

# The Relationship between People's Comments about COVID-19 and Daily Cases

Bingan Chen

March 14, 2022

## Summary of Questions and Results

### Questions

1. What are the people's attitudes toward covid-19 recently?
2. How new cases about COVID increase by day?
3. What is the relationship between them, or, will people's attitude be affected by the change of daily new cases?

### Result

1. The proportions of attitude of people toward covid-19 are almost distributed in a relatively stable way.
2. Cases are increasing a constant rate daily.
3. There is no obvious linear relationship between the attitudes' proportions and increase in covid cases.

# Motivation

In my downtime, I enjoy looking at social media like twitter to keep up with some of the real-time news and what people are saying about it. I suspect that many times the comments twitter tweets me are based on my own browsing preferences and browsing history, rather than a truly fair and comprehensive tweet. And it would be very time consuming and impractical to confirm this suspicion one by one using different accounts. As it happens, I learned the trick of working with `.csv` files in the 163 class. This gave me the idea to use these skills to prove my suspicions more effectively. The new crown virus is one of the most popular topics in the last two years. When I look at tweets on this topic, twitter's push mechanism tends to give me a strong preference for what I view. So I decided to use the official twitter api to extract a certain number of comments and use a machine learning model to evaluate the attitude of each tweet about covid. Since the number of new cases changes every day, I wanted to find out if the increase in cases would make the comments about covid on twitter more negative. That's why this project was done.

# Dataset

- The data gotten by Twitter API (<https://developer.twitter.com/en/docs/twitter-api>)
- Dataset from World Health Organization about the daily cases world-wide that are listed by date, country, etc.

# Method

1. Use panda and DataFrame modification to summarize data from a messy condition that contains lots of useless data, to a organized dataset that contains only useful

rows including country and increase cases by day. The index was set into each date.

2. The Twitter API was applied to get `.json` output for each individual Tweet.
  - (a) Using the generated token to be able to use api and return multiple tweets using loop (each will be recorded as a single **response**). Specifically, the only data fields that are useful, and only the Tweets mentions the specific word `covid`, which will be filter before download the data each time. I used `query` to achieve that goal.
  - (b) All the information along with the following sentimental analysis results will be written into a new `.csv` file for further use. The format need to be adjusted to make the combination of data in the following steps.
3. Use a pre trained Machine Learning Model to apply sentimental analysis to get the attitude for each individual sentences. The numerical value will be calculated while getting other information within the single Tweet `.json`. This column will be added to the `.csv` with tweet's data.
4. The cases data that has been filtered will be combined with the new `.csv` file by the shared index, the date.
5. Visualization:
  - (a) The trend of cases in recent days per day is plotted using `lineplot` vs. time.
  - (b) The daily change of attitude toward covid is plotted using `lineplot` vs. time
  - (c) The relationship between the attitude and the new cases per day.

**Results**

**Impact and Limitations**

**Challenge Goals**

**Work Plan Evaluation**

**Testing**

**Collaboration**