China. The term “China” consistently appears in the top ten terms across political collections and time segmentations. Other China-related terms, such as “Wuhan” and “Chinese,” consistently appeared only in the Center Right and Right collections across time

[33] Bernhard Metzler et al., “Decline of Acute Coronary Syndrome Admissions in Austria since the Outbreak of COVID-19: The Pandemic Response Causes Cardiac Collateral Damage,” *European Heart Journal*, accessed May 5, 2020, https://doi.org/10.1093/eurheartj/ehaa314.

segmentations. Figures 1, 2, and 3 exhibit graphs of word usage for "China," "Chinese," and "Wuhan" across the study's timespan, segmented by political categorization. These provide a visualization of how usage of each of these terms was distributed across the political spectrum.

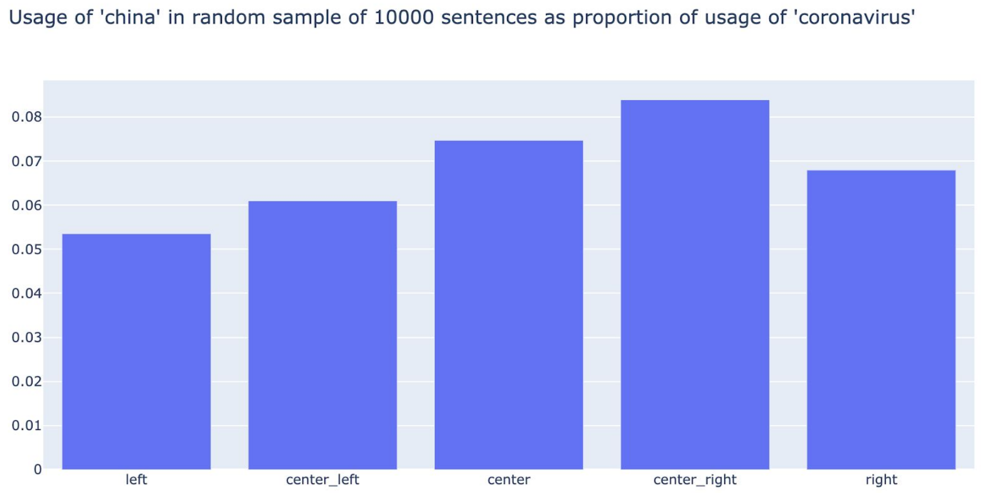

Figure 1: Usage of 'china' in a random sample of 10,000 sentences as a proportion of usage of 'coronavirus'

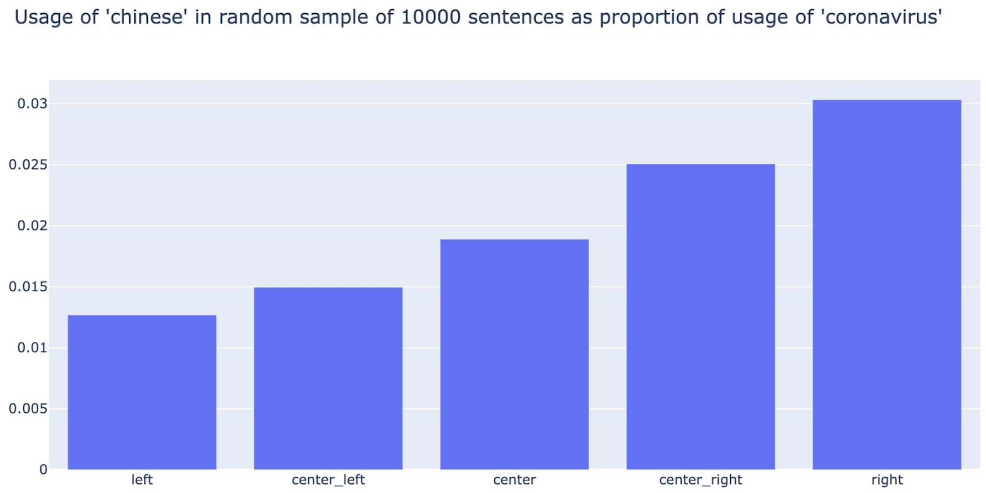

Figure 2: Usage of 'chinese' in a random sample of 10,000 sentences as a proportion of usage of 'coronavirus'

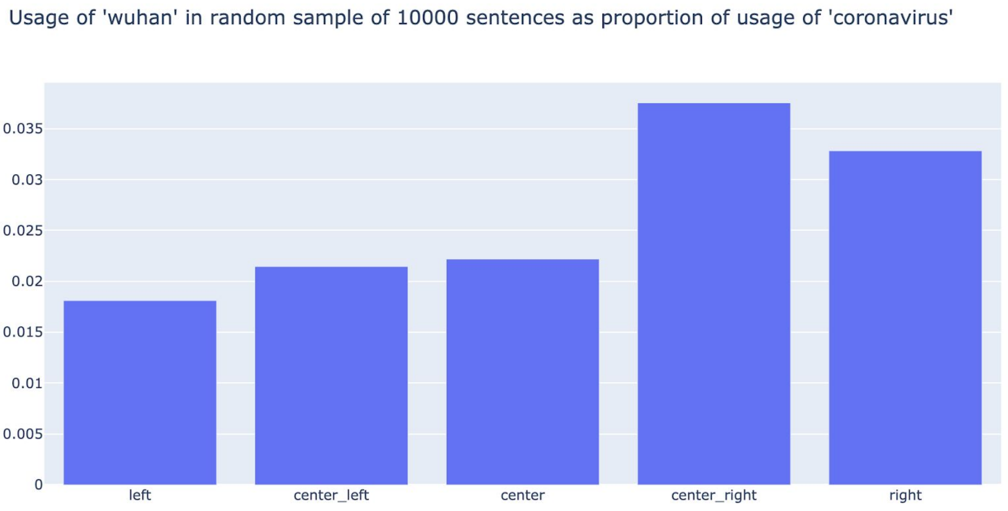

Figure 3: Usage of ‘wuhan’ in a random sample of 10,000 sentences as a proportion of usage of ‘coronavirus’

While usage of ‘china’ and ‘wuhan’ was more evenly distributed throughout the political spectrum, usage of ‘chinese’ consistently increases proportionally to right-wing categorization. One theory behind this, which is supported by the rise of referring to COVID as the “Chinese virus” in Trump-supporting subreddits, is that ‘china’ and ‘wuhan’ are used in less polarized contexts, while ‘chinese’ is more frequently used to stigmatize people of Chinese race or nationality. Further research should explore which phrases frequently surround these terms to better understand why and in which contexts news sources use them.

Also paralleling the results of social media analysis on Reddit, economic concerns were universally outspoken across the political spectrum, as demonstrated in Figure 4, .

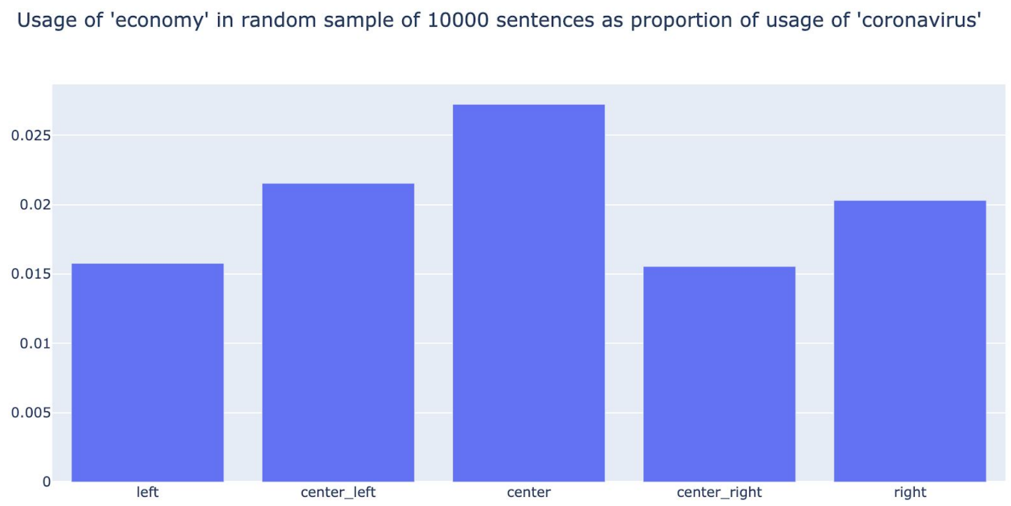

Figure 4: Usage of ‘economy’ in a random sample of 10,000 sentences as a proportion of usage of ‘coronavirus’

There were several other unusual results that warrant further exploration and interrogation of our research methods. Our analysis made it clear that Media Cloud’s political collections are not restricted to English-language new stories. For example, the top bigrams in the Left category are frequently Spanish terms. Even more strangely, these terms, such as “*suscriptor generar*” (“generate subscriber”), “*generar contraseña*” (“generate password”), and “*contraseña nueva*” (“new password”) , seemed to come from non-news-story web content. Similarly, the top Center Right bigrams were often French, such as “*lien langue*” (“discussion link”) and “*lire ligne*” (“read line”) and in some cases German, such as “*abgerufen märz*” (“accessed march”). This issue originated in Media Cloud’s mining or categorization process and likely skewed results due to our focus on English in all language processing. For this reason, the bigram results are not taken to be representative of media trends in respective political collections.

Reflections:

As we noted in our introduction, we viewed our project as an experiment of sorts in writing the history of the present using social media as a primary source. We believe that future historians dealing with the history of the early 21st century will have to engage in analysis of social media, particularly if their interests lie in cultural or social history. However, the analysis of social media differs substantially from previous methods of historical research: the "archive" we deal with is often not primarily textual in nature, contains far more records than a single person or even a large team of people could ever hope to evaluate, and is often not even structured to facilitate effective preservation of the records it contains. Furthermore, rather than being compiled by private organizations, corporations, or government institutions, the social media "archive" is constructed by the actions of its users, which are nonetheless conditioned by the terms, structure, and practices of particular social media platforms. Though studying social media as a primary source might help us understand major cultural and social developments of a particular era, it is difficult, in particular, to identify any representative trends or patterns from a patchwork analysis of individual posts, no matter how popular or influential those posts might be. Thus, computational techniques, particularly those associated with natural language processing, can be helpful, if not essential, for the historical study of social media.

Though we felt satisfied with the final results of our project, and believed that our use of computational techniques allowed us to detect trends and patterns that we would have noticed otherwise, we were also made acutely aware of several issues that almost all digital historians confront. In particular, we discovered that it was extremely difficult to reproduce certain aspects of our analysis because of the limitations of the Reddit API, as discussed in our methodology section. Though our program is able to record and analyze data from individual subreddits at the moment when it runs, it cannot repeat the same analysis a second time, as the front page of every page on Reddit is constantly shifting and the website does not maintain a record of trending posts from particular moments in its past. Thus, it is impossible for other scholars to verify that all of the data we have provided is accurate. However, some of our findings were reproducible: in particular, we ran analyses of posts in the top 200 with particular timestamps and achieved identical results each time.

This particular issue, however, reveals a deeper problem we had to confront: that social media websites are not designed in a way that makes study of older content posted on them feasible, even if they make efforts to preserve those posts. Reddit, like Facebook or Twitter, is explicitly designed to provide an unending stream of new, fresh content to its users, not to allow a user to study the evolution of discourse on any particular subreddit or the website as a whole over time. The website itself, like other social media sites, does not provide a "snapshot" of particular moments in its past, assuming that ordinary users would be more interested in seeing new stories, images, and videos than looking through ones they might have missed. It also does not provide the functionality to search through posts by their timestamp, meaning that any user who wants to look at posts from a particular time range has to rely on a series of workarounds to do so. Though other sites like the Wayback Machine attempt to catalogue the state of popular internet websites over time, they provide a limited picture of the state of a particular website at different points in the past, and one that is not crawlable using its API. While digital data may never be lost, it can be difficult for historians to uncover and reconstruct because of the very nature of social media platforms, which are not designed to allow robust study of their past.

The sheer volume of data on Reddit also proved to be a problem. We were forced to limit our analysis to relatively small samples of posts on a subreddit rather than performing a comprehensive analysis on extremely large quantities of data because of the limitations of the hardware we ran our tests on. It proved impossible for us to process the massive volume of text submitted on a daily basis to reddit.com in a reasonable amount of time with the laptops and personal computers at our disposal. Like scholars in other social science disciplines, we opted to circumvent this problem by finding representative samples of the state of discourse on individual subreddits: besides simply looking at comment sections to top-rated posts, we also looked at comment sections to a limited number of trending posts from particular days and particular timespans. Though we were ultimately satisfied with the results of this approach, our original aim was to conduct a far more comprehensive analysis. This proved impossible with the consumer-grade hardware available to us, but might even be difficult for hardware designed specifically for scientific applications. If feasible, however, we would be interested in doing a similar study in the future using Columbia's scientific computing cluster to analyze all posts from different subreddits in smaller timespans (perhaps a few days to a week). In any case, for many scholars, access to such hardware would often require substantial funding and institutional support, posing another barrier towards conducting digital history for many history departments in the United States today.

Finally, it remains difficult to draw any conclusions about the general attitudes of Americans or English-speaking internet users towards the pandemic over time purely from studying one website. We already noted that Reddit's userbase is not broadly reflective of the American public, featuring a much higher proportion of young and left-leaning users compared to the demographics of the United States as a whole. However, the format of Reddit itself also encourages different kinds of posts compared to other social media platforms: for example, users are often implicitly encouraged to post uncontroversial, yet entertaining content to achieve a large number of upvotes and gain "karma" on the website. Certain communities also ban or (through the downvote system) heavily discourage any content that violates their rules or promotes a certain political perspective opposed to the one that generally dominates discourse on the subreddit. Thus, like other social media networks, Reddit has been accused of creating "echo chambers"[34] where particular political viewpoints are amplified and radicalized by the fact that they go unchallenged by opposing perspectives. However, other social media networks are also structured very differently than Reddit, and promote very different types of content and discourse through their rules, conditions, and unwritten norms. Thus, it is very difficult to generalize discourse on any particular website to the Internet as a whole, much less American society. Because of this, we believe that further research on other social media platforms, like Twitter, Facebook, or Yelp are needed to draw any definite conclusions about how Americans, or English-speaking Internet users as a whole, are discussing the coronavirus.

In the end, we felt that our research on Reddit was an enlightening experience, albeit one that made us painfully aware of the limitations that might be placed on history in the future. In particular, we are concerned that a increasingly large portion of public discourse and cultural memory surrounding major events now only exists in digital form, where it lies in the hands of private corporations who often have little interest in preserving past content on the websites that they manage. There is a major risk that much digital media will never be preserved at all once

[34] Pascal Jürgens and Birgit Stark, "The Power of Default on Reddit: A General Model to Measure the Influence of Information Intermediaries," *Policy & Internet* 9, no. 4 (2017): 395–419, https://doi.org/10.1002/poi3.166.

the servers they are hosted on are retired from service for various reasons. In addition, while the sheer quantity of data present on social media represents a major opportunity for historians, it can only be comprehensively analyzed through the use of advanced computational techniques, which often require access to high-performance computing hardware. Thus, historians who lack institutional support and funding might have a difficult time incorporating computational methods into their research.

The computational methods utilized in this study also equipped us to analyze an immense corpus of news media as a primary source, a project that would not be possible without digital methods. However, this study has exposed some shortcomings of purely computational methods; lingering research questions demand qualitative explanation for the quantitative results we achieved. These qualitative questions, for example, why the disparity between usage of 'china' and 'chinese' straddles party lines, warrant both deeper computational investigation (In what contexts are these terms used?) and broader, more traditional historical methods. A feature to make Media Cloud more useful for digital historians could combine insight into headlines or sentiment analysis with these big-picture n-gram counts. In future research, computational methods more advanced than n-grams (topic modeling, more complex visualizations, etc.) alongside contextual historical analysis will yield a more nuanced and thorough narrative.

It seems clear to us now that histories of our present moment will take a very different form than past historical scholarship: much like what occurred during the print revolution of the early modern era, the present has seen a massive shift in the way that humans record and store records of their time. Still, we retain faith in the ability of historians to adjust to the new demands of historical writing and scholarship. Even if historians of periods after the digital revolution embrace wholly different techniques than ancient or modern historians, they will still be engaged in the same intellectual project as their peers of understanding and analyzing the past. Like other historians, they will eventually discover how to use the best sources and methods available to them. We hope to have made a worthy contribution to this process of adaptation through this report, and we look forward to developing our research further in the future.

Works Cited

Asarch, Steven. “Reddit Has Started Promoting r/CoronaVirus to Combat the Spread of Misinformation on the Platform.” Newsweek, March 3, 2020. https://www.newsweek.com/reddit-fight-coronavirus-misinformation-subreddit-best-news-source-1490292.

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. “Echo Chambers on Social Media: A Comparative Analysis.” *ArXiv:2004.09603 [Physics]*, April 20, 2020. http://arxiv.org/abs/2004.09603.

Gault, Matthew. “Reddit Can’t Quarantine Coronavirus Misinformation.” *Vice* (blog), February 27, 2020. https://www.vice.com/en_us/article/n7jqbx/reddit-cant-quarantine-coronavirus-misinformation.

Isenstadt, Alex. “GOP Memo Urges Anti-China Assault over Coronavirus.” POLITICO. Accessed May 3, 2020. https://www.politico.com/news/2020/04/24/gop-memo-anti-china-coronavirus-207244.

Khondker, HabibulHaque. “Role of the New Media in the Arab Spring.” *Globalizations* 8, no. 5 (October 2011): 675–79. https://doi.org/10.1080/14747731.2011.621287.

Mervosh, Sarah, Denise Lu, and Vanessa Swales. “See Which States and Cities Have Told Residents to Stay at Home.” *The New York Times*, March 31, 2020, sec. U.S. https://www.nytimes.com/interactive/2020/us/coronavirus-stay-at-home-order.html.

Metzler, Bernhard, Peter Siostrzonek, Ronald K. Binder, Axel Bauer, and Sebastian Johannes Reinstadler. “Decline of Acute Coronary Syndrome Admissions in Austria since the Outbreak of COVID-19: The Pandemic Response Causes Cardiac Collateral Damage.” *European Heart Journal*. Accessed May 5, 2020. https://doi.org/10.1093/eurheartj/ehaa314.

Myers, Cayce, and James F. Hamilton. “Social Media as Primary Source.” *Media History* 20, no. 4 (October 2, 2014): 431–44. https://doi.org/10.1080/13688804.2014.950639.

Pew Research Center’s Journalism Project. “Reddit News Users More Likely to Be Male, Young and Digital in Their News Preferences,” February 25, 2016. https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/.

reddit. “Researchers in Austria Concluded That More People Died There in March from Untreated Heart Attacks than from Covid19.” Accessed May 6, 2020.

https://www.reddit.com/r/LockdownSkepticism/comments/ga8gjo/researchers_in_austria_concluded_that_more_people/.

reddit. “Revamping the Quarantine Function.” Accessed April 29, 2020. https://www.reddit.com/r/announcements/comments/9jf8nh/revamping_the_quarantine_function/.

Roberts, Hal, Brittany Seymour, Sands Alden Fish, Emily Robinson, and Ethan Zuckerman. “Digital Health Communication and Global Public Influence: A Study of the Ebola Epidemic.” *Journal of Health Communication* 22, no. sup1 (2017): 51–58. https://doi.org/10.1080/10810730.2016.1209598.

Schild, Leonard, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. “‘Go Eat a Bat, Chang!’: An Early Look on the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19.” *ArXiv:2004.04046 [Cs]*, April 8, 2020. http://arxiv.org/abs/2004.04046.

“Teaching Twitter: The History of the Present | Perspectives on History | AHA.” Accessed April 28, 2020. https://www.historians.org/publications-and-directories/perspectives-on-history/april-2011/teaching-twitter-the-history-of-the-present.

Timberg, Craig, and Allyson Chiu. “As the Coronavirus Spreads, so Does Online Racism Targeting Asians, New Research Shows.” Washington Post. Accessed April 29, 2020. https://www.washingtonpost.com/technology/2020/04/08/coronavirus-spreads-so-does-online-racism-targeting-asians-new-research-shows/.

Tooze, J. Adam. *Crashed : How a Decade of Financial Crises Changed the World*. UK: Allen Lane, 2018.

Appendix: Social Media Findings:

Cosine Similarity (computed from top of all time data):

Coronavirus and Covid19 similarity: 0.37277900052262664

Coronavirus and chinaflu similarity: 0.4046445118947137

Chinaflu and covid19 similarity: 0.3859146704991789

covid19 and lockdown_skepticism similarity: 0.3960043143156913

coronavirus and lockdown_skepticism similarity: 0.4087513487308813

chinaflu and lockdown_skepticism similarity: 0.4056710034251283

Most-Used Words Results:

Top of All Time:

/r/Coronavirus: ('work', 6278), ('make', 4448), ('want', 3614), ('everyone', 3534),

/r/Covid19: ('get', 7560), ('virus', 6143), ('mask', 5774), ('vitamin', 5320), ('study', 4782), ('cases', 4759), ('removed', 4684), ('rate', 4551), ('use', 4499), ('deaths', 4331), ('test', 4297), ('testing', 4283), ('blood', 4078), ('patients', 4029),

/r/China_Flu: ('china', 31802), ('us', 13340), ('masks', 11172), ('chinese', 9733), ('world', 8723), ('government', 8662), ('need', 7528), ('country', 6996), ('right', 6139),

/r/LockdownSkepticism: ('lockdown', 7523), ('even', 5821), ('one', 4965), ('home', 4742), ('know', 3708), ('way', 3383), ('sub', 3300), ('take', 3062),

Early March:

/r/Coronavirus: ('please', 1459), ('need', 1406), ('coronavirus', 1373), ('good', 1326), ('home', 1281), ('fuck', 1254), ('jeff', 1185), ('removed', 1128), ('take', 1028), ('test', 1004),

/r/China_Flu: ('masks', 11399), ('china', 10840), ('know', 6747), ('government', 5513), ('go', 4919), ('mask', 4852), ('right', 4606), ('still', 4546), ('flu', 4328), ('everyone', 4275), ('italy', 3988),

Late March:

r/Coronavirus: ('good', 10772), ('home', 10247), ('right', 9522), ('everyone', 8513), ('stay', 7131),

/r/China_Flu: ('china', 76411), ('masks', 37232), ('know', 24864), ('chinese', 21962), ('government', 21799), ('need', 21413), ('still', 20596), ('world', 20419), ('go', 18828), ('cases', 17453), ('mask', 17394), ('flu', 16217), ('country', 16063),

Early April:

/r/Coronavirus: ('need', 7889), ('removed', 5903), ('good', 5827), ('still', 5636), ('right', 5633), ('home', 4767), ('hope', 4690), ('news', 4680),

/r/China_Flu: ('china', 14949), ('world', 5236), ('chinese', 5034), ('masks', 4926), ('countries', 3381), ('country', 3268), ('mask', 3257), ('ccp', 3132), ('racist', 2611), ('also', 2543), ('trump', 2422), ('government', 2284), ('everyone', 2176), ('many', 2176), ('wear', 2161),

Late April:

/r/Coronavirus: ('think', 3034), ('go', 2709), ('need', 2604), ('really', 2241), ('home', 1978), ('everyone', 1942), ('day', 1916), ('pay', 1853), ('fucking', 1829),

/r/China_Flu: ('china', 1624), ('world', 595), ('ccp', 493), ('taiwan', 481), ('chinese', 440), ('trump', 421), ('country', 333), ('wuhan', 323), ('countries', 274), ('number', 267), ('million', 251), ('right', 243), ('many', 242), ('pandemic', 233), ('news', 226), ('removed', 221), ('funding', 217),

April 29th Trending: (apologies for the lack of numbers: by the time I'd refined my program to start collecting them, the data was no longer available, as the trending posts had changed already)

/r/Coronavirus: work, going, mask, news, back, really, costco, everyone, masks, deaths, see, weeks, getting, need,

/r/Covid19: people, virus, patients, study, know, vaccine, cases, immunity, positive, vitamin, also, much, anyone, remdesivir, risk, r0, population, one, studies, countries, good, testing,

/r/China_Flu: chinese, china, europe, wuhan, million, companies, per, lot, world, 1, investment, us, take, many, acquisitions, assets, lu, european, india, right,

/r/LockdownSkepticism: post, would, lockdown, may, death, posts, approved, please, title, lockdowns, kids, data, lives, video, health, high, cause, virus, listed, primary,

May 4th Trending:

/r/Coronavirus: ('coronavirus', 752), ('please', 723), ('flu', 603), ('removed', 597), ('test', 517), ('could', 490), ('fuck', 482),

/r/Covid19: ('rate', 365), ('countries', 331), ('high', 322), ('italy', 320), ('mask', 284), ('country', 279), ('korea', 269), ('us', 263), ('testing', 249), ('good', 236), ('temperature', 236), ('data', 222), ('number', 221), ('china', 219), ('also', 213),

/r/China_Flu: ('masks', 355), ('cases', 234), ('know', 219), ('government', 178), ('need', 166), ('home', 158), ('everyone', 151), ('healthcare', 135), ('well', 130), ('day', 129),

/r/LockdownSkepticism: ('go', 149), ('think', 136), ('virus', 115), ('take', 100), ('still', 96), ('really', 87), ('live', 85), ('back', 81), ('lockdowns', 78), ('social', 74),

Most-Used Bigrams Results:

Note: I have left out several entries from /r/Coronavirus in the following entries. Those entries come from an automatic comment made in all threads by the AutoModerator, which reminds users to follow certain rules on the subreddit. In the early days of the subreddit, since there were fewer comments and posts overall, those automatic comments tended to crowd out all other possible phrases, meaning that there were few I've provided a list of those bigrams below, drawn from the "top of all time" list: you can see that they all come from the same piece of repeated text by the fact that they have the same number of mentions. I also used 100 posts for the top of all time sample instead of 200, because it seemed like a larger representative sample was needed to get representative bigrams rather than individual words.

Removed bigrams:

(('specific', 'set'), 1405), (('set', 'civil'), 1405), (('civil', 'Personal'), 1405), (('Personal', 'attacks'), 1405), (('attacks', 'accusations'), 1405), (('accusations', 'Repeated'), 1405), (('Repeated', 'offences'), 1405), (('offences', 'Avoid'), 1405), (('Avoid', 'political'), 1405), (('discussions', 'related'), 1405),

Top of All Time:

/r/Coronavirus: (('removed', 'removed'), 2822), (('grocery', 'store'), 2713), (('action', 'concerns'), 2026), (('federal', 'government'), 1677), (('get', 'sick'), 1551), (('medical', 'advice'), 1493), (('work', 'home'), 1487), (('political', 'discussions'), 1483), (('public', 'policy'), 1477), (('feeling', 'anxious'), 1455), (('coronavirus', 'outbreak'), 1433),

/r/Covid19: (('death', 'rate'), 3319), (('New', 'York'), 3180), (('underlying', 'condition'), 2728), (('immune', 'system'), 2625), (('South', 'Korea'), 2610), (('underlying', 'conditions'), 2601), (('good', 'news'), 2422), (('weeks', 'ago'), 2227), (('two', 'weeks'), 2096), (('tested', 'positive'), 2076), (('antibody', 'tests'), 2070), (('infection', 'rate'), 1961), (('blood', 'pressure'), 1961), (('confirmed', 'cases'), 1957), (('mortality', 'rate'), 1839), (('anyone', 'know'), 1603), (('surgical', 'masks'), 1601), (('York', 'City'), 1581), (('blood', 'type'), 1540), (('health', 'care'), 1516), (('wear', 'masks'), 1512), (('fatality', 'rate'), 1505), (('false', 'positives'), 1494), (('antibody', 'testing'), 1447),

/r/China_Flu: (('CHINA', 'CHINA'), 8418), (('healthcare', 'workers'), 4920), (('new', 'cases'), 4327), (('rest', 'world'), 3691), (('Hong', 'Kong'), 3564), (('wear', 'mask'), 3083), (('Chinese', 'government'), 2886), (('wearing', 'masks'), 2652), (('Chinese', 'people'), 2537), (('wearing', 'mask'), 1993), (('toilet', 'paper'), 1946), (('people', 'still'), 1924), (('United', 'States'), 1853), (('last', 'week'), 1814), (('old', 'people'), 1743), (('taking', 'seriously'), 1742),

/r/LockdownSkepticism: (('feel', 'like'), 2245), (('herd', 'immunity'), 2067), (('flatten', 'curve'), 1771), (('people', 'die'), 1630), (('people', 'want'), 1366), (('submission', 'New'), 1355), (('posts', 'team'), 1355), (('listed', 'Posts'), 1355), (('meet', 'high'), 1355), (('standards', 'approved'), 1355), (('see', 'posting'), 1355), (('guidelines', 'take'), 1355), (('number', 'hours'), 1355),

Early March:

/r/Coronavirus: (('action', 'concerns'), 322), (('coronavirus', 'outbreak'), 314),

/r/China_Flu: (('healthcare', 'workers'), 1505), (('Hong', 'Kong'), 793), (('wearing', 'mask'), 667), (('old', 'people'), 651), (('toilet', 'paper'), 623), (('12', '月'), 600), (('confirmed', 'cases'), 591), (('N95', 'masks'), 580), (('taking', 'seriously'), 540), (('hand', 'sanitizer'), 513), (('last', 'week'), 502), (('trust', 'China'), 495), (('general', 'public'), 455), (('make', 'sure'), 454), (('wearing', 'masks'), 442), (('people', 'get'), 438), (('rest', 'world'), 433), (('looks', 'like'), 430), (('2', 'weeks'), 426), (('Home', 'Depot'), 425), (('month', 'ago'), 414),

(('set', 'related'), 308), (('related', 'coronavirus'), 308), (('outbreak', 'focused'), 308), (('focused', 'politicians'), 308), (('politicians', 'rather'), 308), (('rather', 'public'), 308), (('public', 'policy'), 308), (('policy', 'discretion'), 308), (('discretion', 'repeat'), 308), (('repeat', 'reliable'), 308), (('reliable', 'sources'), 308)

Late March:

/r/Coronavirus: (('toilet', 'paper'), 2352), (('social', 'distancing'), 1627), (('action', 'concerns'), 1361), (('coronavirus', 'outbreak'), 1055), (('related', 'coronavirus'), 1053), (('public', 'policy'), 1039), (('medical', 'advice'), 1031), (('feeling', 'anxious'), 1029), (('last', 'week'), 1020),

/r/China_Flu: (('CHINA', 'CHINA'), 4968), (('new', 'cases'), 1004), (('rest', 'world'), 869), (('Chinese', 'government'), 837), (('wearing', 'masks'), 700), (('following', 'since'), 669), (('Chinese', 'people'), 616), (('Hong', 'Kong'), 598), (('everyone', 'else'), 490), (('blame', 'China'), 484), (('million', 'people'), 456), (('wet', 'markets'), 456), (('three', 'weeks'), 450), (('every', 'country'), 447), (('two', 'months'), 429), (('watching', 'since'), 420), (('around', 'world'), 406), (('3', 'weeks'), 393), (('Chinese', 'Communist'), 387),

Early April:

r/Coronavirus: (('stay', 'home'), 1161), (('hazard', 'pay'), 725), (('get', 'sick'), 645), (('going', 'get'), 586), (('people', 'dying'), 530), (('make', 'sure'), 522), (('death', 'threats'), 493), (('covid', '19'), 478), (('get', 'back'), 472), (('lot', 'people'), 464), (('good', 'news'), 447), (('pretty', 'much'), 408),

/r/China_Flu: (('wet', 'markets'), 624), (('rest', 'world'), 554), (('Chinese', 'people'), 510), (('Chinese', 'government'), 478), (('South', 'Korea'), 368), (('human', 'human'), 360), (('United', 'States'), 328), (('wet', 'market'), 323), (('healthcare', 'workers'), 313), (('Hong', 'Kong'), 298), (('common', 'sense'), 274), (('called', 'racist'), 261), (('civilized', 'world'), 248), (('made', 'China'), 235), (('whole', 'world'), 227), (('2', 'months'), 225), (('medical', 'supplies'), 225), (('China', 'would'), 221), (('people', 'wear'), 217), (('new', 'cases'), 202), (('got', 'ta'), 191), (('coming', 'China'), 188), (('western', 'countries'), 184),

Late April:

/r/Coronavirus: (('social', 'distancing'), 488), (('hazard', 'pay'), 485), (('going', 'get'), 280), (('Filtering', 'Facepiece'), 253), (('research', 'papers'), 252), (('San', 'Francisco'), 252), (('United', 'States'), 251), (('turned', 'away'), 247), (('people', 'protesting'), 225), (('looks', 'like'), 223), (('last', 'week'), 214), (('stupid', 'people'), 206), (('people', 'get'), 205), (('second', 'wave'), 197), (('people', 'would'), 196), (('people', 'getting'), 194), (('common', 'sense'), 194), (('stimulus', 'check'), 191),

/r/China_Flu: (('human', 'human'), 74), (('whole', 'world'), 70), (('exactly', '50'), 64), (('rest', 'world'), 49), (('human', 'transmission'), 46), (('long', 'time'), 44), (('deflecting', 'blame'), 42), (('people', 'like'), 41), (('sign', 'petition'), 41), (('acknowledge', 'Taiwan'), 40), (('5', 'accuracy'), 39), (('Wuhan', 'Virus'), 38), (('every', 'country'), 38), (('per', 'minute'), 36), (('North', 'Korea'), 35), (('Maybe', 'China'), 33), (('Virus', 'pandemic'), 32), (('much', 'better'), 32), (('isolated', 'treatment'), 32), (('Taiwan', '039'), 32), (('quot', 'Taiwan'), 32), (('withheld', 'knowledge'), 32), (('knowledge', 'highly'), 32), (('highly', 'infectious'), 32), (('many', 'countries'), 32), (('Article', '63'), 32), (('spread', 'rumors'), 32),