CMPT 459 - Final Project

Masks and Covid-19

A report on mask sentiment and the spread of covid

Author: Eric Gao
August 7th, 2020

Background and Motivation

During the current pandemic, we as a society have seen a multitude of different issues regarding civil rights and freedoms be brought forward by concerned citizens. One such issue is the mandatory wearing of protective masks in order to limit the spread of COVID-19. Given the many people who protest the new laws, it would be interesting to know if having more dissenting citizens was somehow correlated to the spread of the virus in certain areas.

To this end, we went to twitter to measure the general sentiment of citizens across the United States. As stated above, we use a sentiment lexicon to grade each US tweet that's related to masks as an accumulation of their word's negative/positive sentiment ratings. After rating the individual tweets, we roll them up based on dates and states in order to preserve the timeline of changing opinions on masks over time.

Once the tweets are prepared, the data on covid cases is changed to match. By default, the case, and death statistics in the source data are listed as overall cases and deaths in each county in each day. Since we are more interested in infection rates then fatality rates, the deaths column is dropped. The base data is also split into counties along with states, meaning that we need to roll-up the data once more in order to match the tweet sentiment data. Once that is complete, we simply merge the two tables which now share the same date and state indexes. If there are any gaps in the data, we simply leave the overall Sentiment equivalent to the mean between the previous and next recorded sentiment.

Problem Statement

To begin with, we believed that generally, over time, the sentiment would first worsen before the cases began to grow, with a general 2 week gap between any change in sentiment and the gradual increase of confirmed cases per day. To this end, we can graph the number of new confirmed covid cases per state against the current sentiment towards masks of the same state. This way we can attempt to do 2 things, find correlations between the two datasets and perhaps categorize states based on their mask-infection rates.

Datasets

Covid related tweets - https://github.com/thepanacealab/covid19_twitter
Subjective Word Sentiments - http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
New York Times live covid report data - https://github.com/nytimes/covid-19-data

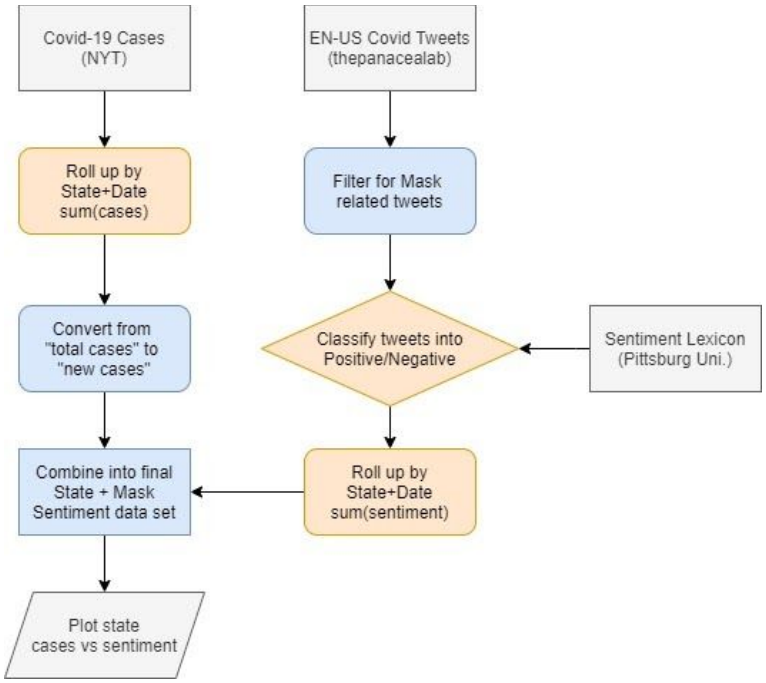
These were the three primary datasets that were used in this project. The first is one of the two provided tweet datasets recommended by the professor. This dataset was chosen due to its pre-computed categorizations allowing quick filtering of country and language. As the project focuses on english tweets in the United States, the existing tweet tags allowed for a

much faster hydration of the dataset. After basic hydration, tweets related to masks were further filtered out as per the project's interests.

The second dataset is a lexicon of words that have been generated through a research paper at the University of Pittsburgh in 2005. It contains 8000 words that were identified as carrying a strong or weak positive or negative sentiment. This is used to filter and grade the tweets from the covid related dataset. After the words are cleaned and lemmantized, they are compared with the dataset, a weakly positive word is given a score of 0.5, and a strongly positive word, given 1. Negative sentiment words have the same but scaled negatively, aka -0.5 for weak negative, and -1 for strongly negative words. The sentiment of all words in a tweet are added up to give the tweet's overall sentiment score.

The last dataset is a daily covid cases and deaths tracker that is released by the New York Times. The dataset contains information on each county's covid cases and deaths over time. This is used to gain an overall view of the covid spread in each state.

Architecture & Pipeline



(fig. 1) Overall Project flow

The project relies on a number of python programs to run each module in the chart above. From loading the tweets, to cleaning and classifying the text. Each part is modular and can be taken out and used elsewhere. The final datasets are saved as .pkl files, used by pandas to save dataframes, however prior to this, all files are saved as standard csvs or jsons.

Methodology

The first task after pulling mask related tweets originating from the US was to classify the tweets with a sentiment. The tweets were classified based on a sentence analyzer that compared keywords (individual) up to the generated sentiment lexicon referenced in the “Data Sets” section. Like stated in the Scenario, the words were first classified as either strongly, or weakly associated with sentiment and then given a rating between -1 and 1.

Roll-up was used for most of the data sets, as described in the “Data Sets” section of the report. First Rolling up the tweets based on their state, accumulating the overall sentiments by summing them up. Since the scale of sentiments went from -1 to 1 we could determine that if a state had more negative tweets about masks, the overall sentiment rating would be more negative, and if they had more positive tweets, then the overall would be more positive. Also, if rolled up into a more granular dataset, like cities for example. The number of tweets per day, per city is negligible at best and would make a bad representation of what could be a larger portion of people. In many instances, a single person’s tweet could shape the sentiment of an entire town for a day or two.

Next was to drill down the confirmed covid cases dataset to match the mask sentiment dataset. What we wanted to know is if the number of cases per day was increasing or decreasing and if they had some relation to the mask sentiment. However, the data provided was the total number of cases up to a certain day within each county, this meant first rolling up to combine counties from the same state, before drilling down/transforming the case data from total into new_cases per day only.

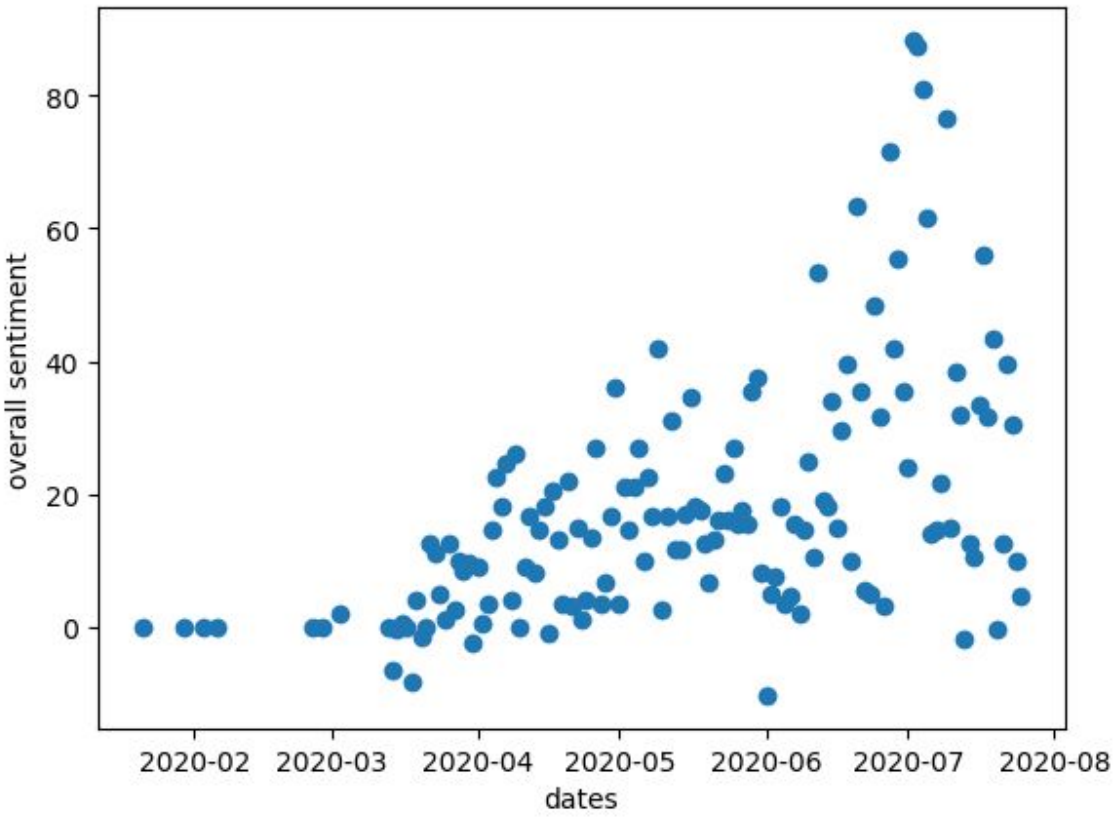
Once we had the datasets prepared, we combined them in order to search for global patterns. To do this, we created 2 dataframes, 1 for each dataset, and merged them together based on the (date, state) index. Each state had a unique entry per day referring to their # of new cases and overall mask sentiment on that day. Then we separated the overall dataframe based on state, resulting in 51 dataframes, each for 1 state with indexes on days between January 21st and July 27th. Once all the state data frames were created we then graphed their data points to analyze what sorts of patterns we could see based solely on their plotted graphs.

Evaluation

1. Overall Analysis of USA

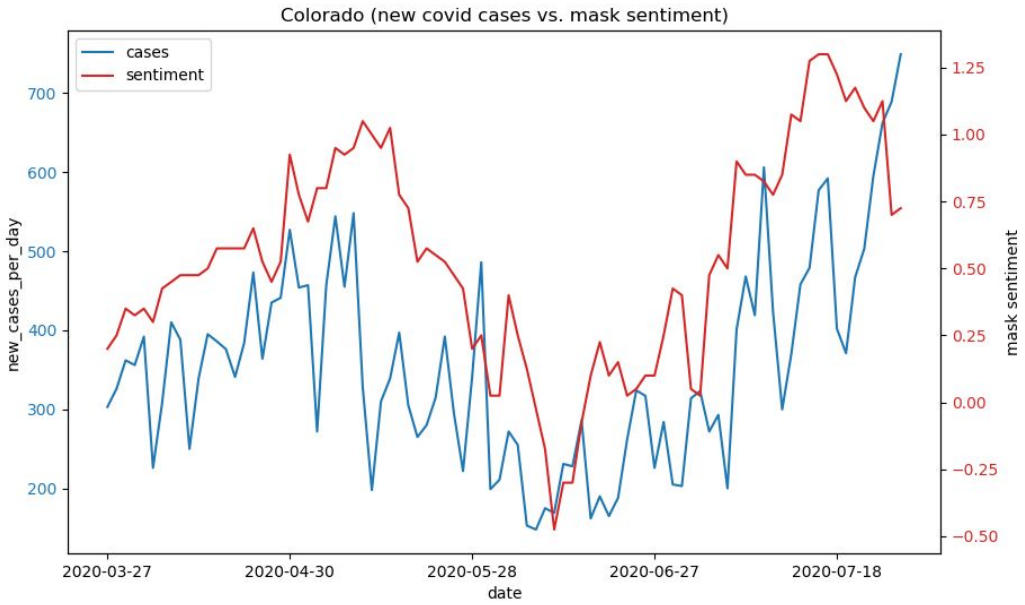
Before drilling down to individual states, I mapped out the overall sentiment of people in the US on the topic of wearing masks. Just to see if there is an overall trend:

(fig.2) Overall Sentiment in the US

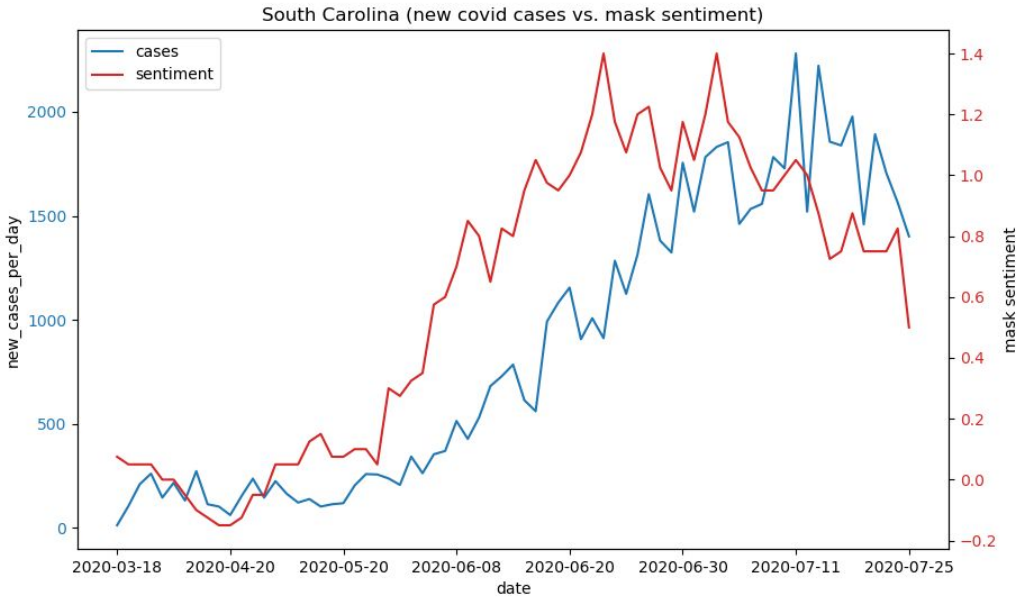


2. Analysis on States

Overall, in states that had sufficient data points (aka tweets), we see a strong correlation between the overall sentiment of masks and the number of new cases that appear daily. Generally, as cases go up, the sentiment will follow, and if sentiment drops, cases drop too.

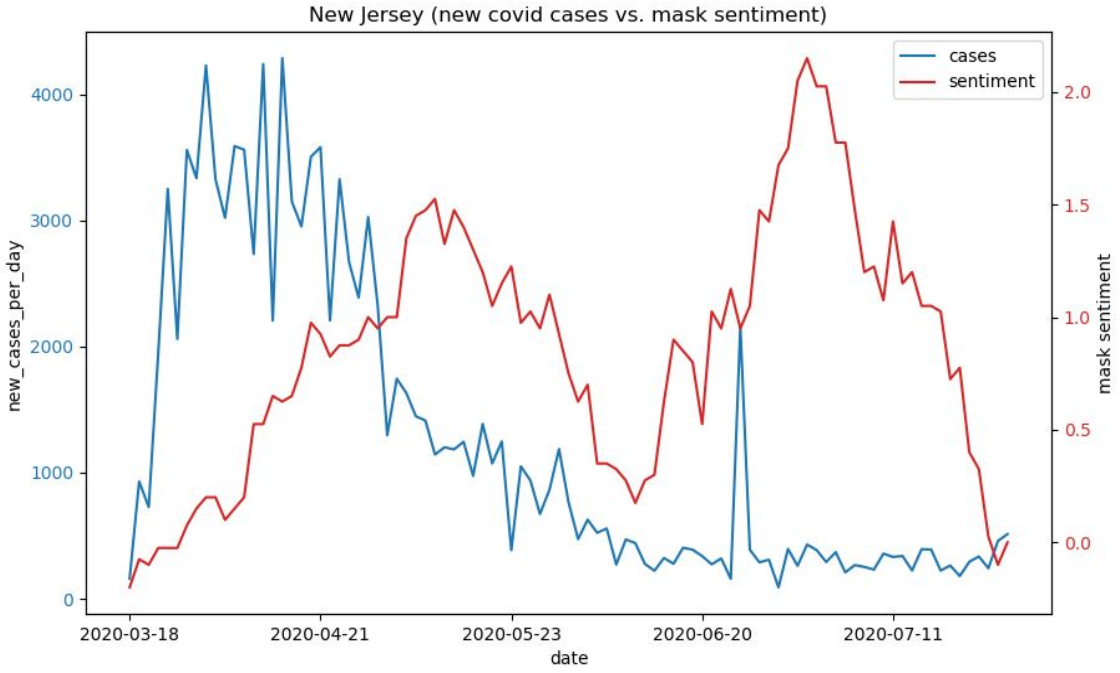


(fig.3 Colorado)

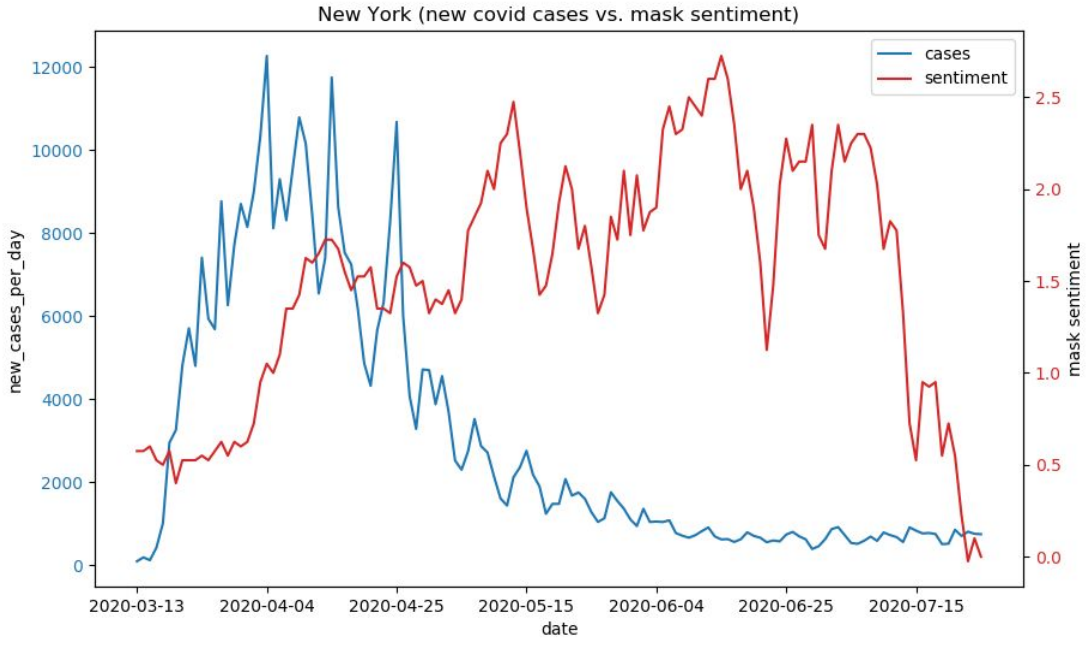


(fig.4 South Carolina)

There are a few exceptions to this, for example New York and New Jersey, who had an early spike of cases before the new cases slowed down. In these states, the sentiment towards masks generally stays at a high ish level across the board with little variation:

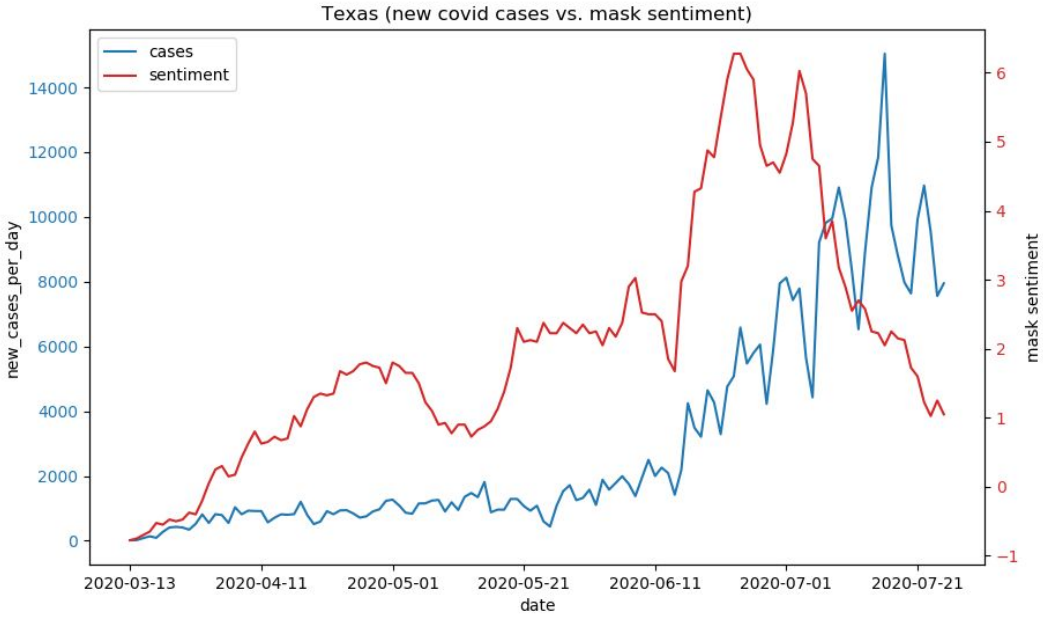


(fig.5 New Jersey)

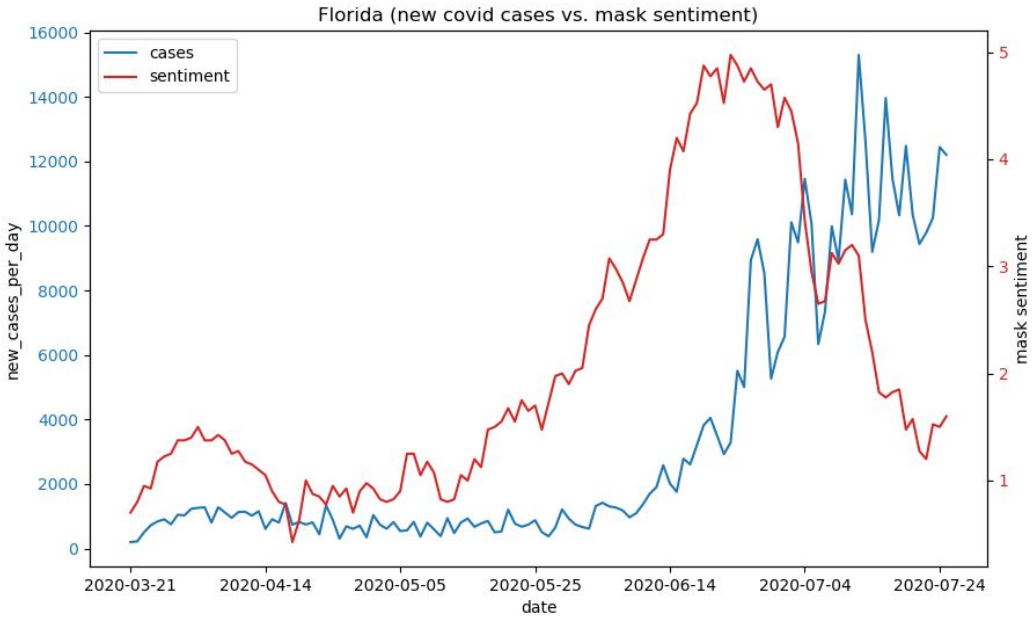


(fig.6 New York)

Then there were states that seemed to follow the trend up until the past month or so, where the sentiment drops rapidly despite the overall number of cases per day still rising quickly. Typically these states are from the southern half of the US, such as Texas or Florida

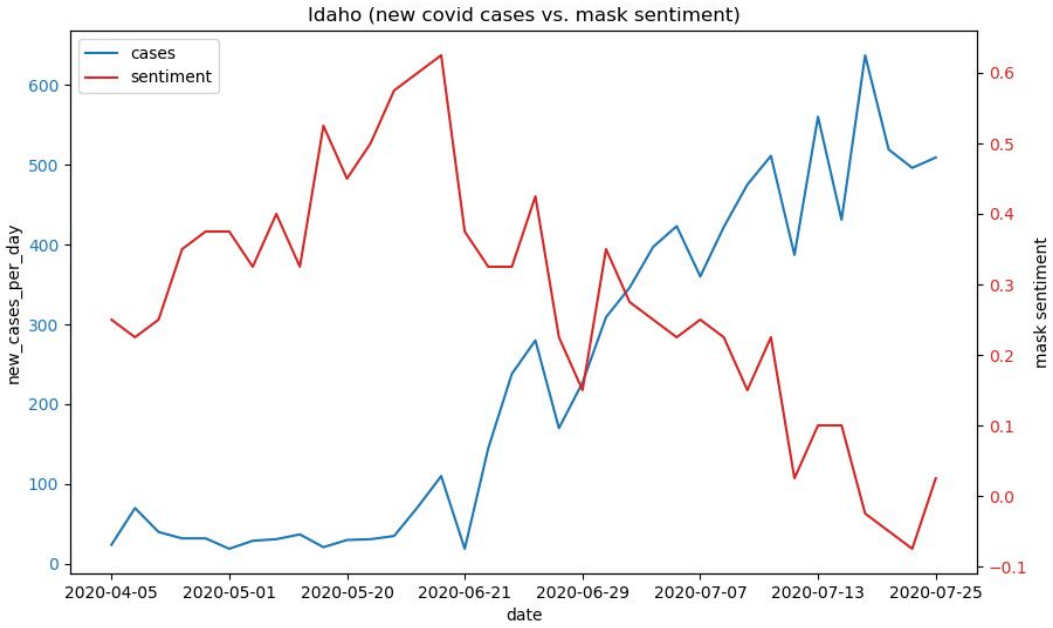


(fig.7 Texas)

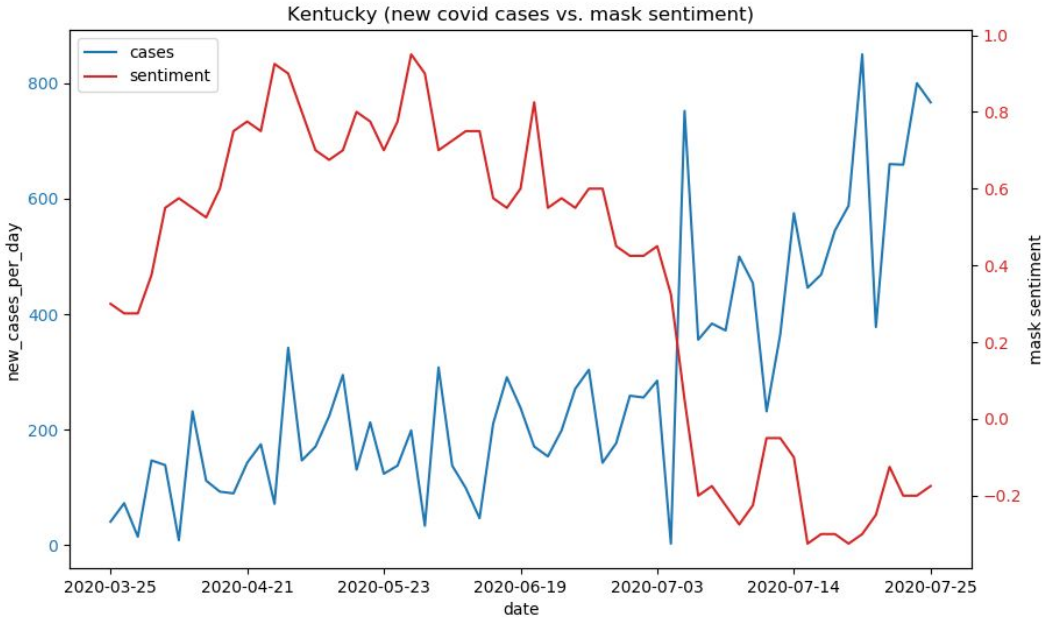


(fig.8 Florida)

And finally, there are many states who started off strong, with low cases and relatively positive sentiment towards masks, but as the quarantine dragged on, their sentiments slowly grew negative and cases began to spread faster and faster.



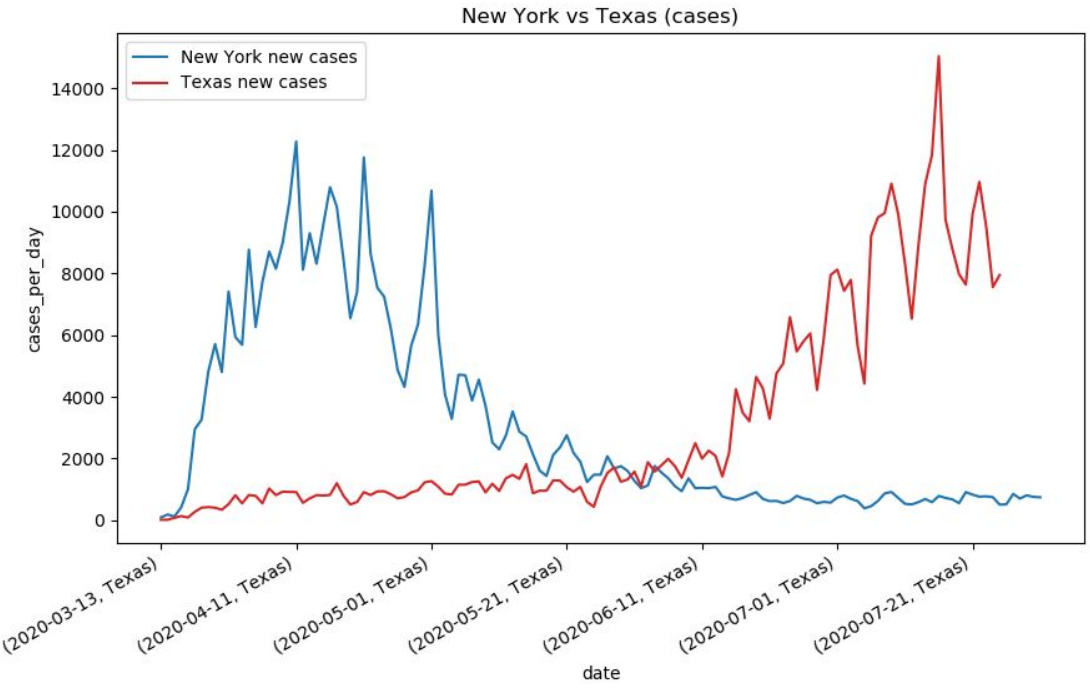
(fig.9 Idaho)



(fig.10 Kentucky)

3. Analysis between States

The illustrate the importance of mask sentiment, we can see the difference in cases over time of New York and Texas. Two very populous states with very differing sentiments. Their outcome in cases over time shows how important it is to have a positive view on masks and actively use them to slow the spread of covid-19



(fig.11 New Cases per day of New York vs. Texas)

Results

When starting this project, I didn't expect to find anything concrete, especially since the sentiment analyzer/lexicon was rather simplistic and unable to identify sentiment in phrases. However, as is shown by multiple States, there are consistent patterns that seem to reflect on how states and their peoples have reacted to the widespread use and popularization of masks. It's very telling that most states that are recovering have relatively stable positive opinions on masks, while those that slip are almost immediately punished by rising cases.

However, some shortcomings that I realized when working on this project include, overall data size being too small for various less populous states. For example, Wyoming, had less then 10 tweets about masks and covid that were traceable to the state and therefore the plot created by its sentiments is heavily skewed. Other graphs that look too simple have similar problems, not enough tweets to draw from, despite the recent news. A probable explanation for this, is the geo tagging system used by twitter, which tags many rural posts with a general US tag or city tag without indicating state. These tweets were cleaned out when I was setting up the

data and may have affected less populous areas more than those with large urban city centers, like New York or Texas.

The weakness of the sentiment analyzer may have been avoided by using a machine learning model to better predict the sentiments of tweets based on previously defined positive/negative sentiment tweets. If those tweets were covid related, then the model may become even more accurate. But as it stands, the general keyword based sentiment analyzer has already revealed many interesting patterns in the relation between mask sentiment and the spread of the coronavirus, giving further proof that masks are correlated to slowing the spread of covid-19.

References

https://github.com/thepanacealab/covid19_twitter August 2020

```
@dataset{banda_juan_m_2020_3757272,
  author       = {Banda, Juan M. and
                  Tekumalla, Ramya and
                  Wang, Guanyu and
                  Yu, Jingyuan and
                  Liu, Tuo and
                  Ding, Yuning and
                  Artemova, Katya and
                  Tutubalina, Elena and
                  Chowell, Gerardo},
  title        = {{A large-scale COVID-19 Twitter chatter dataset for
                   open scientific research - an international
                   collaboration}},
  month        = may,
  year         = 2020,
  note         = {{This dataset will be updated bi-weekly at least
                   with additional tweets, look at the github repo
                   for these updates. Release: We have standardized
                   the name of the resource to match our pre-print
                   manuscript and to not have to update it every
                   week.}},
  publisher    = {Zenodo},
  version      = {20.0},
  doi          = {10.5281/zenodo.3723939},
  url          = {https://doi.org/10.5281/zenodo.3723939}
}

g``
```

Theresa Wilson (2008). Fine-Grained Subjectivity Analysis. *PhD Dissertation, Intelligent Systems Program, University of Pittsburgh.*

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proc. of HLT-EMNLP-2005.

https://github.com/nytimes/covid-19-data August 2020
Contributors:
Mitch Smith, Karen Yourish, Sarah Almukhtar, Keith Collins, Danielle Ivory and Amy Harmon have been leading our U.S. data collection efforts.

Data has also been compiled by Jordan Allen, Jeff Arnold, Aliza Aufrichtig, Mike Baker, Nikhil Baradwaj, Robin Berjon, Matthew Bloch, Nicholas Bogel-Burroughs, Maddie Burakoff, Christopher Calabrese, Andrew Chavez, Robert Chiarito, Carmen Cincotti, Alastair Coote, Matt Craig, John Eligon, Tiff Fehr, Andrew Fischer, Matt Furber, Ariana Giorgi, Rich Harris, Lauryn Higgins, Jake Holland, Will Houp, Jon

Huang, Danya Issawi, Jacob LaGesse, Hugh Mandeville, Patricia Mazzei, Allison McCann, Jesse McKinley, Miles McKinley, Sarah Mervosh, Andrea Michelson, Blacki Migliozzi, Steven Moity, Richard A. Oppel Jr., Jugal K. Patel, Nina Pavlich, Azi Paybarah, Sean Plambeck, Carrie Price, Scott Reinhard, Thomas Rivas, James G. Robinson, Michael Robles, Alison Saldanha, Alex Schwartz, Libby Seline, Shelly Seroussi, Rachel Shorey, Anjali Singhvi, Charlie Smart, Ben Smithgall, Steven Speicher, Michael Strickland, Albert Sun, Thu Trinh, Tracey Tully, Maura Turcotte, Bella Virgilio, Miles Watkins, Phil Wells, Jeremy White, Josh Williams, Jin Wu and Yanxing Yang.