SENTIMENT ANALYSIS OF COVID-19 SOCIAL MEDIA POSTS

A STUDY BY:

LUKEN WEAVER, ALEX KLAPHEKE,
JON GODIN & REZA FARROKHI

PROBLEM STATEMENT

The COVID-19 response has been largely regional and state-based in nature. Some states have enacted strictly-enforced stay-at-home policies, while others have provided guidelines. It would be worthwhile to compare the sentiment analysis of social media posts across geographic regions and compare them to both the local policies on social distancing and the occurrences of the pandemic in those areas.

KEY QUESTIONS:

What does the sentiment about COVID-19 across time look like for each of the geographies?

Do changes in sentiment align with changes in policy?

METHODOLOGY

OUR ORIGINAL PLAN WAS TO SOURCE THE DATA FROM TWITTER POSTS

- There is a github repository containing on ongoing collection of tweets IDs associated with the novel coronavirus (https://github.com/echen102/COVID-19-TweetIDs). This would normally be an ideal source for COVID-19 sentiment analysis.
- However, the repository is now up to over 80 million tweets in English alone, which posed several problems for us
- We did not currently have the storage capacity to handle that volume level of tweets
- Cost – a Premium Twitter Developer account would be necessary to analyze this many tweets
- The two free versions of Twitter Developer have opposing limitations:
  - One product allows for a larger number of tweets to be scraped, but only for the past 7 days, which would not meet the purposes of our project
  - The other allows full access to all tweets, but only for a small number of pulls to be scraped, which would limit our scope
- Finally, very few Twitter posts are geotagged by their posters, limiting our ability to focus on narrow geographies

ENTER REDDIT:
ENDPOINTS OF INTEREST: R/NYC & R/HOUSTON

Our program can sample data from any subreddit. As a proxy for geographic data, we chose to sample from general discussion subreddits for localities and decided on a direct contrast and comparison of two in particular.

Criteria:

- 1. A contrasting element regarding how Covid-19 was affecting the area and how it was being responded to
- 2. A minimum amount of data to work with, which caused us to focus on cities

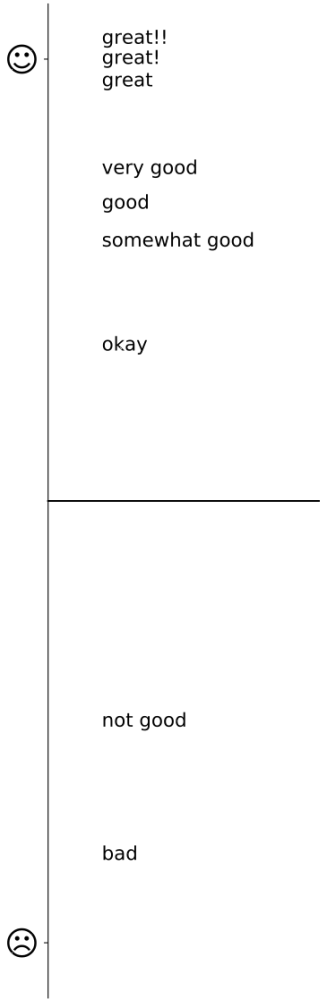
New York City, NY, and Houston, TX, have interesting contrasts both in terms of policies and case numbers, despite both being among the largest cities in the US.

New York City has multiple dedicated subreddits. We chose r/nyc as it has the largest user base.

Note: Reddit does have weaknesses as a source. Generally speaking, 90% of those visiting a subreddit are “lurkers”, with only 1% being responsible for the majority of content. The user base is also presumably less diverse than Twitter.

DATA GATHERING

- The Reddit API has two databases for **original posts** (submissions) and **comments**, respectively.
- We chose to focus on comments, as we felt they were more likely to express sentiment. Submissions were needed to ensure topic relevance.
- Our data gathering module first searches the submissions database for the most commented posts using a keyword or set of keywords in a given time frame (for our baseline, the keywords we used were “covid-19”, “coronavirus”, “quarantine”, and “pandemic” for the 80 day period between February 2nd and May 10th) inside given subreddits (the local subreddits for Houston and NYC)…
- … extracts the unique link ids of those posts…
- … then performs a query on the comments database to pull all comments relating to those link ids (i.e. all comments for those posts) .
- The data for each subreddit were then balanced by taking the smallest sample (Houston had fewer comments than New York) and then sampling all subreddits for that number of data points to be fed into our sentiment analysis.



SENTIMENT ANALYSIS

We utilized the VADER sentiment analyzer: (Hutto & Gilbert 2014)

- Matches 7,500 keywords, which are human-annotated for polarity (☺ /) & intensity (how strongly felt).

- Incorporates contextual information, such as:
  - **Negation** not
  - **Intensifiers** very really !
  - **Hedges** somewhat pretty ?
  - **Emoticons** :) :(

- Works on unprocessed text, so no need to tokenize, etc.

SENTIMENT ANALYSIS: INTERPRET WITH CARE

Sentiment analysis only tells you what the current mood is, *not* why it is that way.

> For example, in late April, r/nyc showed negative sentiment at the word “Fauci”—not from displeasure with *him*, but rather in response to reports indicating that Trump might fire Fauci.

Positive and negative sentiments can cancel out, so the score reflects consensus, or lack thereof.

Also, because sentiment scores are usually reported as averages, they can often look relatively flat or neutral on average, such as might occur if two different camps of reddit posters have very different sentiment that end up basically cancelling each other out.

CORONAVIRUS CONTEXT

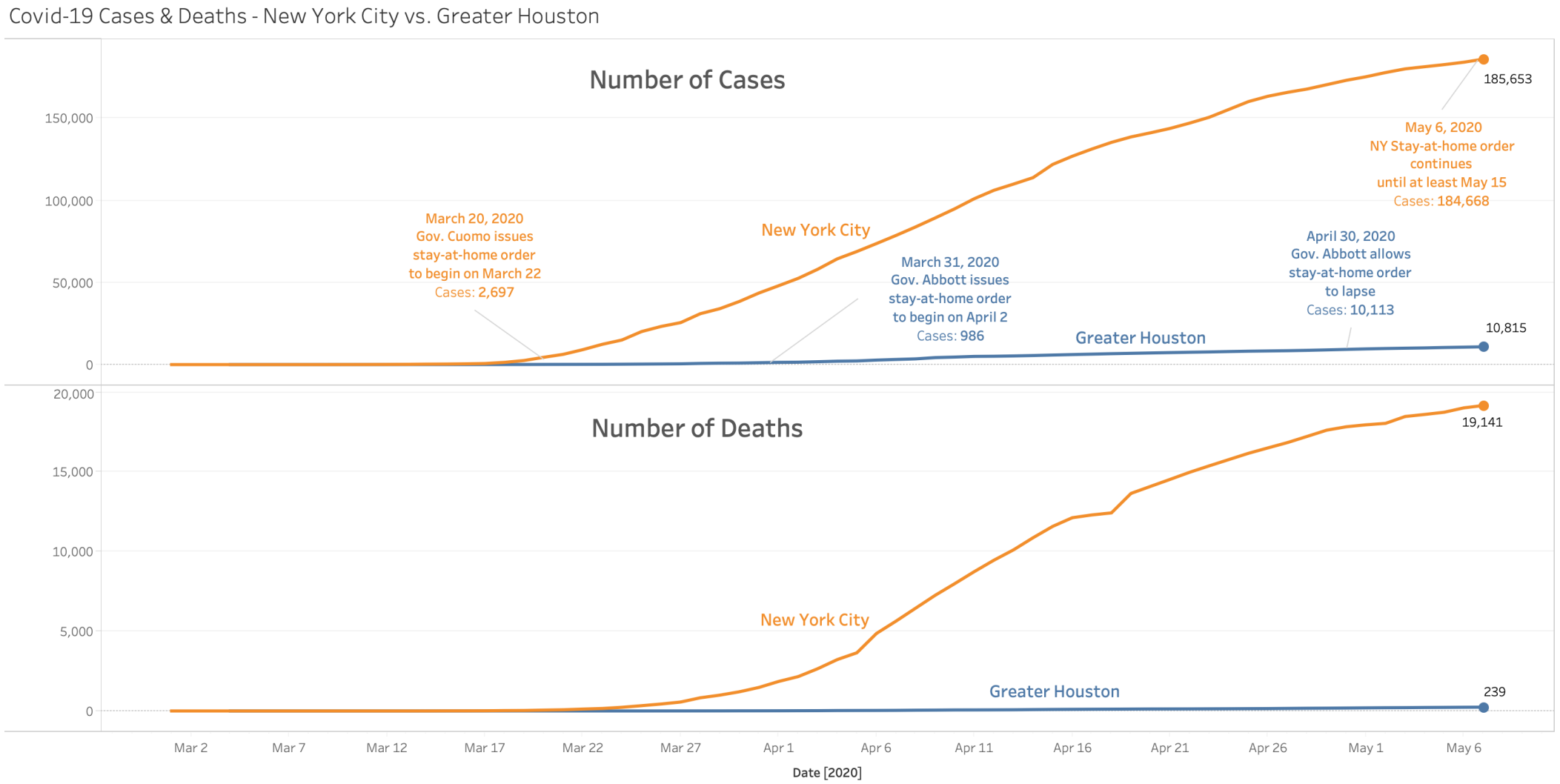
CASES & DEATHS IN NEW YORK CITY VS. GREATER HOUSTON

TIMELINE OF KEY EVENTS RELATED TO COVID-19

Date	Event
March 15	CDC issues guidance recommending against gatherings of 50+ people for 8 weeks
March 16	President Trump issues new guidelines urging avoidance of gatherings of 10+ people and to restrict discretionary travel
March 17	President Trump tells reporters “this is a pandemic”
March 20	New York Governor Andrew Cuomo issues state-wide order that all non-essential workers must stay at home, effective March 22
March 25	Deal struck on Capitol Hill regarding economic stimulus bill
March 31	Texas Governor Greg Abbott issues stay-at-home order, effective April 2
April 14	President Trump announces decision to halt US funding to the World Health Organization (WHO)
April 17	Governor Abbott announces phased reopening of the Texas economy starting April 20
April 26	Governor Cuomo lays out broad outline for gradual restart of the state, allowing low-risk businesses upstate to re-open in mid-May. No suggestions of loosening restrictions in NYC in the near future.
April 30	Governor Abbott allows stay-at-home order for Texas to lapse
May 1	In Texas, all retail stores, restaurants, movie theaters, and malls were allowed to reopen with limited capacity
May 15	New York state stay-at-home order set to expire

Sources: https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic_in_the_United_States , https://www.nytimes.com/interactive/2020/us/states-reopen-map-coronavirus.html , https://www.texastribune.org/2020/04/28/texas-reopening-restaurants-greg-abbott/

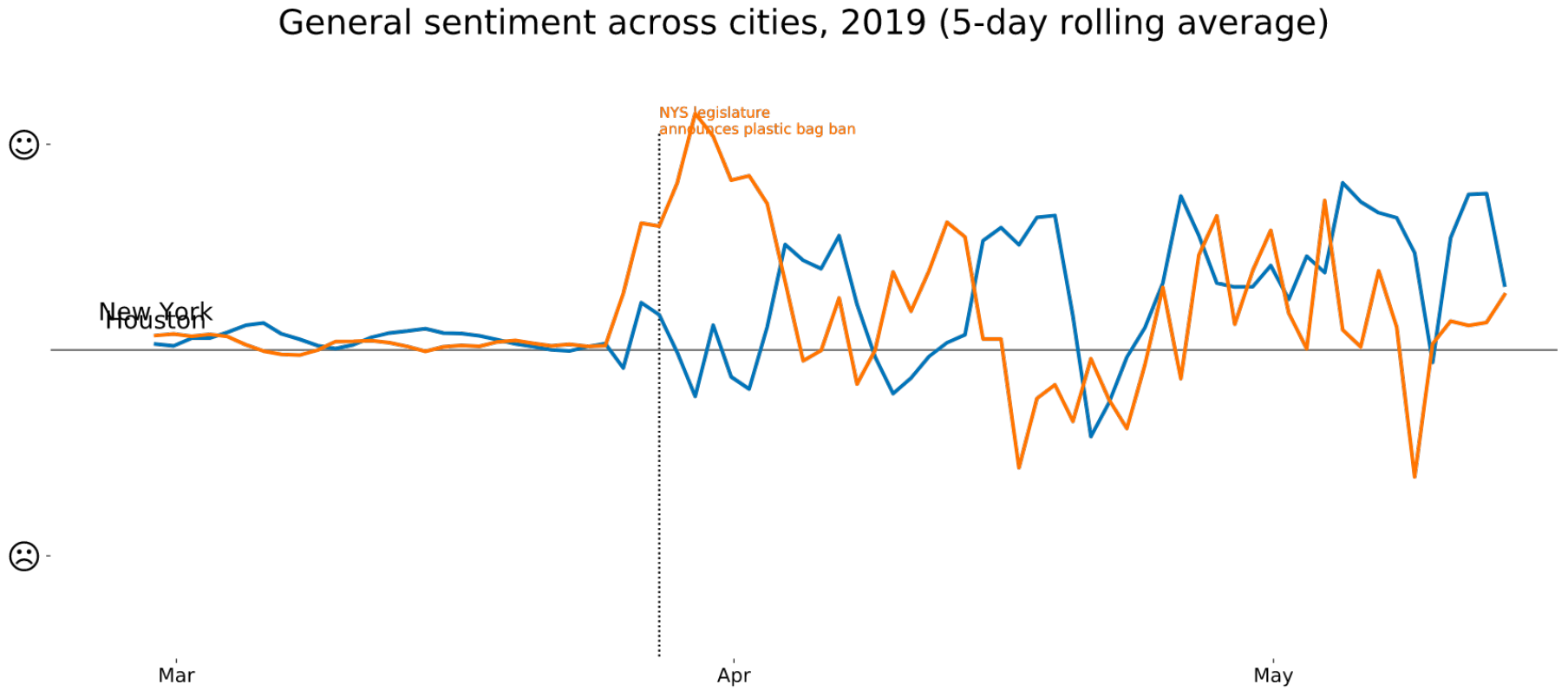
COVID-19 CASES & DEATHS BY REGION



Source: New York Times, based on reports from state and local health agencies https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

SENTIMENT ANALYSIS RESULTS

COMPARISON OF SOCIAL MEDIA POST SENTIMENT OVER TIME IN NEW YORK CITY AND HOUSTON, TX

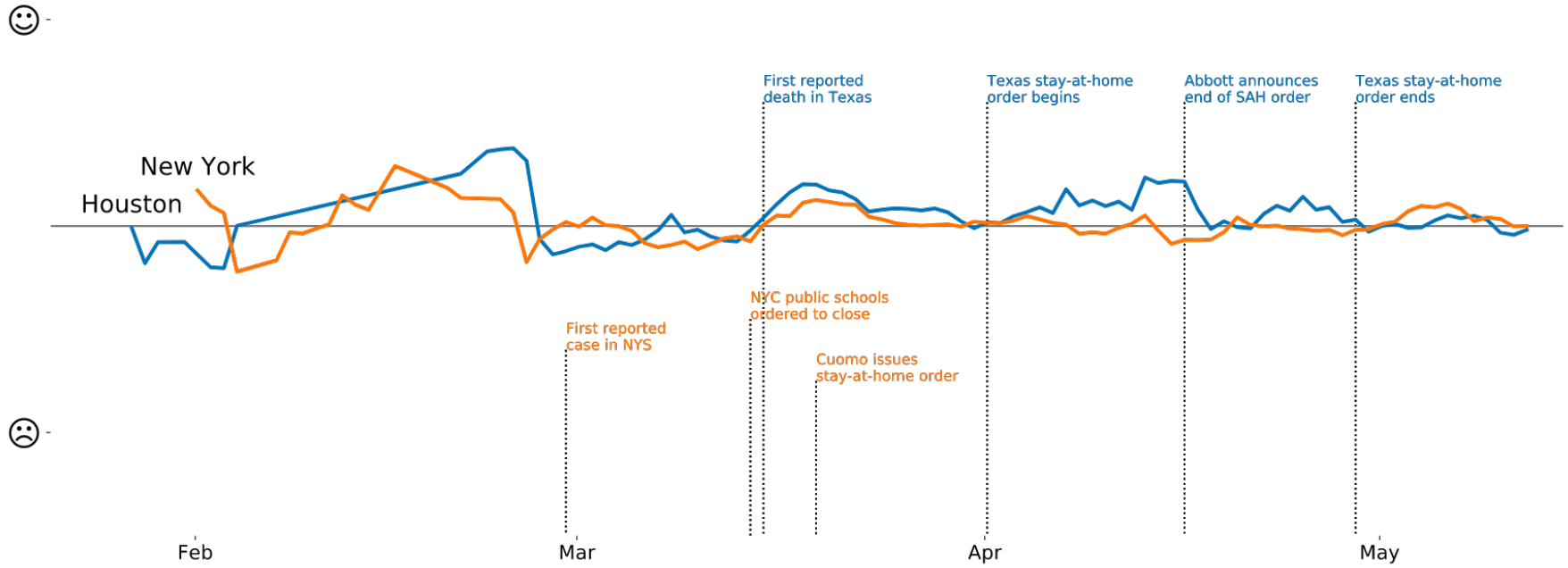


WE USED THE SAME TIME PERIOD FOR 2019 AS A BASELINE

For reasons we don't quite yet understand, sentiment in both Houston (blue) and NY (orange) was relatively flat throughout March 2019, then became more volatile thereafter.

The largest positive spike in sentiment occurred when the New York state legislature announced a plastic bag ban.

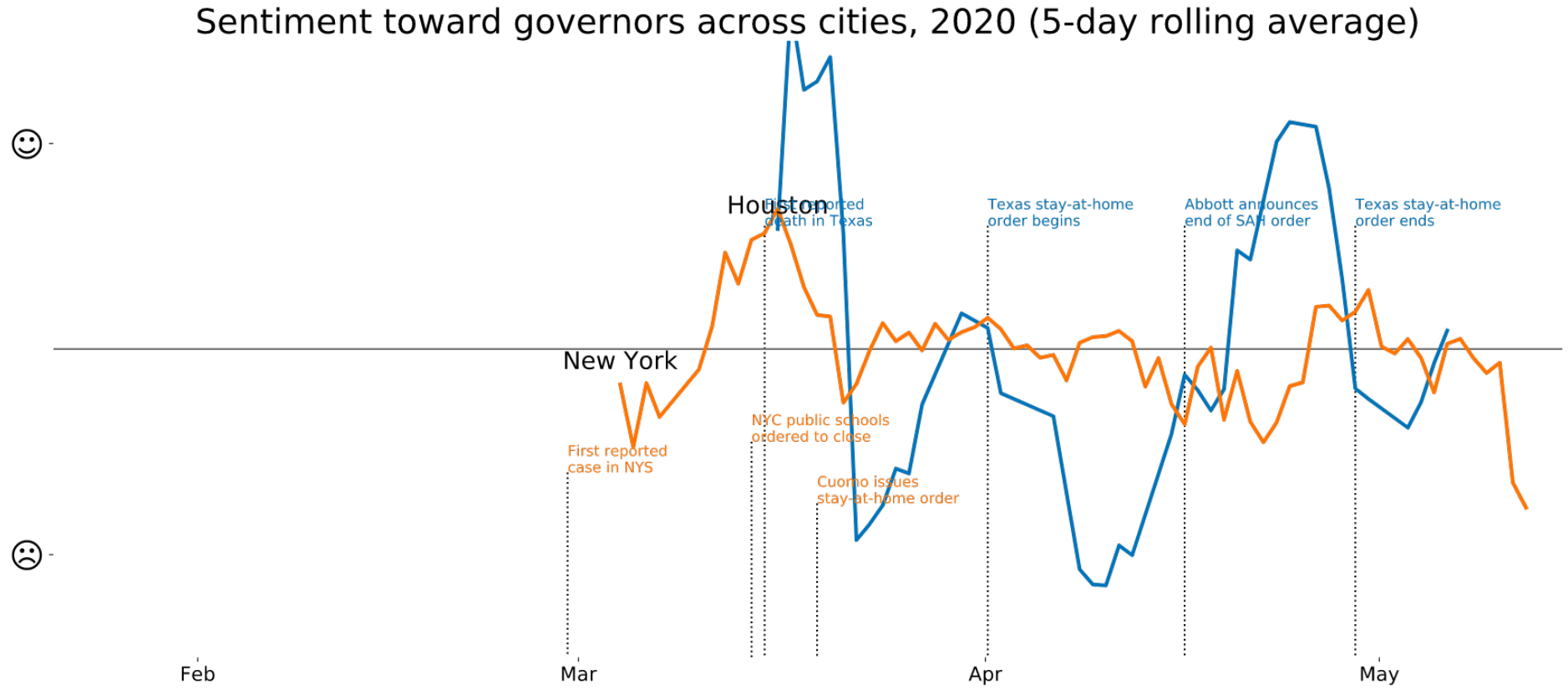
COVID-19 sentiment across cities, 2020 (5-day rolling average)



THE 2020 PERIOD COVERING THE COVID-19 CRISIS

Overall, sentiment in both cities, taken across all COVID-related posts, was generally neutral, with perhaps a slight lean towards the positive.

So, although we can uncover strongly polarized opinions beneath the surface, on the whole, it's difficult to see a clear picture emerge, even with the vastly different rates of cases and strength of restrictions in the two cities.



WE NEXT LOOKED AT POSTS MENTIONING THE GOVERNORS

Here, a very clear story emerges. Despite relative few cases and deaths in Houston, when Gov. Abbott issued the stay-at-home order in early April, sentiment plummeted in Houston. Sentiment in Houston also rose dramatically when it became clear that the stay-at-home order would be short-lived (among the shortest in the country).

In contrast, sentiment in New York in posts involving Gov. Cuomo remained relatively steady and neutral, if not slightly positive, despite global highs in cases and deaths and very strong restrictions about social distancing being in place.

CONCLUSIONS & NEXT STEPS

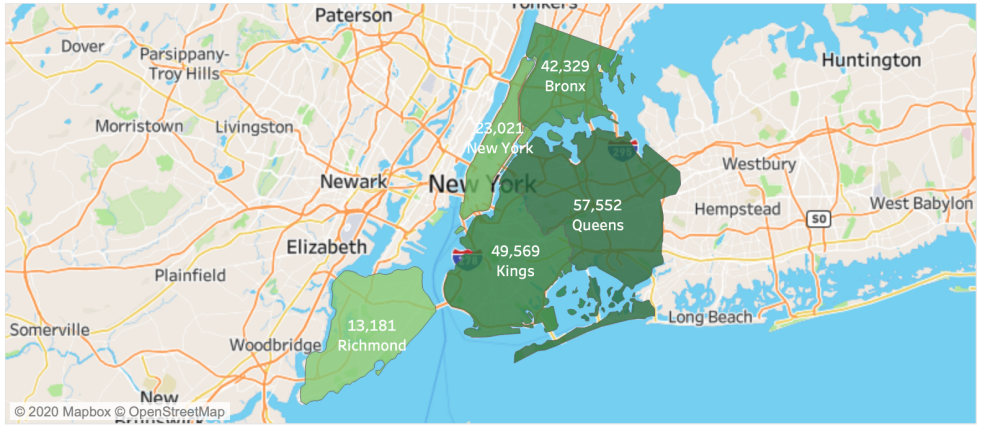
- In summary, we do feel like we created a working proof-of-concept that social media posts can be analyzed for sentiment in the face of a global (or even local) crisis
- The story was clearest when focusing on posts mentioning each of the respective state governors, where Texans tended to bristle at any suggestion of lockdown restrictions, while New Yorkers seemed more generally accepting given the high incidence of COVID in the five boroughs of New York.
- Sentiment analysis is tricky, however, so you must be careful when aggregating sentiment across and within potentially diverse local populations - many effects can get cancelled out, or the sentiment can be directed toward the opposite of the keywords used.
- For future steps, using Twitter posts holds potential due to the huge volume of COVID-related tweets that have been curated. However, the proper data infrastructure needs to be in place to handle such a large trove of data, and a Premium Twitter Developer account will be required for the analysis.

APPENDIX

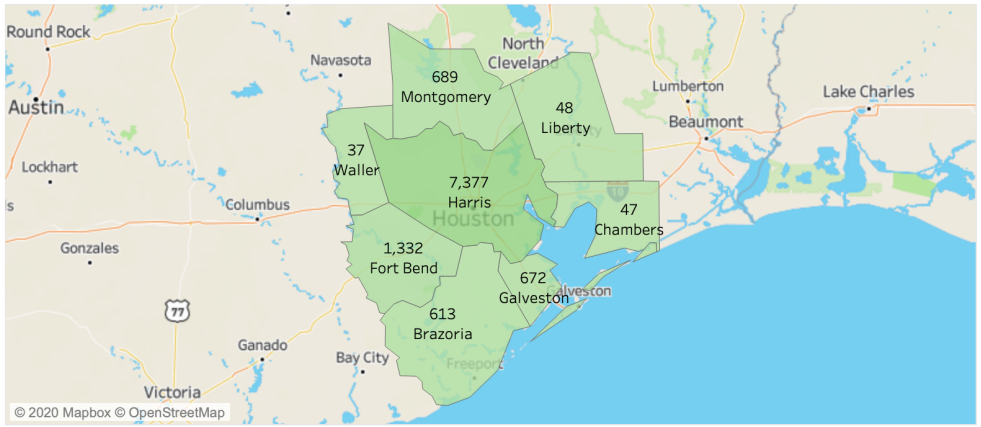
CASES & DEATHS BY COUNTY, NEW YORK CITY & GREATER HOUSTON

Cases

NYC Cases - Count Approximations - May 7, 2020

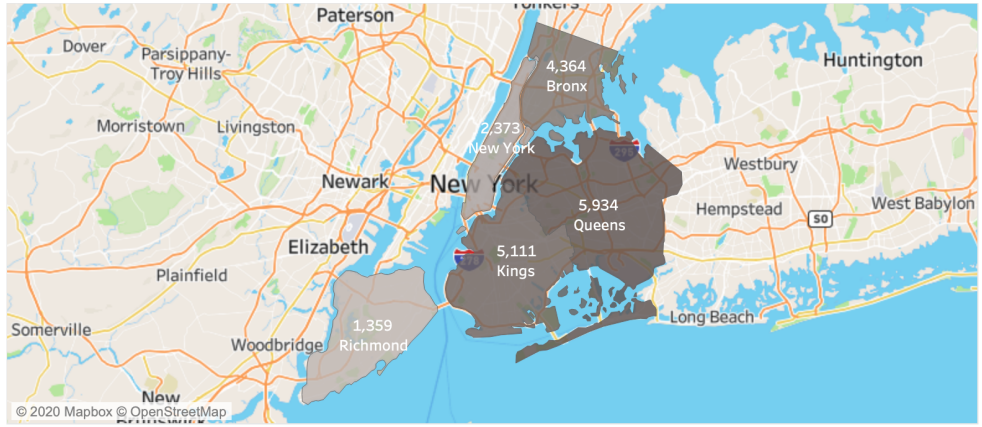


Greater Houston Cases - May 7, 2020



Deaths

NYC Deaths - Count Approximations - May 7, 2020



Greater Houston Deaths - May 7, 2020

Day of Date May 7, 2020

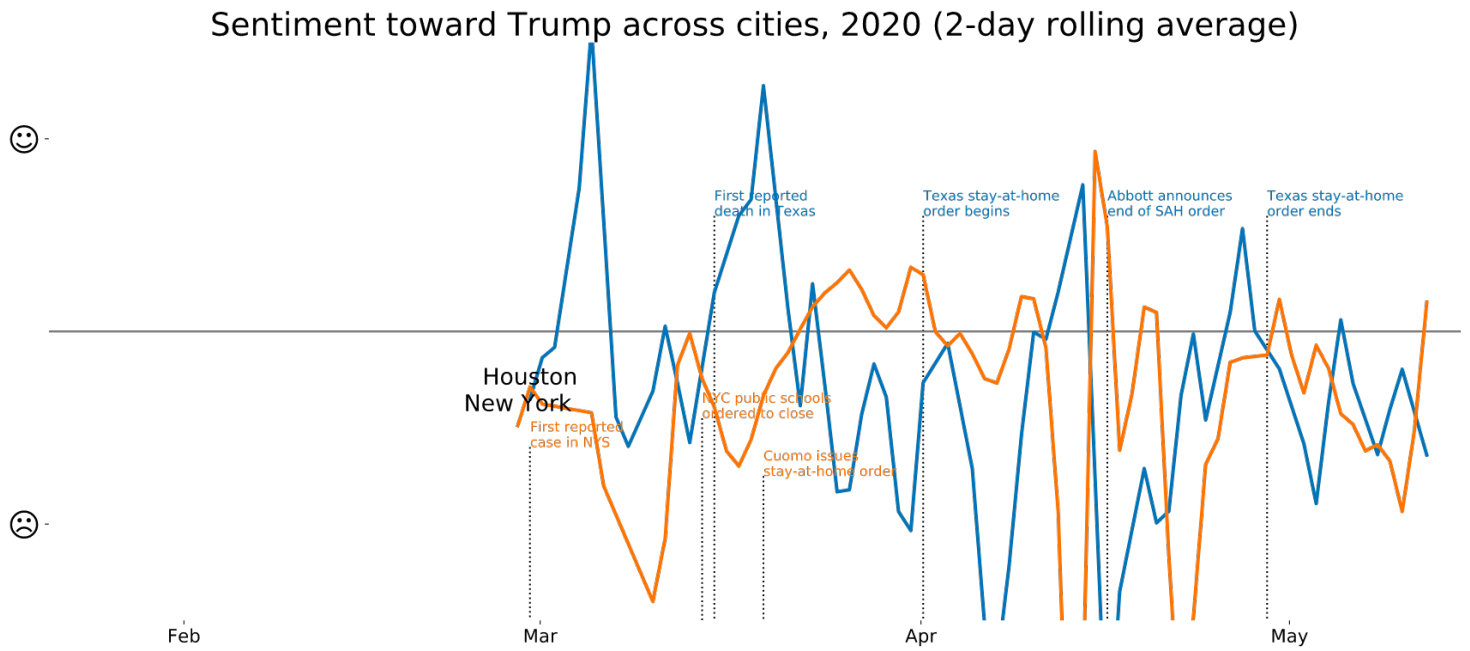
Show history

Source: New York Times, based on reports from state and local health agencies https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

Counts by County on the NYC map are approximations based on share of cases reported on the NYC Health page, multiplied by the NYC totals from the New York Times data above. Use exact values with caution, these are not official.
https://www1.nyc.gov/site/doh/covid/covid-19-data.page

SENTIMENT IN POSTS MENTIONING PRESIDENT TRUMP

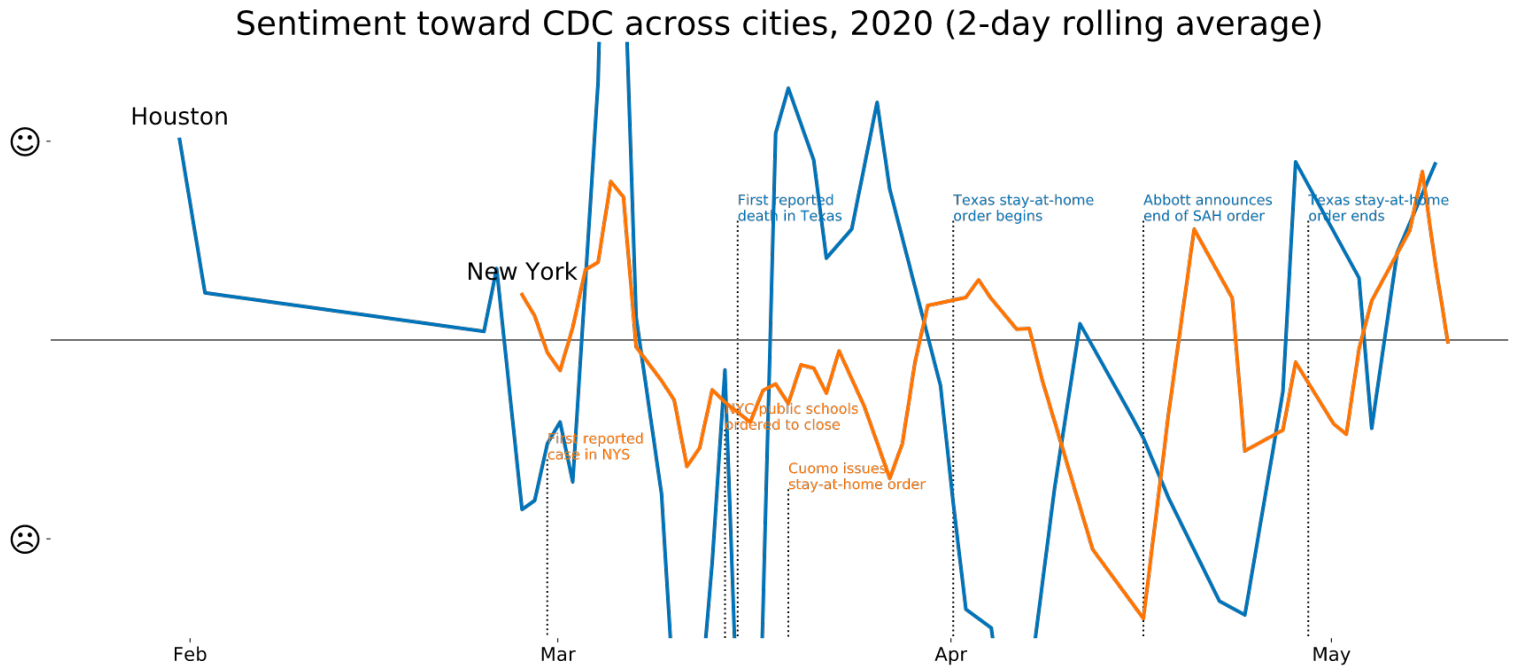
The sentiment for these posts was highly volatile in both cities, often in opposite directions.



SENTIMENT IN POSTS MENTIONING THE CDC

In early days, the CDC seemed to be the primary source of scientific perspectives about the emerging pandemic, and the sentiment tends towards the positive.

However, the volume of posts including the CDC diminished over time as Dr. Fauci emerged as the face of the scientific community, leading to higher volatility in the sentiment scores.



THANK YOU