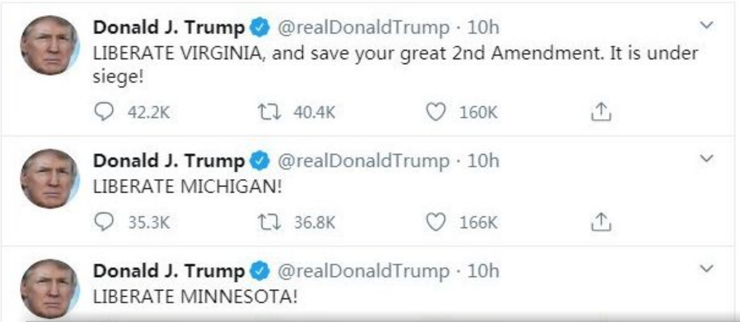
Mapping Coronavirus Sentiment

Is there a relationship between regional Twitter Sentiment and regional Coronavirus outcomes?

Outline

1. Introduction
2. Problem Statement
3. Data Collection
4. Data Processing
5. Sentiment Analysis Models
6. Maps
7. Takeaways
8. Further Analysis

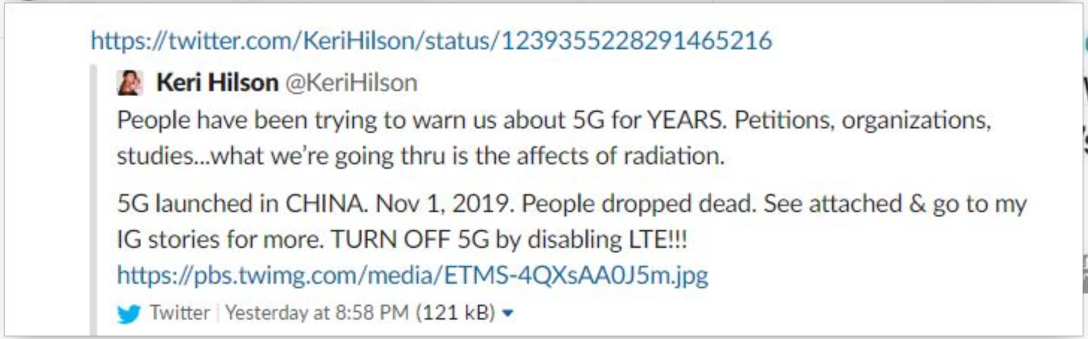
Introduction/Context



Donald J. Trump @realDonaldTrump · 10h
LIBERATE VIRGINIA, and save your great 2nd Amendment. It is under siege!
42.2K 40.4K 160K

Donald J. Trump @realDonaldTrump · 10h
LIBERATE MICHIGAN!
35.3K 36.8K 166K

Donald J. Trump @realDonaldTrump · 10h
LIBERATE MINNESOTA!

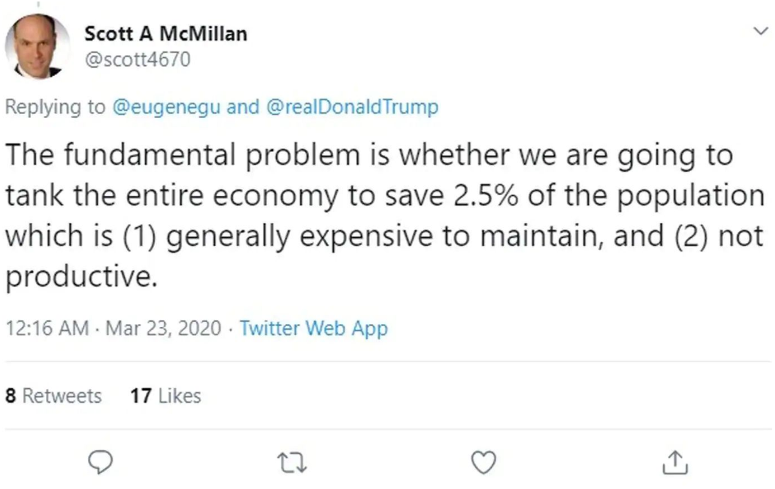


https://twitter.com/KeriHilson/status/1239355228291465216

Keri Hilson @KeriHilson
People have been trying to warn us about 5G for YEARS. Petitions, organizations, studies...what we're going thru is the affects of radiation.

5G launched in CHINA. Nov 1, 2019. People dropped dead. See attached & go to my IG stories for more. TURN OFF 5G by disabling LTE!!!
https://pbs.twimg.com/media/ETMS-4QXsAA0J5m.jpg

Twitter | Yesterday at 8:58 PM (121 kB)



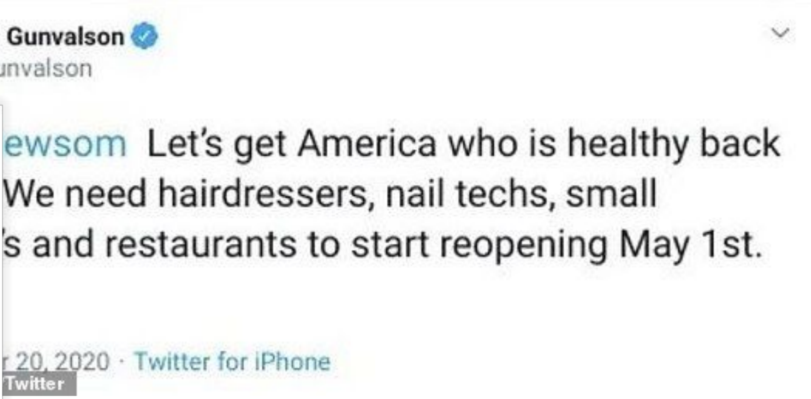
Scott A McMillan
@scott4670

Replying to @eugenegu and @realDonaldTrump

The fundamental problem is whether we are going to tank the entire economy to save 2.5% of the population which is (1) generally expensive to maintain, and (2) not productive.

12:16 AM · Mar 23, 2020 · Twitter Web App

8 Retweets 17 Likes



Vicki Gunvalson
@vgunvalson

ewsom Let's get America who is healthy back
We need hairdressers, nail techs, small
's and restaurants to start reopening May 1st.

r 20, 2020 · Twitter for iPhone
Twitter

Problem Statement

Can we predict the severity of the COVID outbreak in a region using Twitter data?

Twitterscraper

- Collaborated with DSI-11-NY, provided
- Added date based custom batch retrieval capability, and more efficient functions
- Able to grab tweets from specific latitudes and longitudes
    - Regions: SD, SF, Bakersfield, Chico, Redding, and Sacramento
- Grabbed ~140,000 total tweets from the last 3 months
- From the COVID, Coronavirus, Quarantine, etc. hashtags

Data Collection Issues/Assumptions

- Majority of tweets have no location data
- We are only capturing a minority of tweets for a given hashtag in a region
- Coronavirus outcomes: LA times county level data
- Limited to CA

Data Processing

- summarization of data pre-processing
    - duplicate tweet removal
    - dropped nulls in model using TextBlob (maintained 99% of data)
    - Lemmatization and tokenization by Spacy and NLTK
    - clean text function (lower case, remove websites, remove trailing characters, tokenization, lemmatization)

Sentiment Analysis Models

Sentiment Analysis: sPacy + nltk.opinion_lexicon

Used sPacy NLP library for :

- Tokenization
- Lemmatization
- Stop words
- Named entities

Nltk.opinion_lexicon

- 4800(approx) negative and 2000 positive words
- Each tweet was parsed and analyzed
- Count based sentiment (-1 negative, 0 neutral, 1 positive)

Sentiment Analysis: TextBlob

Focused on Sentiment Analysis polarity score

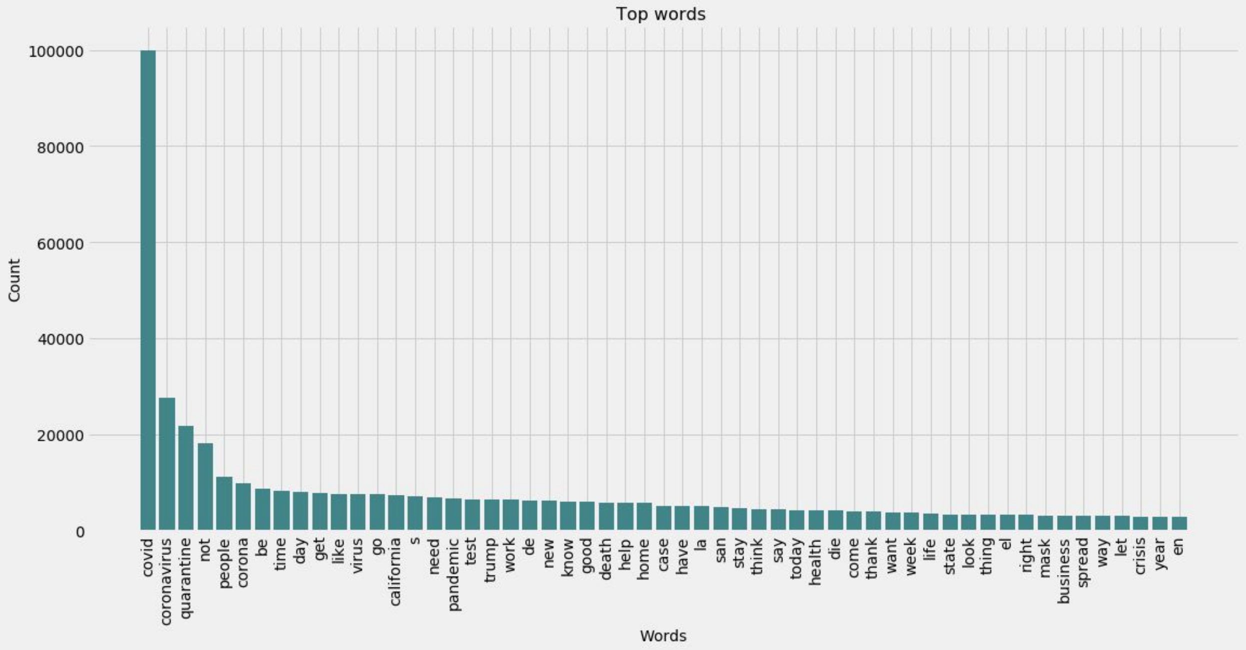
- score between [-1.0 (negative), 1.0 (positive)]

Other TextBlob Features:

- Part-of-speech tagging
- Tokenization
- Lemmatization
- Word frequency
- Spelling corrections
- Much more!

TextBlob

Model #1: sPacy,Nltk.opinion_lexicon corpus



- Identified Sentiment with NLTK's opinion_lexicon corpus
- 67%/33% train test split
- Labeled data with:
  - 37% neutral, 32% negative, 30% positive
- CountVectorizer, Logistic Regression
- Gridsearched over these CountVectorizer hyperparams :
  - Max features, n_gram range, min_df, max_df

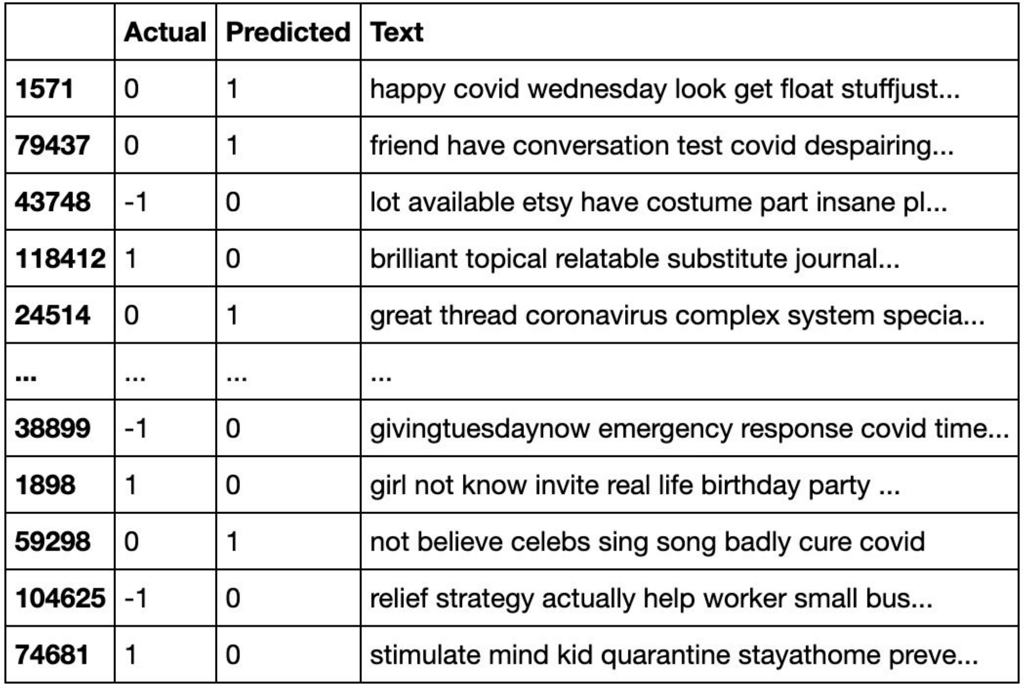
Results of Sentiment Analysis Model #1

Best Accuracy Scores:

Crossval: 88%
Training: 96%
Test: 90%

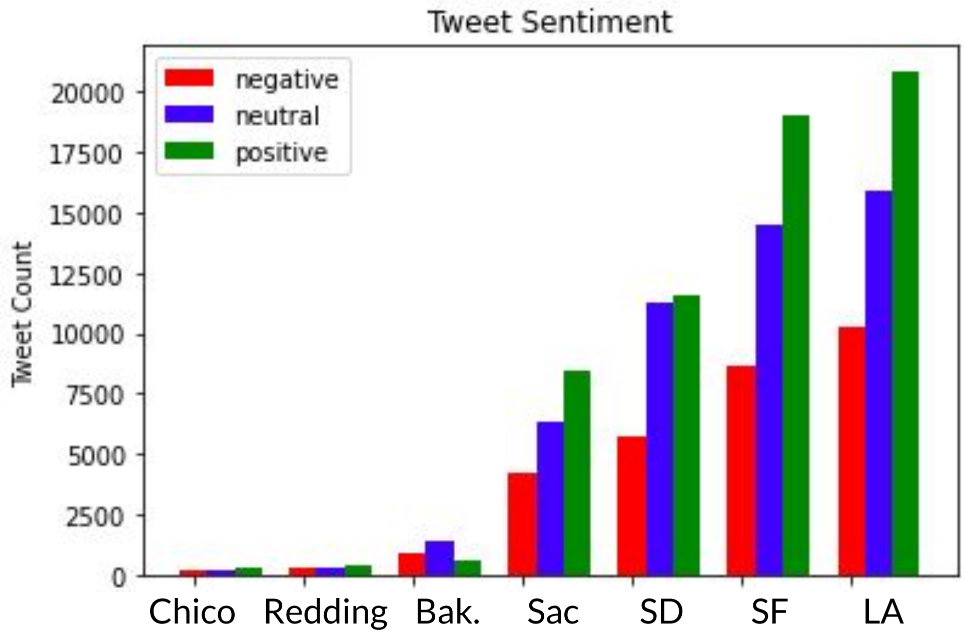
Best Hyperparameters:

max_features=15000
gram_range=(1, 2)
max_iter=1000



	Actual	Predicted	Text
1571	0	1	happy covid wednesday look get float stuffjust...
79437	0	1	friend have conversation test covid despairing...
43748	-1	0	lot available etsy have costume part insane pl...
118412	1	0	brilliant topical relatable substitute journal...
24514	0	1	great thread coronavirus complex system specia...
...	...	...	...
38899	-1	0	givingtuesdaynow emergency response covid time...
1898	1	0	girl not know invite real life birthday party ...
59298	0	1	not believe celebs sing song badly cure covid
104625	-1	0	relief strategy actually help worker small bus...
74681	1	0	stimulate mind kid quarantine stayathome preve...

Model2: NLTK, TextBlob, Logistic Regression



- Labeled Sentiment with TextBlob
    - Simple Interface
    - 3 Labels: Negative, Neutral, Positive
- 75%/25% train test split
- Gridsearched Logistic Regression Model hyparameter C for regularization

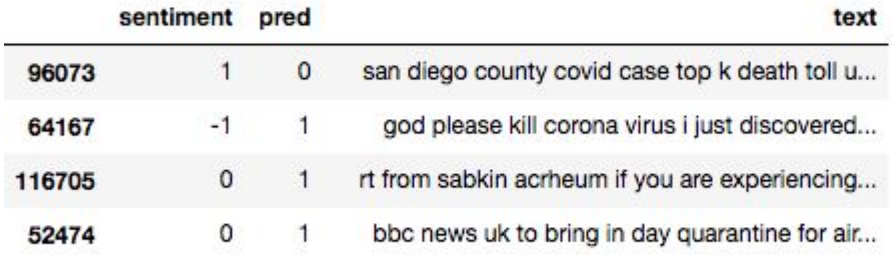
Raw Sentiment Score	
mean	0.07
std	0.25
min/max	-1/1
75th percentile	0.2

Results of Sentiment Analysis Model #2

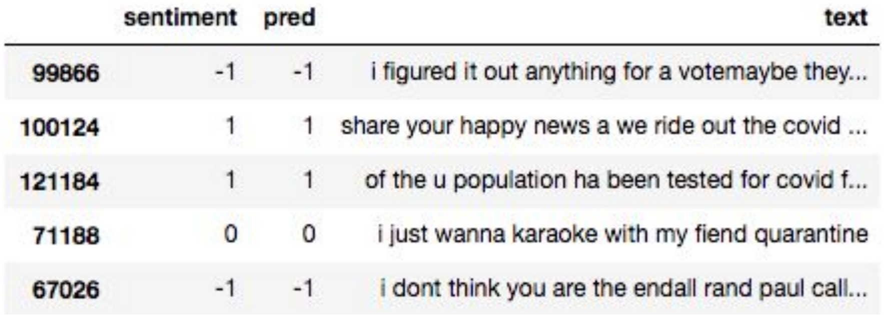
Best crossval score: 0.8456
Best TRAIN score: 0.9667
Best TEST score: 0.8578

- High variance
- Gridsearch chose C = 10

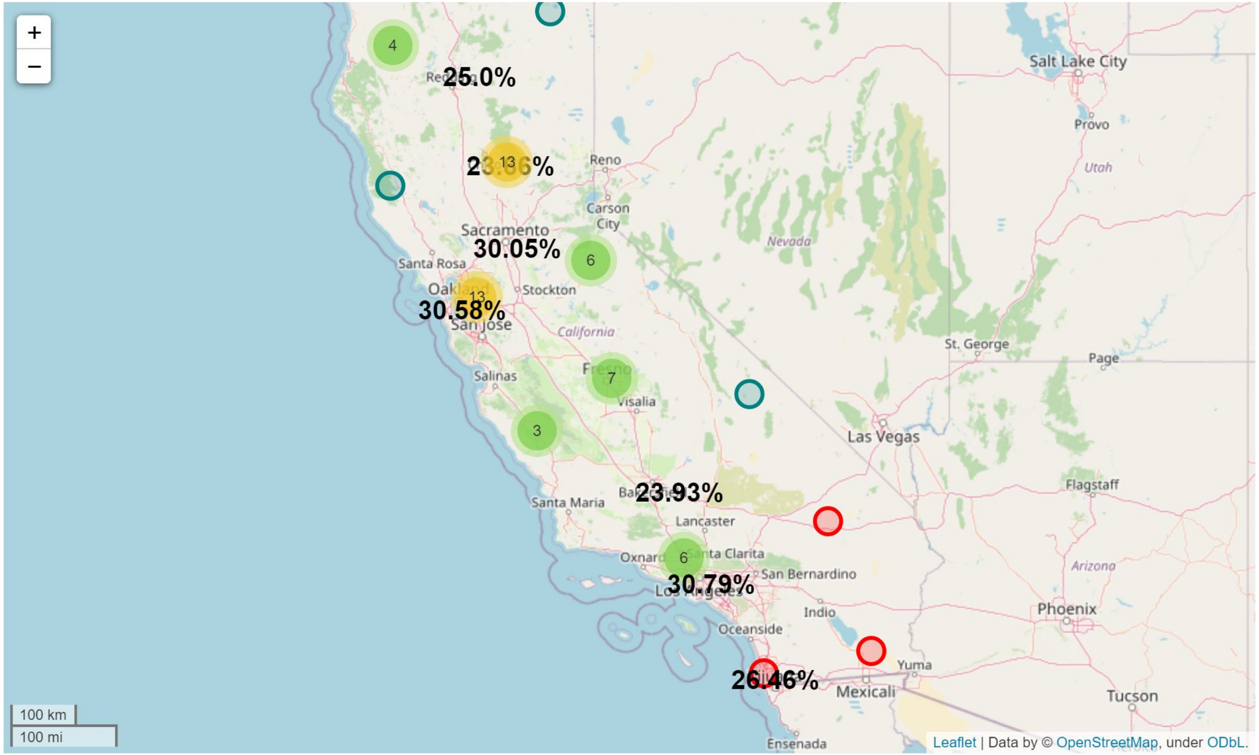
Can improve sentiment labels in original data set and gridsearch through more hyperparameters with more powerful computer



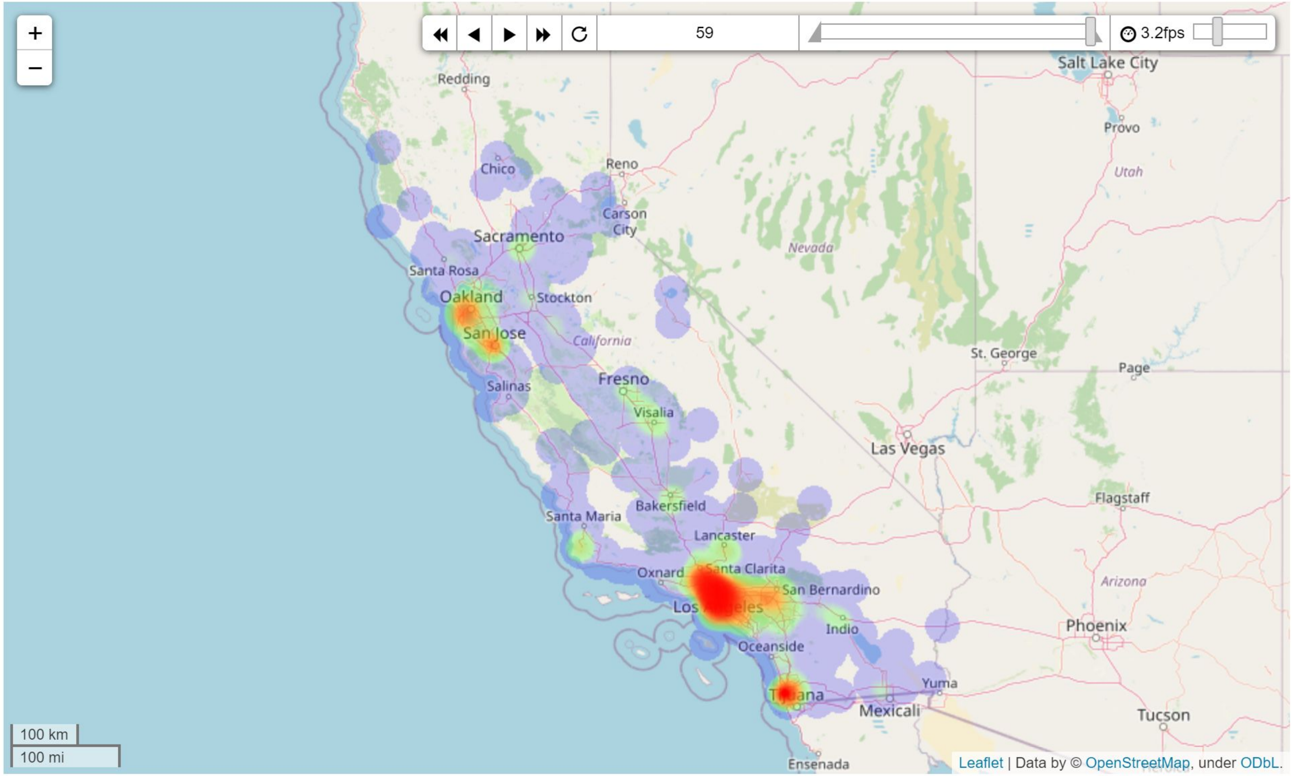
	sentiment	pred	text
96073	1	0	san diego county covid case top k death toll u...
64167	-1	1	god please kill corona virus i just discovered...
116705	0	1	rt from sabkin acrheum if you are experiencing...
52474	0	1	bbc news uk to bring in day quarantine for air...



	sentiment	pred	text
99866	-1	-1	i figured it out anything for a votemaybe they...
100124	1	1	share your happy news a we ride out the covid ...
121184	1	1	of the u population ha been tested for covid f...
71188	0	0	i just wanna karaoke with my fiend quarantine
67026	-1	-1	i dont think you are the endall rand paul call...



Twitter Sentiment Mapped by Region



Timelapse Mapping (3/16 - Present)

Takeaways

- We were unable to find any clear relationships between twitter sentiment and the severity of an outbreak in a region, and thus couldn’t predict the severity of an outbreak

Issues/Further Analysis

-Lots of conflating factors getting in the way of making any assertions regarding twitter sentiment's relationship with coronavirus outcomes

-Data collection issues, need either mathematical tools to make assertions despite said issues, or need to get creative with data collection

-Utilize additional computing power for modelling via an external resource

-Could become very useful with a few tweaks

Questions?