Method - Data Collection

Using the keyword "coronavirus" for querying, we collected news articles published on news websites in the United States from the beginning of 2020, where the first cases of coronavirus were reported, until April 19, 2020. We used a third-party API called Currents API which generated data in JSON format. For sources, we looked at the top 15 U.S news websites measured by unique monthly visitors (Statista, 2020), excluding the news aggregators (Yahoo News, Google News) and topic-specific newspapers (The Wall Street Journal). The free version of the API, despite returning some useful features including Title, Date, Description and URL, did not allow us getting the full text of the articles. Therefore, we found a resolution by putting all the URLs available from the previous step into the newspaper3k module in Python, which enabled scraping the necessary text. Furthermore, after checking the data quality from each source, we narrowed them down to these 6 news outlets: CNN, The New York Times, Fox News, The Guardian, USA Today and The LA Times.

For Reddit data, we searched for all comments from the subreddit CoronavirusUS from February when it was created until April 19, 2020. We used the pushshift.io Reddit API to build the URL with the relevant parameters which returned a page of JSON objects.

In view of Twitter data, the standard API only allows search queries of up to 7 days which means that we cannot scrape data in the normal way for more historical data. However, thanks to the support from the team at crowdbreaks.org, who is tracking Twitter trend on COVID-19 in real time, we managed to get all tweet IDs that they had collected since January in the U.S. By creating a Python environment and executing the download command following their guidelines, we retrieved the tweets from the provided IDs, also in JSON format.

We used both the jsonlite package in R and the json module in Python to extract key features from each JSON result, namely the date and text attributes. Unfortunately, because of Twitter's restriction allowing only 180 tweets every 15 minutes, we calculated a total download time of 15 days using 2 API keys we had. On that account, we decided to split the data in half by stratified sampling, keeping the distribution of number of tweets per week, in order to speed up data collection and obtain the data in time for the report. The Twitter data were collected until April 16, 2020.

For data cleaning, we removed duplicated data, texts that were too short and kept only texts written in English. Consequently, 14,149 news articles, 352,842 Reddit comments and 109,329 tweets were ready for analysis. The data distribution can be seen in Figure 1. The number of news reported on COVID-19 in the U.S were relatively low in the first 6 weeks (week 3 - week 8), and the subreddit CoronavirusUS was not even created until week 7. However, there was a dramatic increase starting from week 9 and peaked in week 11, 12 and 13 for Twitter, Reddit and News respectively. This was the end of March when the number of infected cases in the country surged.

A screenshot of a video game

Description automatically generated

Method - Pre-processing

For text data preparation, we focus on quanteda package in R, which is one of the most popular R packages for the quantitative analysis of textual data. Firstly, we created a corpus for the data set of each source. A corpus consists of a collection of documents (each document is an article, a Reddit comment or a tweet), and the document variables which describe the characteristics of the document, for instance published date and source. Subsequently, we tokenized each corpus, which separates the text into into its single words (also called terms or tokens).

We pre-processed the data sets by lowering upper case letters and removing the numbers, punctuations, symbols, URLs and separators. The main idea is to reduce the final amount of terms extracted, which is important in order to improve the accuracy of both topic modeling and sentiment analysis. If two words are similar it is convenient to combine them as one unique word. Moreover, if a word is not relevant for the analysis, it can be removed from. Hence, we also implemented stopword removal and lemmatization technique. Stopwords are words that appear in texts but do not give the text a substantial meaning (e.g., "the", "a", or "for") and lemmatization deals with the inflected forms of words by replacing them with their base forms.

In the next step, we generated a document-feature matrix (also known as document-term matrix) for each source. It represents how frequently terms occur in the corpus by counting single terms. We kept only the top 5% of the most frequent features (minimum term frequency set at 0.95) that present in less than 10% of all documents (maximum document frequency set at 0.1) to focus on common but distinctive features.

Method - Sentiment Analysis

To investigate sentiments in news articles, reddit comments and tweets, we first split a document into tokens. We calculated sentiment scores with the opinion lexicon developed by Bing Liu et al., which is included in the quanteda.dictionaries and the tidytext package. The Bing lexicon is a list of English words categorized into positive and negative categories. It comprises 2,005 positive terms and 4,781 negative terms.

For each source, we calculated the sentiment score as the percentage of the difference between the number of positive terms and negative terms operating on the total opinion terms

On the topic-based level, we assigned the sentiment value for each topic word with an integer of +1 for positive and -1 for negative, and the probability of a word in a topic given by the STM topic model as the term weight of the word. As a result, topic sentiment score is calculated as

An overall topic sentiment score is computed by multiplying the sentiment value by the probability of words and summing the products in a chosen topic. Multiple topics could have the same words with different probability values. Therefore, the sentiment score of a topic would be distinguished from others even if the same set of words appear in different topics.

Analysis – Sentiment Analysis

We examined the sentiment in news, Reddit and Twitter about coronavirus from week 3 to week 16 of 2020.

A close up of a map

Description automatically generated

Overall, the opinion towards COVID-19 of all three sources remained negative throughout the period. There was a similar pattern between news articles and Twitter, as the sentiment scores were significantly low in week 3 but showed an upward trend in the subsequent weeks and wavered at around -20 from week 12 to week 16. For Reddit, the sentiment of comments did not have much changes, as it fluctuated at around -30 and -20.

Using STM model with the same number of topics of 25 for all three sources, we calculated the sentiment score of each topic. Again, the general opinions in most topics were negative. For further analysis, we ignored the topics that only had emotional words but did not contain any interpretable topics (e.g. Topic 21 of Twitter with high probability words like "dumb", "stupid", "kid", "hate").

A screenshot of a video game

Description automatically generated

Reddit had some extremely negative sentiments, which showed that people were significantly worried about the fatality and the spread of the virus (Topic 4 of Reddit) and about the clinical symptoms of patients with COVID-19 (Topic 22). At the same time, news and Twitter had a few notable though not extreme negativity, with topic 22 and 23 in news and topic 6 and 11 in Twitter, whereas the other topics in these two sources were close to neutral. Twitter users appeared to have the same concerns as Reddit users, while the news media gravitated towards blaming politician and worry on social security. The difference in sentiment scores also indicates that Reddit comments convey more emotional expression than news articles and Twitter posts do.