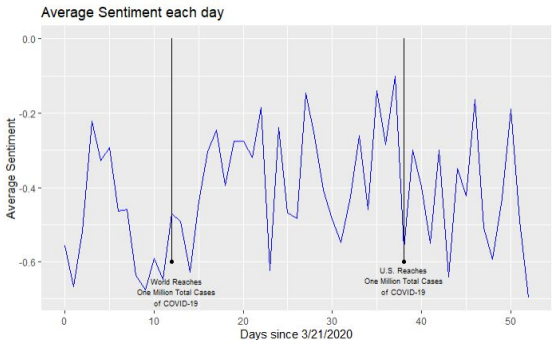
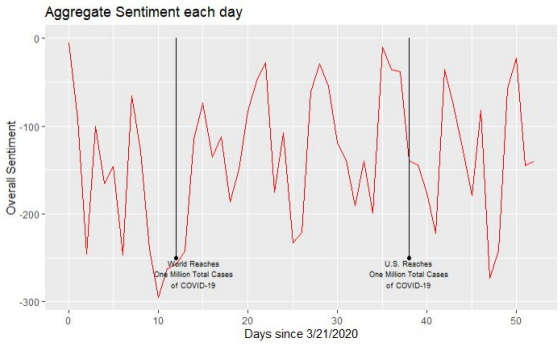
Team Members: Richard Yim (richyim555@g.ucla.edu), Susan Chen (susannalily@g.ucla.edu), Emily Hou (emilyhou1@gmail.com), Cassandra Tai (cassandratai1234@gmail.com), Vicki Truong (vickitruong29@gmail.com)

CRoSs_ValidatEd DataFest 2020 Writeup

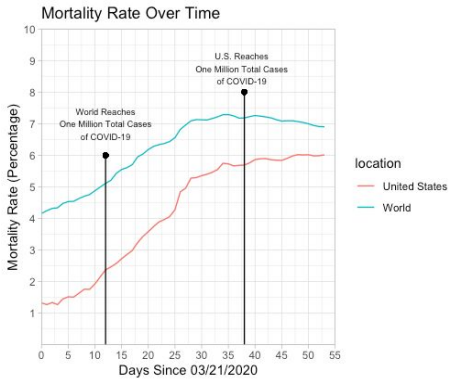
Coronavirus (COVID-19) has garnered significant media attention worldwide since the first outbreaks in late December of 2019. The situation has quickly escalated in affecting almost every continent to such an extent that the World Health Organization (WHO) declared COVID-19 a pandemic on 3/11/2020. As the ongoing situation develops, news stations have been providing continuous coverage from a gamut of perspectives of the pandemic. We used this as the topic of our discussion. We obtained a dataset with world news headlines from 3/21/2020 to 5/12/2020 as well as data regarding worldwide confirmed cases and deaths and performed sentiment analysis in R to explore a possible relationship between headline sentiment and mortality rates both worldwide and within the US.

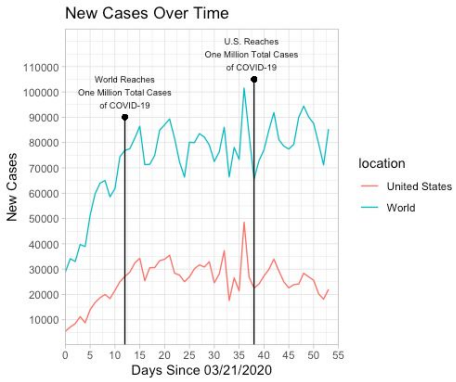
To do so, we calculated the daily mortality rate and new cases reported for those diagnosed with coronavirus both globally and nationwide. Additionally, we performed text mining on the news headlines with the help of package "tidytext": we found the frequency of keywords and also used a dataset of words from the package assigning sentiment scores between -5 to 5. We summed and averaged the overall score of headlines and word frequency each day. The daily aggregate score reflects the magnitude of articles written along with its general feelings. The daily average score was calculated from the non-zero sentiment scores, since many words had neutral scores of 0. However, one caveat from our analysis was that we assigned the sentiment value based on individual words not based on context, so positive superlatives on negative outcomes and vice versa would not have the correct sentiment scores. From this, we executed further investigation to determine any trends or influence between the news sentiment, word frequency, and coronavirus mortality rates.





From the graphs above, we can see that both the average sentiment scores and aggregate sentiment scores are negative throughout this time period, though the scores do fluctuate.





When comparing the sentiment graph to the graph of mortality rates, we can see that there appears to be no link between the two. So the sentiment of the news does not reflect the state of affairs.

Data Sources:

https://www.kaggle.com/gabrielmilan/multipurpose-world-news-dataset | https://github.com/owid/covid-19-data/tree/master/public/data/ | https://arxiv.org/pdf/1103.2903.pdf