**Appendix (Dataset Description)**

The data comprises of covid19 tweets, 42,000 in total, from Twitter and data from John Hopkins’s GitHub (<https://github.com/CSSEGISandData/COVID-19>).

**Twitter Data**

* date – date and time of tweet post
* name – name/Twitter handle/username of the tweeter
* sources – device poster used to post the tweet (converted to factor with 4 levels – iPad, iPhone, Android and Web)
* isretweeted – is the post a retweet or not (TRUE = retweet, FALSE = not a retweet)
* text – text/content of the tweet
* length – character count of each tweet
* favorites – the number of favorites a tweet received
* retweets – the number of retweets a tweet received

Below are the keywords that we selected for text mining and analysis:

* ishealth – # of occurrences for the text “health”
* ispandemic – # of occurrences for the text “pandemic”
* isvirus – # of occurrences for the text “virus”
* isemergency – # of occurrences for the text “emergency”
* isdeaths – # of occurrences for the text “dead” and “deaths”
* iswho – # of occurrences for the text “who” and “world health organization”
* iscdc – # of occurrences for the text “cdc” and “centers for disease control”
* isnih – # of occurrences for the text “nih” and “national institutes of health”
* isdisease – # of occurrences for the text “disease”
* isquarantine – # of occurrences for the text “quarantine”
* isrecover – # of occurrences for the text “recover”
* isban – # of occurrences for the text “ban”
* iscoronavirus – # of occurrences for the text “coronavirus”
* iscovid19 – # of occurrences for the text “covid19”
* iswash – # of occurrences for the text “wash”
* isracist – # of occurrences for the text “racist” and “racism”
* isasian – # of occurrences for the text “asian”
* ischinese – # of occurrences for the text “chinese” and “china”
* isinfectious – # of occurrences for the text “infectious” and “infections”

**John Hopkins’s Github**

Covid19 data were also cloned from John Hopkins’s GitHub repository. The two datasets that we utilized from the repository were the time\_series\_19\_covid-Confirmed and time\_series\_19\_covid-Deaths. These time series datasets include: Province/State, Country/Region, Lat, Long, and a time series between 1/22/2020 - 3/16/2020.

Since China, Italy and US are considered the epicenters of the crisis we grouped the two time series datasets by these countries.

* ch.confirmed – total # of confirmed cases in China
* it. confirmed – total # of confirmed cases in Italy
* ot.confirmed – total # of confirmed cases in other countries
* us.confirmed – total # of confirmed cases in United States
* ch.deaths – total # of deaths in China
* it. deaths – total # of deaths in Italy
* ot.deaths – total # of deaths cases in other countries
* us.deaths – total # of deaths in United States

The repository data sources include:

* World Health Organization (WHO)
* DXY.cn. Pneumonia. 2020
* BNO News
* National Health Commission of the People’s Republic of China (NHC)
* China CDC (CCDC)
* Hong Kong Department of Health
* Macau Government
* Taiwan CDC
* US CDC
* Government of Canada
* Australia Government Department of Health
* European Centre for Disease Prevention and Control (ECDC)
* Ministry of Health Singapore (MOH)
* Italy Ministry of Health