Data Mining Final Project

Corona Virus Heatmap Localization | Comparing Twitter data with Worldometer Data

TEAM MEMBERS:

Megha Babariya
Allen Sun
Varun Sura
Srushti Buddhadev

Professor:

Hui Yang

AGENDA

1. Data Collection:
    a. Worldometer
    b. Twitter
2. Data Preprocessing
3. Sentiment Analysis
    a. Textblob
    b. SentiStrength
    c. Vader
    d. AFINN
    e. Ensemble
4. Emotion Analysis
5. Results and conclusions
6. Future Directions

AGENDA

1. **Data Collection**
2. Data Preprocessing
3. Sentiment Analysis
4. Emotion Analysis
5. Results and conclusions
6. Future Directions

Worldometer Data Collection

- Collects data from https://www.worldometers.info/coronavirus/country/us/
- Uses Beautiful soup library
- Key Challenge:

  Table was modified with new columns over time so generalized script had to be written (No hardcoding).

USA State	Total Cases	New Cases	Total Deaths	New Deaths	Active Cases	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests
USA Total	1,592,723	+22,140	94,936	+1,403	1,127,711	4,812	287	14,117,870
New York	364,249	+1,619	28,758	+110	272,484	18,724	1,478	2,622,544
New Jersey	152,096	+1,082	10,747	+156	134,311	17,124	1,210	531,950
Illinois	100,418	+2,388	4,525	+146	95,782	7,925	357	642,713
Massachusetts	88,970	+1,045	6,066	+128	55,092	12,908	880	489,953
California	85,885	+2,081	3,512	+87	67,171	2,174	89	1,428,360
Pennsylvania	68,151	+724	4,822	+71	56,473	5,323	377	368,743
Michigan	53,009	+659	5,060	+43	19,715	5,308	507	438,565
Texas	51,651	+979	1,423	+21	19,852	1,781	49	788,734
Florida	47,471	+527	2,096	+44	37,737	2,210	98	772,669
Maryland	42,323	+777	2,123	+42	37,394	7,001	351	215,330
Georgia	39,801	+946	1,697	+22	37,764	3,749	160	402,940
Connecticut	39,017	+587	3,529	+57	29,224	10,944	990	190,718

Twitter Data Collection

Overview of the process

What libraries were used ?

How much data was collected (Date period, size) ?

What problems did you run into and how did you solve them ?

Eg: 1) Rate of tweets collection 2) Data size

Tweets Type	Number of tweets	Size of dataset
Original dataset	101,718,655	1 TB
Filtered dataset	15,099,967	3GB

AGENDA

1. Data Collection
2. **Data Preprocessing**
3. Sentiment Analysis
4. Emotion Analysis
5. Results and conclusions
6. Future Directions

Data Preprocessing

Worldometer

- Dropping useless columns

Twitter data

- Converting the locations to state
- Cleaning up text
- Removing missing values

Tweets Type	Number of tweets	Size of dataset
Original dataset	101,718,655	1 TB
Filtered dataset	15,099,967	3GB
Preprocessed dataset	14,067,351	2.5GB

Merging twitter and worldometer data chronologically

Tweet Preprocessing Techniques

Original Tweet

We're Staying Unified In The Fight Against **COVID** 19! 😁Salute To @naviomusic For His Contribution To The National Taskforce & For Supporting Fellow Ugandans.

Read Review To His New Album "Strength In Numbers" Here ~ https://bit.ly/2T42YGi

#SINalbum
#StrengthInNumbers
#UGHipHop

Lower case
Remove URL address
Remove Unicode
Remove stop words
Remove hashtag
Remove numbers
Remove emoticons
Remove punctuation

After Preprocessed

staying unified fight covid salute naviomusic contribution national taskforce supporting fellow ugandansread review new album strength numbers sinalbum strengthinnumbers ughiphop

AGENDA

1. Data Collection
2. Data Preprocessing
3. **Sentiment Analysis**
4. Emotion Analysis
5. Results and conclusions
6. Future Directions

TextBlob

- Naive sentiment analysis algorithm

- Python library that offers a simple API to access its methods and perform basic NLP tasks.

Accuracy = 0.41

	Negative	Neutral	Positive
Precision	0.48	0.42	0.38
Recall	0.21	0.45	0.78
f1_score	0.29	0.43	0.51

SentiStrength Method

- SentiStrength method is used to determine both, positive scores and negative scores of a particular set of text.
- Calculated using SentiStrength software with our own Ground truth data.
- Calculated Accuracy by comparing the rating given by a human-opinion(by us) with that of rating given by the SentiStrength Algorithm.
- Accuracy = 0.46
- Precision = 0.35
- Recall = 0.56
- F1-Score = 0.43

Lexical analysis based algorithms

1. Vader (Valence Aware Dictionary & Sentiment Reasoner)

    a. Lexicon and Rule based sentiment analysis tool.
    b. Popular for Twitter Data
    c. VADER uses a combination of a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as either positive or negative.

2. AFINN

    a. Corpus of English words that are manually rated for Valence with an integer between -5 and 5.
    b. This technique consists of pre-labelled words (2477) classified as +ve , -ve and Neutral.
    c. The comparative score is then calculated based on the total score obtained and the total number of Tweets.

Vader

Accuracy = 0.55

	Negative	Neutral	Positive
Precision	0.69	0.5	0.46
Recall	0.5	0.45	0.78
f1_score	0.58	0.47	0.58

Afinn

Accuracy = 0.58

	Negative	Neutral	Positive
Precision	0.72	0.54	0.46
Recall	0.58	0.48	0.7
f1_score	0.64	0.51	0.55

Ensemble Method

- Combines the first 3 algorithms. Final score is calculated using:
  - 20% weightage to text blob score
  - 40% weightage to afinn score
  - 40% weightage to vader score

- Accuracy = 0.54

	Negative	Neutral	Positive
Precision	0.69	0.54	0.42
Recall	0.46	0.52	0.74
f1_score	0.55	0.53	0.54

Comparing all algorithms

- Accuracy order: Afinn (0.58) > Vader ~ Ensemble (0.55) >> sentistrength(0.46)>TextBlob (0.41)
- F1-score order: Afinn > Vader > Ensemble > sentistrength> TextBlob

ML Models	Accuracy	Weighted-F1	Macro-F1	Micro-F1
afinn	0.58	0.57	0.58	0.58
vader	0.55	0.54	0.55	0.55
text_blob	0.41	0.41	0.41	0.38
sentistrength	0.46	0.28	0.46	0.45
ensemble	0.54	0.53	0.54	0.54

- Overall Afinn is the best.

	Negative	Neutral	Positive
Precision	0.72 (Afinn)	0.54 (Afinn, Ensemble)	0.46 (Afinn, Vader)
Recall	0.58 (Afinn)	0.52 (Ensemble)	0.78 (Vader, TextBlob)
F1-Score	0.64 (Afinn)	0.53 (Ensemble)	0.58 (Vader)

- Class-wise performance breakdown:
  - Afinn best for negative predictions
  - Ensemble best for neutral predictions
  - Vader best for positive predictions

AGENDA

1. Data Collection
2. Data Preprocessing
3. Sentiment Analysis
4. **Emotion Analysis**
5. Results and conclusions
6. Future Directions

Emotion Analysis

- The tweets collected can be better classified using emotion analysis.
- Uses lists of keywords which can be used to classify the emotion.
- Preprocessing includes making text lowercase, removing the symbols, removing the stop words.
- Classifies tweets as happy, sad, frustrated, or neutral.

```
In [9]: # determing emotion using check_emotion function
        data['emotion'] = data['text'].apply(check_emotion)
        data.head()
```

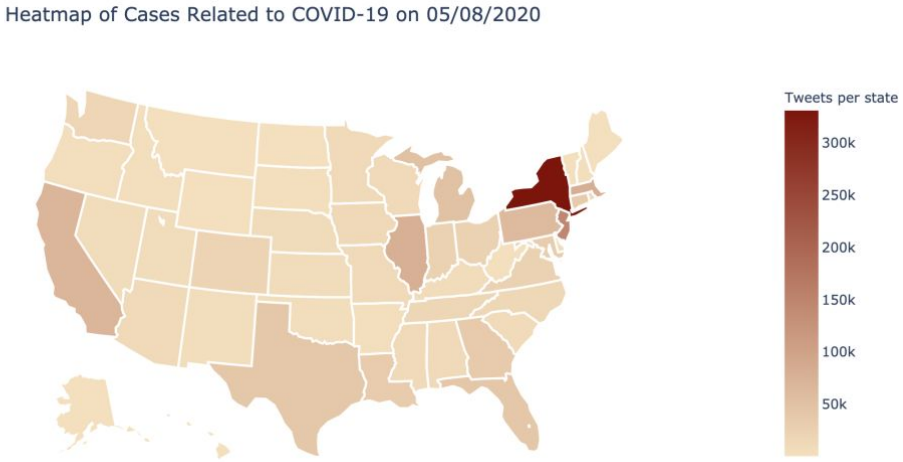
Out[9]:

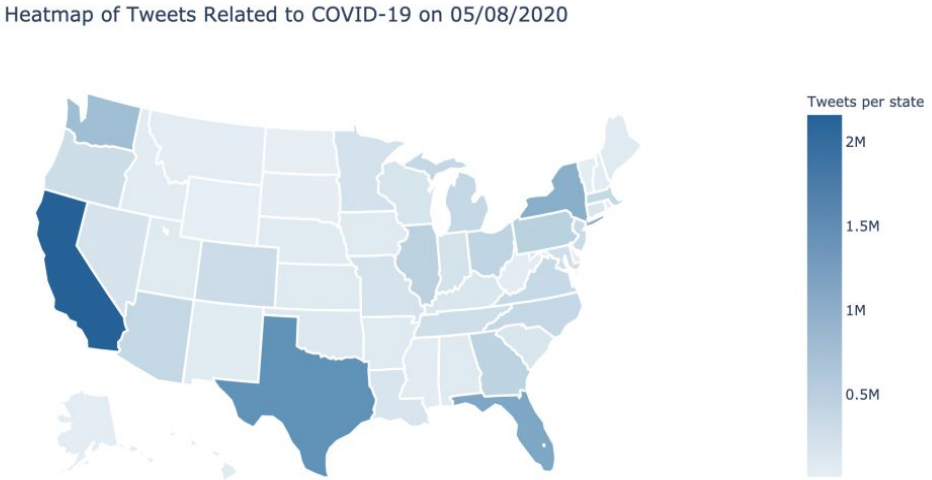
	text	emotion
0	coronavirus panel recommends use hydroxychloro...	neutral
1	swear covid stuff going back church wanted pas...	happy
2	retweet information injustice toward black peo...	sad
3	police raids syracuse mosque continues holding...	happy
4	half us covid 19 deaths wyoming would nyc will...	neutral

AGENDA

1. Data Collection
2. Data Preprocessing
3. Sentiment Analysis
4. Emotion Analysis
5. **Results and conclusions**
6. Future Directions

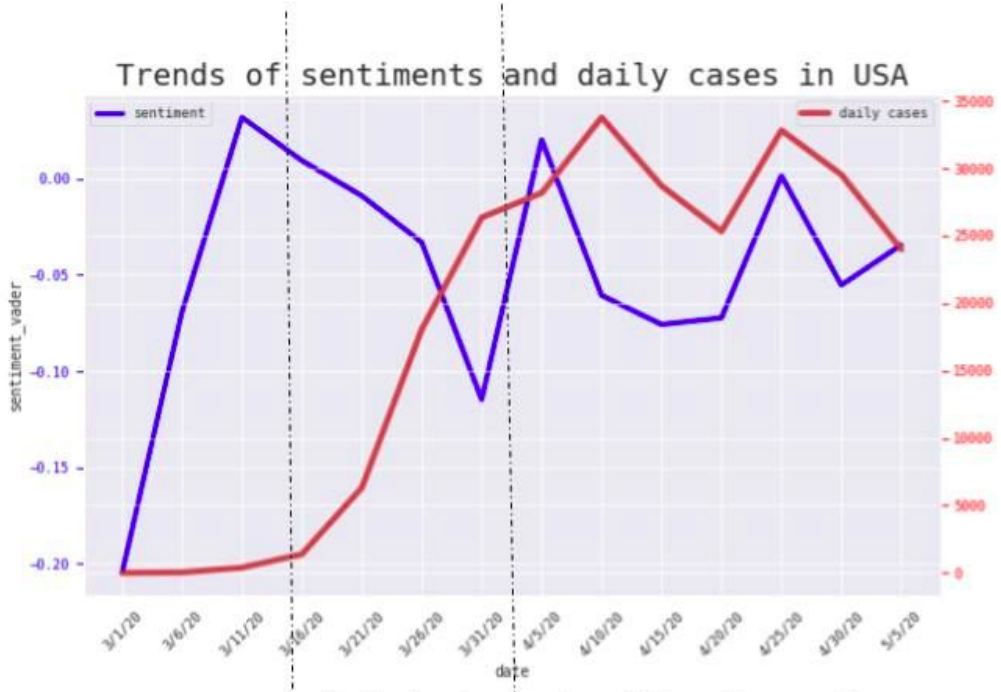
Heatmap: Tweets vs Cases





Conclusion: There is a strong correlation between number of tweets and number of cases in a state with some exceptions such as Louisiana (right of Texas).

Correlation: Daily Cases vs Daily Sentiment



Not many cases initially.

Sentiment is high.

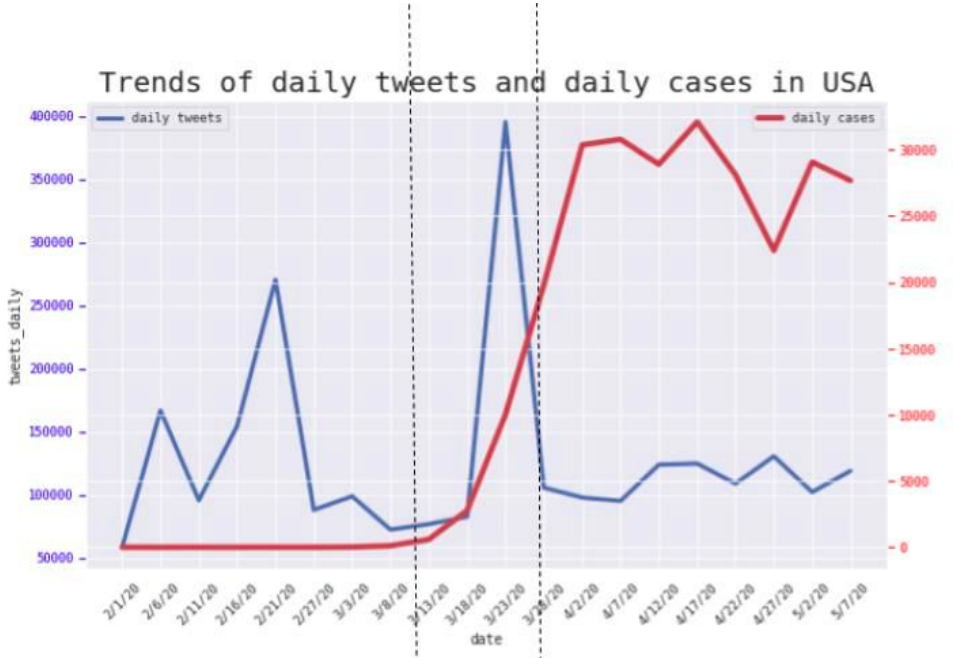
Sudden burst of cases.

Sentiment shoots down,

Curve flattens. Cases go down.

Sentiment starts increasing again.

Correlation: Daily Cases vs Daily Tweets



Initial stage:
Curiosity drives #Tweets high
#Cases is low

Crucial stage:
#Cases shoots high
#Tweets also shoots high

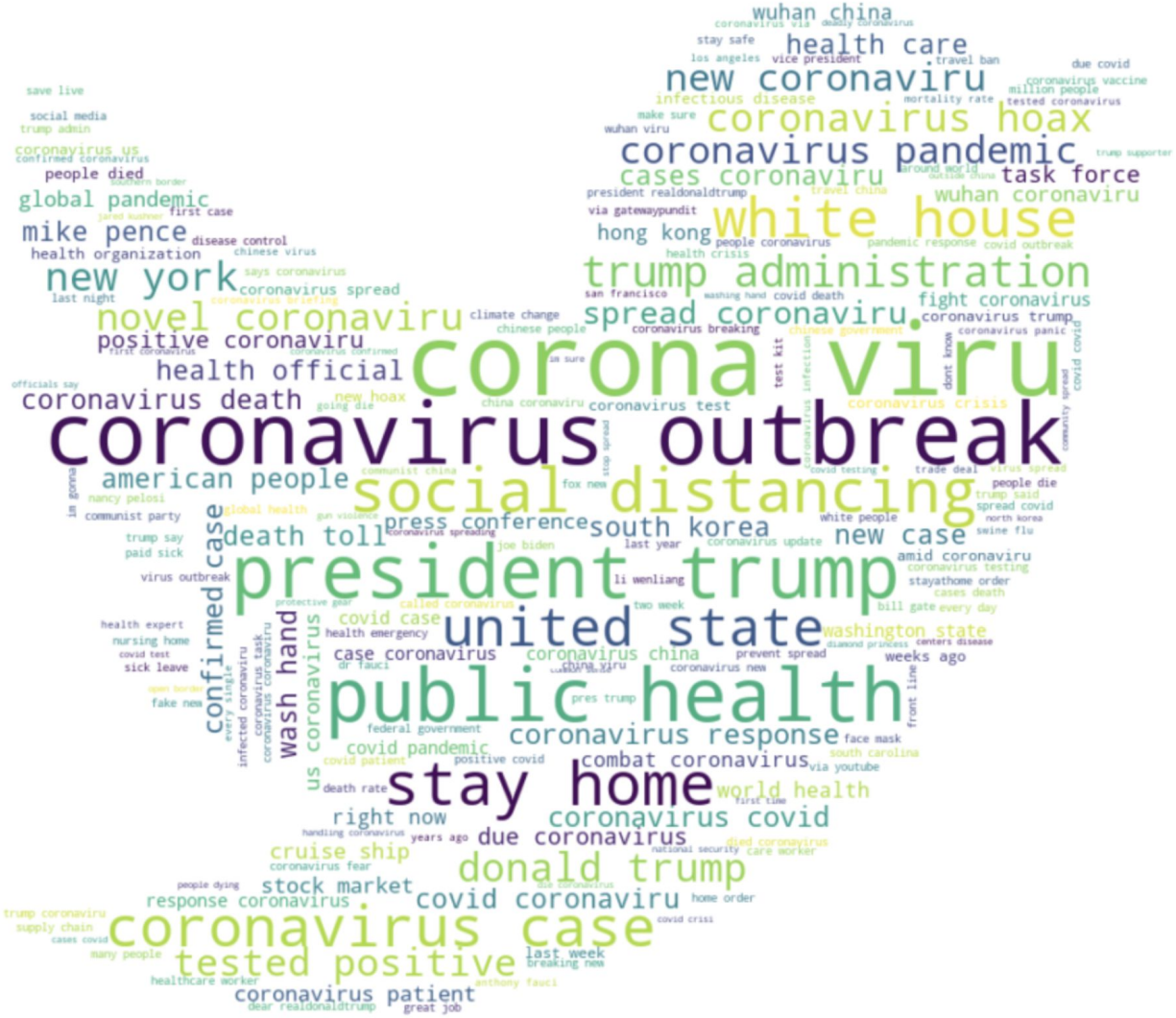
Final stage:
#Cases becomes constant
#Tweets becomes constant

AGENDA

1. Data Collection
2. Data Preprocessing
3. Sentiment Analysis
4. Emotion Analysis
5. Results and conclusions
6. **Future Directions**

Future Work:

- Can improve the accuracy of the emotion analysis by analysing more datasets for keywords.
- Can try supervised model by using a larger curated dataset for training.
- Can try using clustering techniques to improve the emotion analysis process.



Thank you!
Questions ?