Coronavirus Tweets Text Mining

Chang Shen
supervisor: Robert Mcdougal

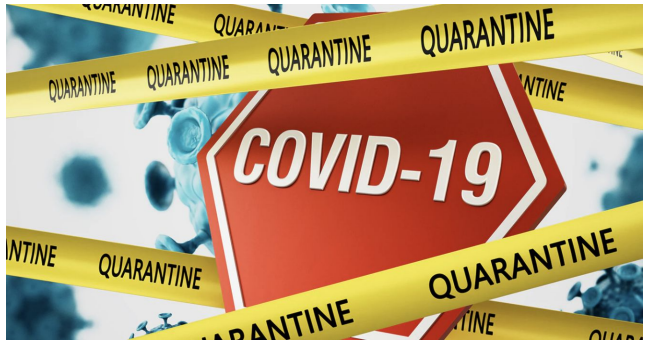
Contents

Social media and coronavirus

1.Background

Coronavirus and social media

COVID-19
Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).
It has spread worldwide, leading to an ongoing pandemic since Dec 2019.



Social Media
Since there wasn't any proved effective medicine or vaccine back then, CDC initiated a quarantine guideline.
Social Media became an essential resources for people to communicate with each other.
Remote work and learning increased the anxiety in public .

Research questions

Whether people's attitude(sentiment) to COVID-19 has changed from March to April?

When people talk about the COVID-19, what topics they care about?

Can we label the sentiment of covid relateds tweet automatically with the model trained by the coronavirus tweets NLP dataset?

If there any way we can map coronavirus tweets to the location sent? Conduct geographical wise text analysis?

what our data looks like

2. Data description

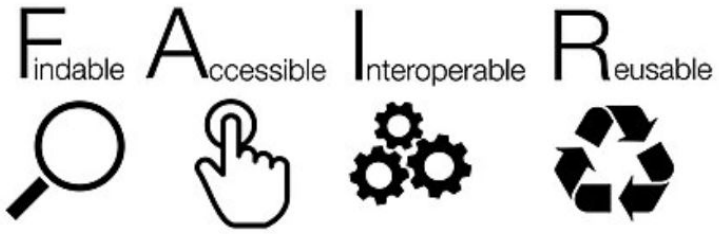
Resources and FAIRness

Data Resource
Coronavirus tweets text data from Kaggle public data Repositories. See the data source here Coronavirus tweet NLP dataset.

How they collected the data
Collected from Twitter stream with a purpose of scientific research and all the confidential information(name/contact/demographic) was removed from the dataset. The metadata of the data is also available and licensed on Kaggle.

Data format
It's stored in .csv format

FAIRness
This dataset is following the FAIRness principle



Data description

Data Overview

We have 41,157 tweets in total posted from 2020-03-16 to 2020-04-14, labelled with 5 different sentiments.

Variable	Description	Type	Example
UserName	The user ids, correspond to information of different users.	Integer	3799
ScreenName	The user ids, correspond to different users' twitter names.	Integer	48751
Location	Where did the twitter sent from, input by users. Can be country level/state level/random sentence	String	London
TweetAt	The date a user posted the twitter in yyyy-mm-dd format	Date	2020-03-16
OriginalTweet	The twitter text - free text	String	Was at the supermarket today. Didn't buy toilet paper. #Rebel#toiletpapercrisis #covid_19 https://t.co/eVXkQLIdAZ
Sentiment	The sentiment labelled manually, catgorical variable with 5 classes "Neutral" , "Positive", "Extremely Negative", "Negative", "Extremely Positive"	String	Neutral

Make computer understand the meaning of the words

3. Exploratory Data Analysis

Data preprocessing

Missing data
Only Location column contains missing data, the missing rate is 20.87%
Might be **self selection bias**, can't impute with the mode. Code it as a category "Unknown"

Standardization and data correction/deduplication
We realized that in original data set the location columns contains various location granularity and not in standard format(states/country/coordinate/joke).

There are 12,220 unique locations in 41,157 entries, among them 9,406 only appear once in our data.

First challenge !
"Outside, standing on a bucket"
"Our one precious planet"
"OMG am I the fattest one here"
"worldwide"
"Salt Lake City, UT (But left my heart in San Francisco)"

Data preprocessing

Text processing

U.S.A -> usa
car, cars, car's, cars' -> car

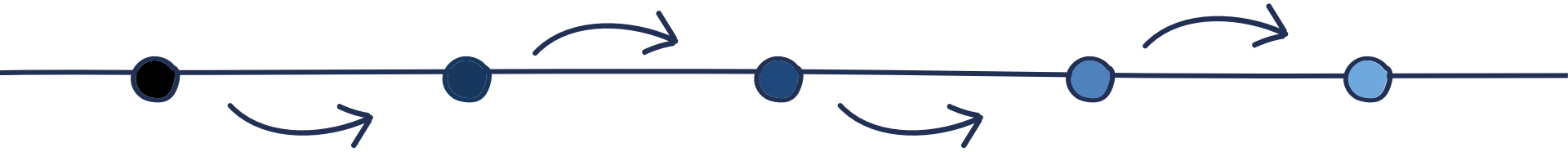
to lowercase/remove punctuation/get stem

超市-> “ ”

remove non-english

@bis634 -https://bis634/->””
stay safe #covid -> covid

Extract hashtags and mentions as a new variable and remove url



Remove stopwords, i.e. "it", "can", "we"

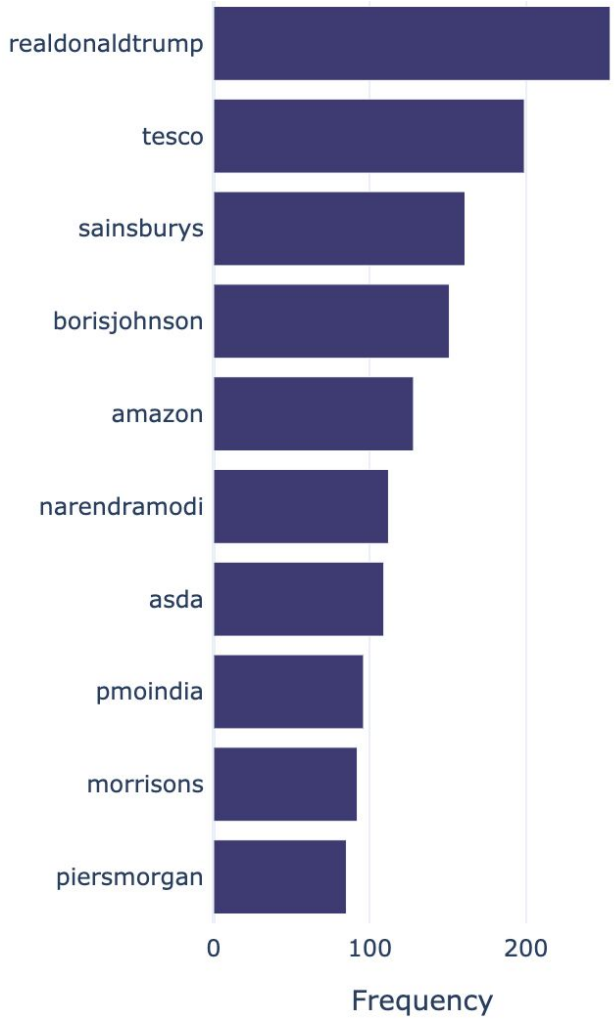
we can success -> success

Tokenize

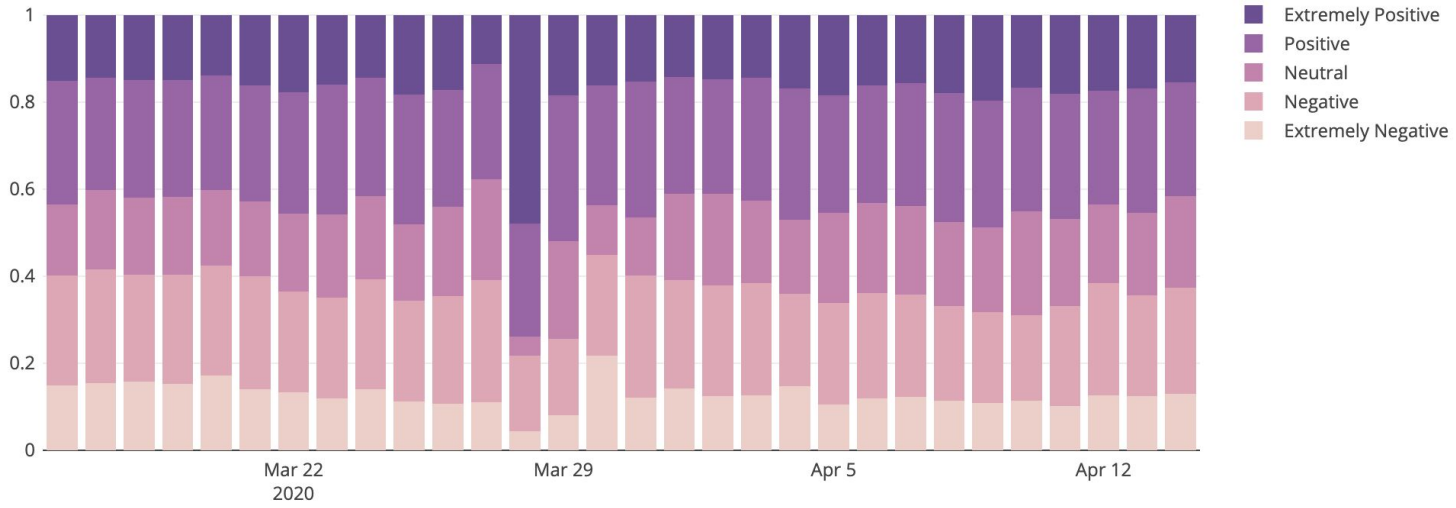
“corvid change life” -> “covid”, “change”, “life”

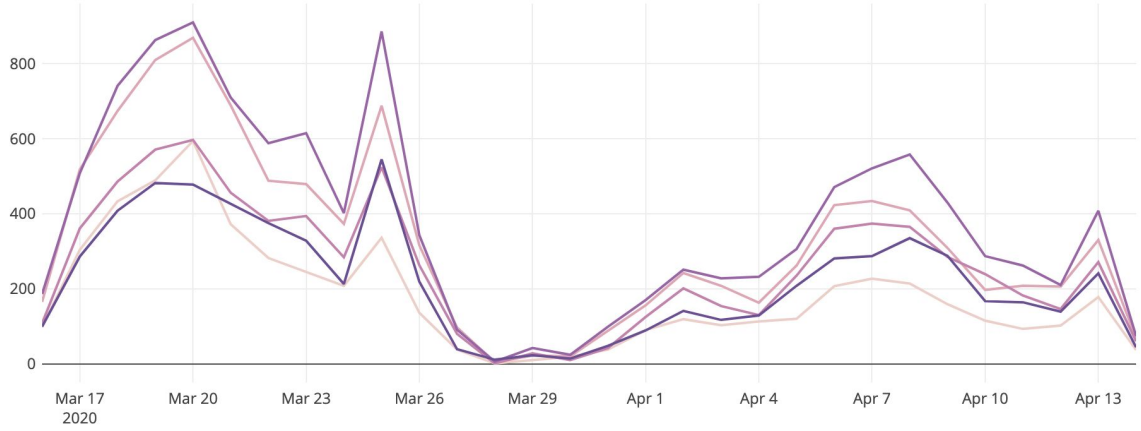
Data preprocessing

Top 10 locations where the tweets came from



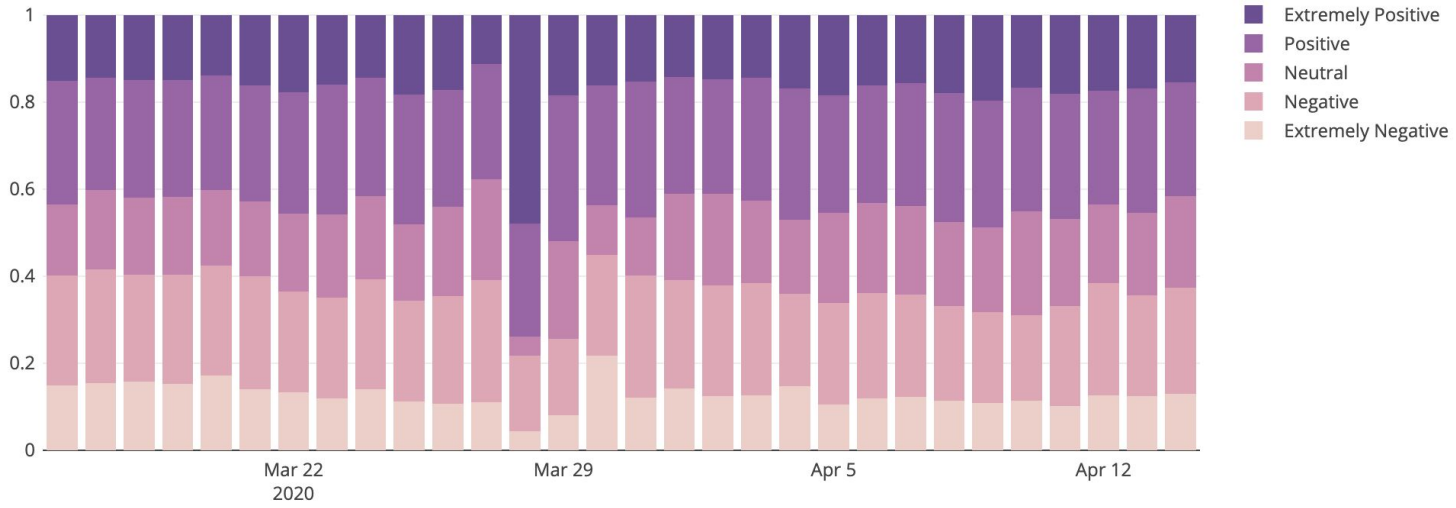
Sentiment distribution

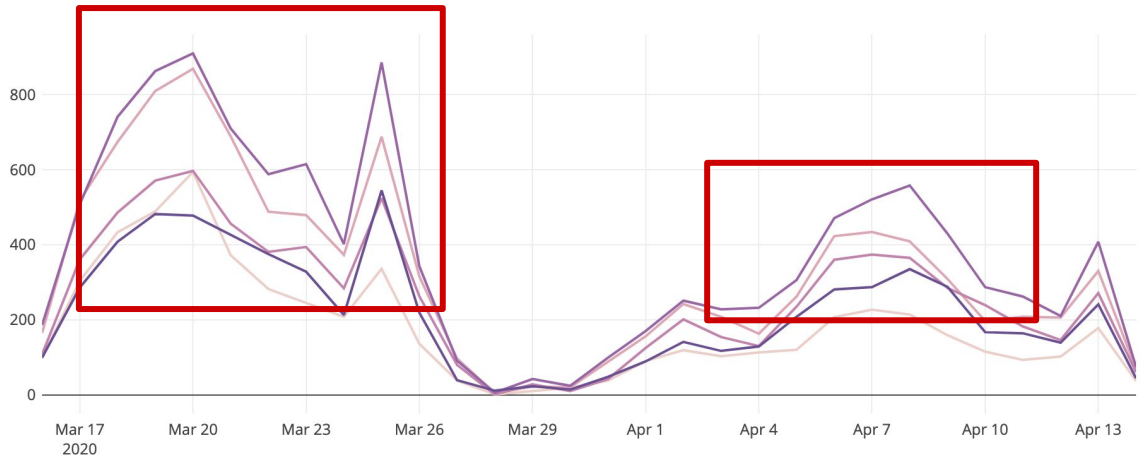




Sentiment	counts
Positive	11422
Negative	9917
Neutral	7713
Extremely Positive	6624
Extremely Negative	5481

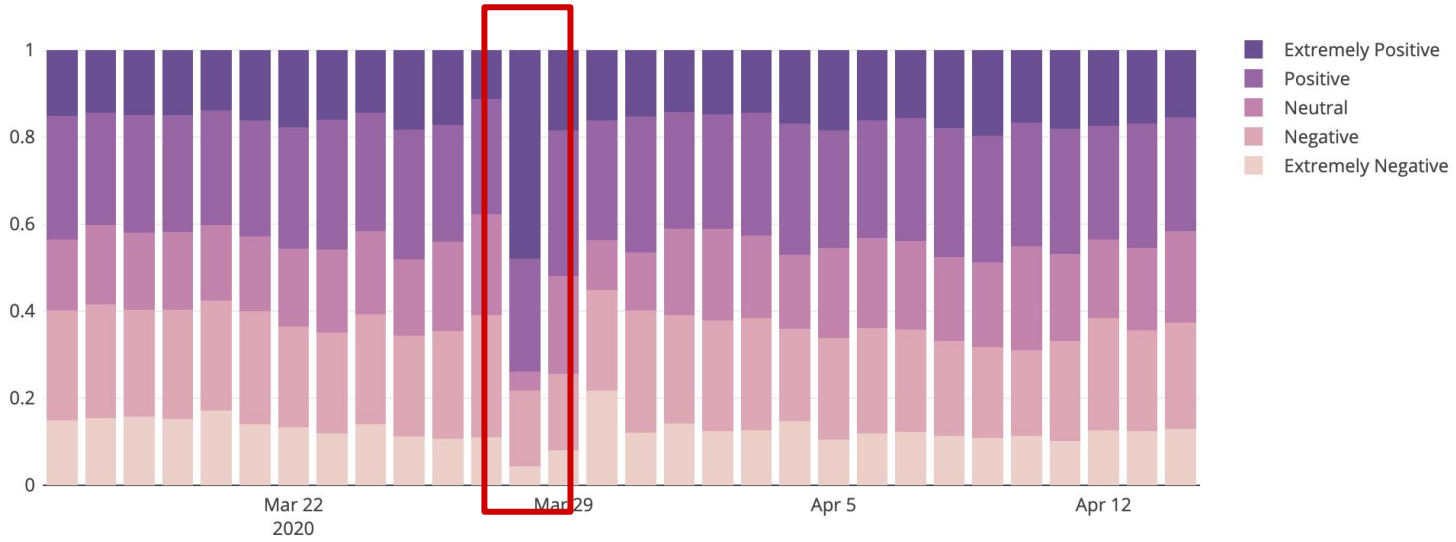
Sentiment distribution

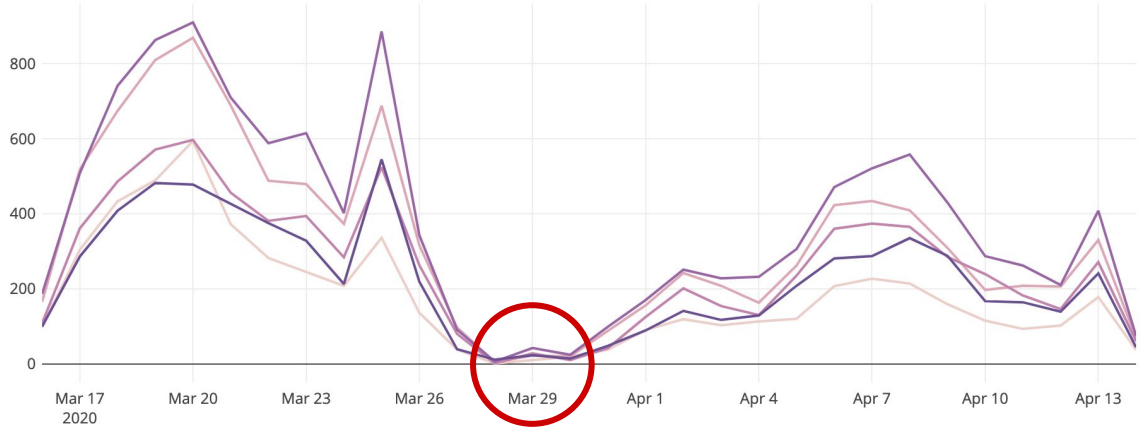




Sentiment	counts
Positive	11422
Negative	9917
Neutral	7713
Extremely Positive	6624
Extremely Negative	5481

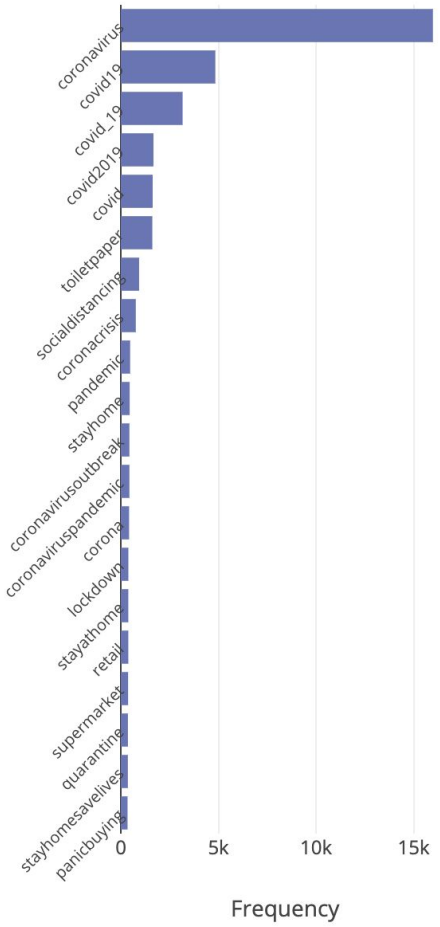
Sentiment distribution

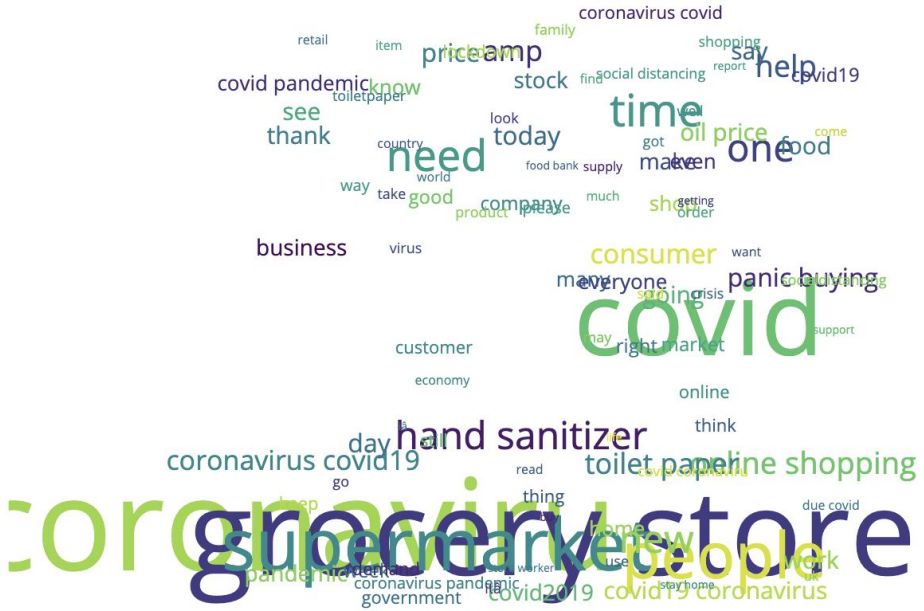




Sentiment	counts
Positive	11422
Negative	9917
Neutral	7713
Extremely Positive	6624
Extremely Negative	5481

Wordcloud and hashtag





Build the sentiment text classification models

4. Modelling and API

Methods

LDA

To further explore the topics in the coronavirus tweets. And the similarity between topics

Bert

Bert sequence text classification model.

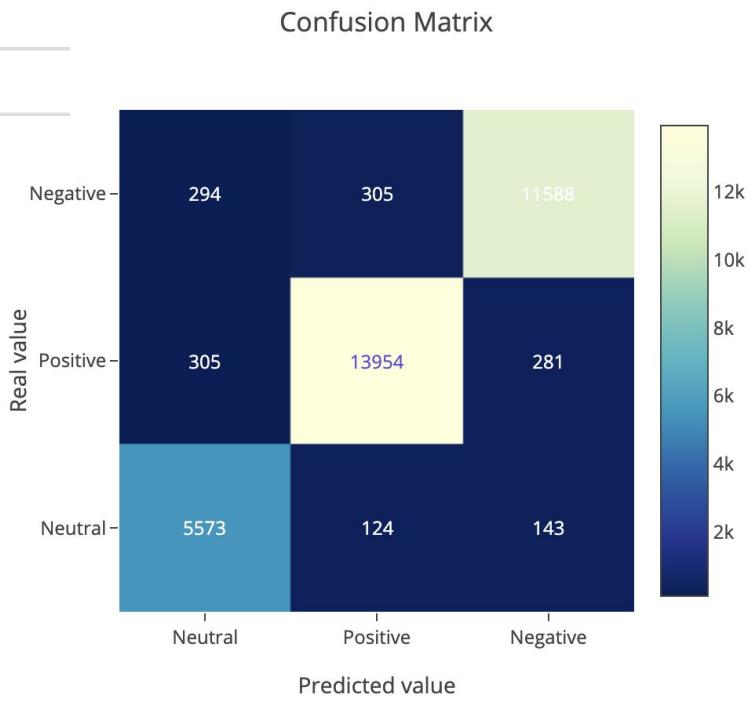
Build on pretrained bert has high accuracy.

Bert

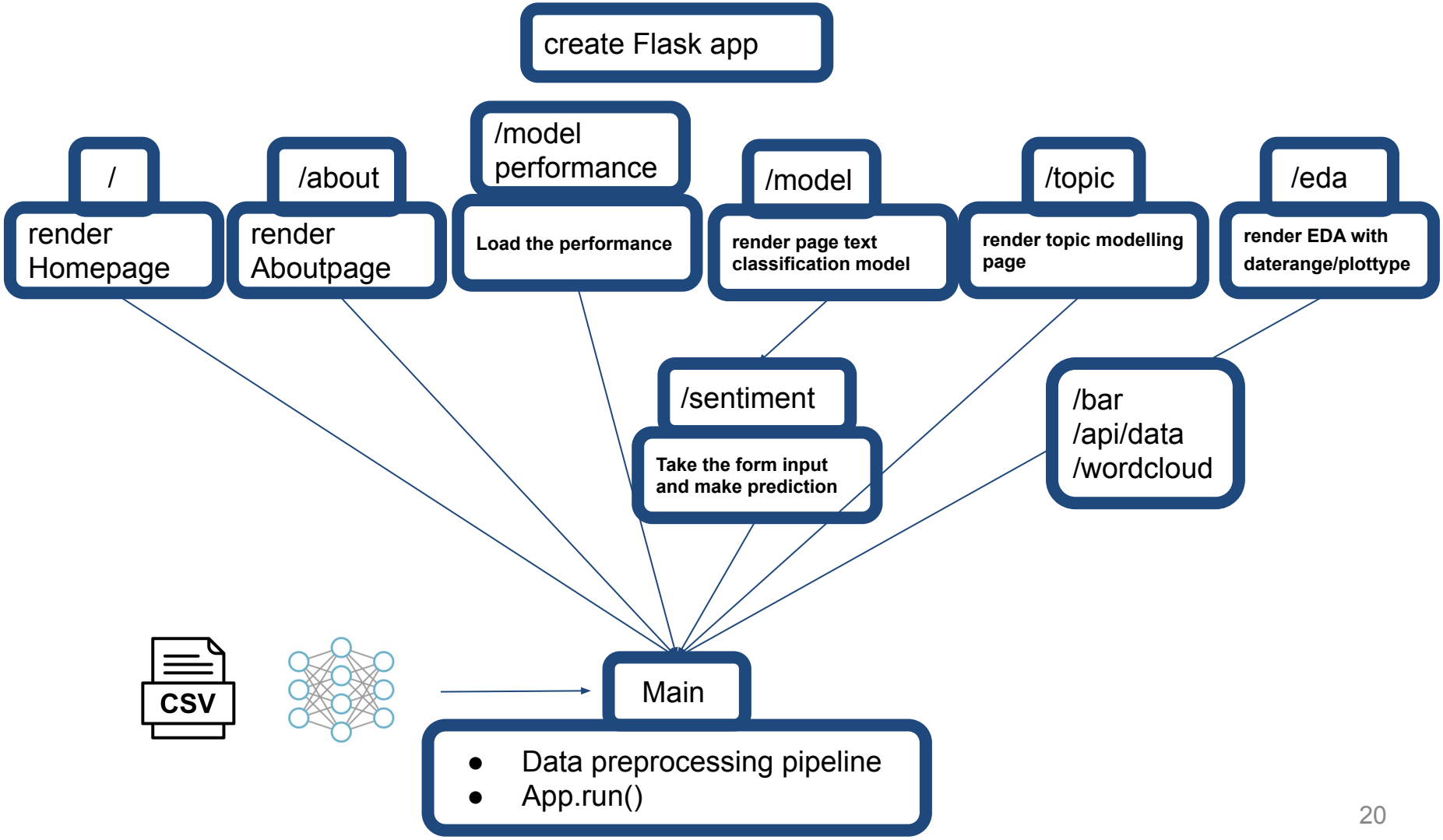
Parameters setting

Parameter	Value
Model	Bert For Sequence Classification
Optimizer	Adam
Learing Rate	2e-5
Epoch	4
Batch	64
Train:Validate	9:1

Test accuracy: 90.1%



API server


create Flask app
/
render Homepage
/about
render Aboutpage
/model performance
Load the performance
/model
render page text classification model
/topic
render topic modelling page
/eda
render EDA with daterange/plottype
/sentiment
Take the form input and make prediction
/bar
/api/data
/wordcloud
csv
Main
Data preprocessing pipeline
App.run()

Frontend

https://bis634.herokuapp.com/

Potential next steps and what I learned/superising

5. Takeaways

Takeaways

Interesting Findings

1. People's topics in March were around supermarket/toiletpaper while in April we focus on the social distancing more
2. People like to mention @realdonaldtrump when they send COVID-19 related tweets.
3. The sentiment distribution at the end of March is very different and # of tweet was 1/10 of 20th, March.

Difficulties

1. Location variable is too dirty and not applicable
2. Tweets are different from other text. Need to deal with the url/mentions/hashtags
3. Find the reason for the abnormal data is difficult.
4. API need to load large data, take times

Future work

1. We can conduct sentence embedding and dimension reduction to evaluate the similarity of two tweets.
2. Can improve the API and the frontend to make it more efficient.

Study objective

Coronavirus Tweets Text Mining and Sentiment Analysis

Data analysis on COVID-19 tweets. Topic trending and Sentiment distribution

Apply machine learning models to classify the tweets sentiment and detect negative emotions

create a public interactive visualization tool to present the analysis

Latent Dirichlet Allocation

Hashtags is a way to identify the topics in tweets.

Topic Modelling

1. Statistical model for discovering the abstract "topics" that occur in a collection of documents.
2. Unsupervised + Bayesian 20% from topic A and 80% from Topic

See the topic modelling result for k=5 demo here