Public Opion and Media Portrayal of Events related to the Novel Coronavirus (COVID-19)

Yehong Deng (ydeng2), Yuming Liu (yumingl),
David Owusu-Antwi (dowusu), Zhen Yuan (zheny)

CAPP 30122 – March 16, 2020

1 Overview

Our goal is to analyze the relationship between public opinion and media portrayal of information, specifically in relation to the Novel Coronavirus (COVID-19). For this particular case study, we restrict public opinion to Twitter (i.e., publicly available tweets) and media portrayal to various news articles (freely available, subscription-based, traditional, etc.) obtained through Google News. We collect all our data for the month of February. We carry out analysis through measuring keyword frequency in tweets and news article titles. In particular we examine whether or not different news sources capture public opinion, and in what ways. We collect news from nine sources: the Washington Post, the Wall Street Journal, the Verge, Bloomberg, Reuters, Buzzfeed, CNBC, Vice-News, and CNN. Then, we retrieve the 20 most frequently used words from tweets and news titles per day. Finally, we visualize the relationships between keywords from the news and from tweets, using three different visualization methods. By collecting the contents from Twitter and different sources of news, we can determine whether people and news media share similar concerns about the spread of COVID-19.

2 Commands

Run these commands before running application on Linux to collect all necessary dependencies.

```
sudo apt install python3-pyqt5
pip install tweepy
pip install newsapi-python
pip install seaborn
```

To execute application, enter the directory `data-visualization` and run (`./gui.py`) directly, or run one of the following commands:

```
ipython gui.py
python gui.py
```

3 Overall Structure

3.1 Data Collection

Primary Architects: Yuming, Yehong

(a) `twitter_api.py`: uses tweepy module for the Twitter API to crawl Twitter for the requested information.

(b) `google_news_api.py`: with the Google News API, crawls Google News for articles given a set of keywords, news sources, and a date-range to consider. **Google limits the date-range to at most one month from the current date.** To specify the date range, see line 13 (`[dates.append(’2020-02-’` str(i)) for i in range(11,30)]+) and for testing, change 11 to 20, for example.

3.2 Data Cleaning and Transformation

Primary Architect: Zhen

(a) `cleaning.py`: cleans scraped data for analysis, reading from CSV files (scrapped from Twitter and Google News) into dataframes. Defines two sets of stopwords to neglect during keyword collection (`stopwords`: trivial words, `stopwords_for_news`: contains word in news source names):

- with `read_csv`, recursively finds all CSV files present in the `tweets_data` and `news_data` folders, iterating over the files and converting to a single tweets dataframe and a single news dataframe,
- with `split_time`, formats timestamps for tweets and news records,
- with `get_top_K_words`, retrieves the top K (modifiable, but set as a constant to 20) keywords in tweets and news dataframes, respectively, by frequency of occurrence; returns a list of keyword-frequency pairs.

(b) `transformation.py`: imports cleaned data from `cleaning.py` and transforms data into dictionaries of lists corresponding to x-axis and y-axis data, for plotting.

3.3 Data Visualization

Primary Architect: David

(a) `plotting.py`: reads transformed data from `transformation.py` to draw different plots using seaborn and matplotlib (i.e., correlate frequency, cumulative frequency, keyword matching).

(b) `gui.py`: embeds generated seaborn figures as interactive plots with selection control panels to update each plot by keyword or news source selection.

4 Accomplishment

Our project used the Twitter API to scrape public tweets on COVID-19 based on a set of hashtags. We used the Google News API to scrape news articles, from the specified news sources, on COVID-19. In particular, we scraped the top 20 most popular news articles per day for most of the month of February (2/11/2020 - 2/29/2020). We crawled over 17,000 unique tweets and over 1,800 news articles.

We show our results by three interactive visualizations: a frequency correlation graph, cumulative frequency graph, and keyword matches graph. The frequency correlation is a scatter plot of keyword frequency in news article titles versus keyword frequency in tweets, with interactive selection among the top 20 words appearing in the news titles to view how often those words appear in tweets. The cumulative frequency graph plots the frequencies of the top 20 keywords from tweets per day, with interactive selection of which keywords to plot frequencies for. The keyword matching graph plots, for the top 20 keywords appearing in both tweets and news article titles, the number of matches between keywords in tweets and keywords in a particular media source per day. For the cumulative frequency and matches, the users can multi-select and compare different legend.