Yale University

Coronavirus Tweets NLP Analysis

Chang Shen

Supervisor: Robert McDougal

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first case was identified in Wuhan, China, in December 2019[6]. The relevance of the COVID-19 global pandemic in Mar, 2020 has raised the attention of researchers all over the world.

Since there wasn't any proved effective medicine or vaccine back then, CDC initiated a quarantine guideline for COVID-19 prevention. Most of the schooling and work were moved online until now. Social Media became an essential resources for people to communicate with each other. It certainly made our lives convenient. However, it's a double-edged sword. Misinformation, fake news and extreme emotion can mislead the public. In a way, remote work and learning increased the anxiety in public because of the social media.

One of largest online public communities Twitter is an important resource for people to learn about this new virus and communicate with each other. It's a very significant to detect the negative sentiment, rumors and misinformation from those tweets to avoid the public anxiety and the topic trending during the pandemic. This research is meant to be a social behavior study focusing on the Coronavirus Tweets Text Mining and Sentiment Analysis and will be flexible to be transferred to other social media text datasets, representing and communicating the information with the public.

1.2 Data resources and FAIRness

The dataset I am using it's a coronavirus tweets text data from **kaggle public data Repositories**. See the data source here Coronavirus tweets NLP - Text Classification. It was collected from **Twitter stream** with a purpose of scientific use and all the confidential information was removed from the dataset. The data was stored in **.csv** format. The metadata of the dataset is also available and licensed on Kaggle. The sentiment is manually labelled by Aman Miglani.

FAIRness This dataset is following the FAIRness principle:

1. **Findability**: It's a public datasets, thus the data and metadata can be easily found by anyone, searching on Kaggle public dataset with key word text classification and tweets coronavirus.

2. **Accessibility**: Since it's a public datasets, no permission need to request. Data and metadata can both be easily downloaded from Kaggle API.

2. **Interoperability**: The data is stored in .csv format and the meta data is stored in json. Since it's social media data, it doesn't have any vocabularies and ontology related to it.

2. **Reusability**: This data is described with a clear and accessible data usage license with file descriptions, column descriptions and license specified. The provenance is provided(Twitter).

1.3 Research Questions

According to the dataset we have, some potential research questions to ask are

- Whether people's attitude(sentiment) to COVID-19 has changed from March to April?
- When people talks about the COVID-19, what topics they care about?
- Can we label the sentiment of covid related tweets automatically with the model trained by the coronavirus tweets NLP dataset?
- If there any way we can map coronavirus tweet to it sender location? Conduct geographical wise text analysis?

2 Exploratory Data Analysis

2.1 Original Data Statistics and Manipulation

2.1.1 Data Overview

Size and format: The data is stored in .csv format and with 41,157 entries, with a size of 10MB.
The data has 6 variables, including:

- **UserName**: The user ids, correspond to different users' real word names.
- **ScreenName**: The user ids, correspond to different users' twitter names.
- **Location**: Where were the tweets sent from, ranging from states in United States to other countries like Canada, Sydney,Australia
- **TweetAt**: The date a user posted the tweets in yyyy-mm-dd format.
- **OriginalTweet**: The tweets text - free text, unconstructured data.
- **Sentiment** The sentiment labelled manually, catgorical variable with 5 classes ”Neutral” , ”Positive”, ”Extremely Negative”, ”Negative”, ”Extremely Positive”

Potential Information, the orignal dataset already has many useful datasets specified(location/date, etc), however there are still other important variable need to be derived, for example, we need to extract the the hashtag data from the OriginalTweet variable to know the topic for this twitter and probability some Part of Speech tags need to be applied and besides it may require us to extract the emoji information and conduct name entity recognition to get more information for modelling.

2.1.2 Data Preprocessing

There are a lot of data processing work need to be done before starting to analyze the data.

- **Missing data**
  There are some missing data in the original dataset, need to be dealt with, in **Location** variable, the missing rate is more than 20%. Since in social media the choice of not revealing the location can be a kind of self selection bias, so the missing data mechanism is not MCAR. If imputing the data with mode/model prediction, we may lose information the user behavior. Thus here, instead of any imputation, we code the missing location as a new category ”Unknown”.
- **Standardization and data correction**
  For the location variable, the data isn't in standard format for example, in some entries, it only shows the country, and not in standard format(USA, United States, etc), while for other entries, it also shows the city/states of the country like LA, USA. And since it can be customized by the user, there are some weird location like ”outdoor, standing on the bucket”.
  We applied regex expression split the city and country and map the information. However, the usable location information is still limited and not a good variable to analyze on, there was my first challenge when I tried to analyze the text data. The most frequent locations showed in Figure 1.
  I also corrected the mis-recorded information like 28-03-2021 for Tweet date.
- **Text Mining**
  For text data, the messy nature makes it difficult to process, the strategy we used here is:

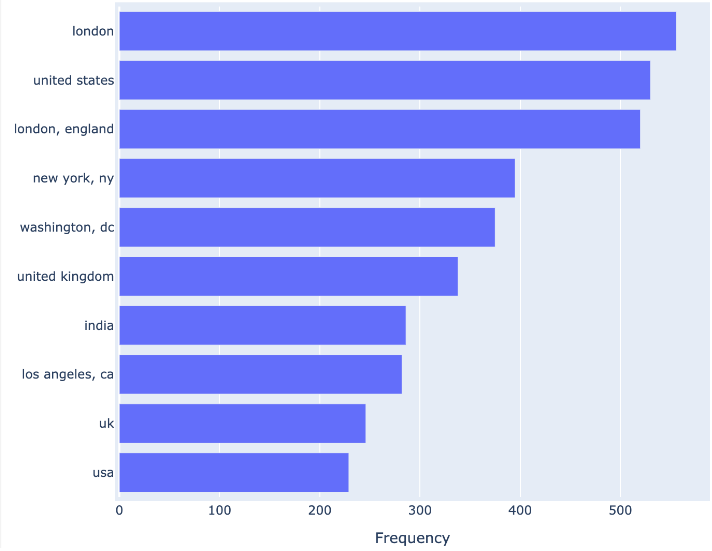


Figure 1: *Where the tweets came from top 10 locations frequency*

- Standardize to lowercase/remove punctuation/get stem
- Remove meaningless stopwords like ”it”, ”can”, ”we”
- Remove non English Tweets, since most of the tweets were sent in English, we decided to focus on the English tweets to make sure the data cleanness.
- Tokenize the sentence
- Lemmatize each token
- Analyze on N-gram
- Extract the hashtag and mentions, remove the url

2.1.3 Descriptive Statistics

Here are some essential statistics from our datasets, since the main variable we used is free text, the detailed statistics of it will be discussed in the next section.

Variable	# unique	# Missing	Mode
UserName	41157	0	NA
ScreenName	41157	0	NA
Location	12220	8590	london
TweetAt	30	0	2020-03-20
OriginalTweet	40748	0	*NA*
Sentiment	5	0	Positive
Hash	15398	0	coronavirus
Mention	10323	0	realdonaldtrump

Table 1: *Descriptive statistics*

In summary, with the data preprocessing, we acquired a data of better quality and a better understanding of the tweets dataset we have. As we can see in the descriptive statistics:

1. Just as our previous guess, the **UserName** and **ScreenName** are just random user identifiers, no more interestinng information in it. Thus, this two variables would not be applied in the following analysis part.
2. For **Location** variable, at the begining, when I accessed the dataset, my intention was to conduct a geographical analysis on tweets and the tweets' sentiments and compare with the COVID-19 case at different states. However, after several of attempts and failures I finally give up on this plan. The Location information in social media posting is not applicable and also wouldn't be included in the following analysis.
3. The tweets collected is ranging from 2020-03-16 to 2020-04-14, 30 days in total. While in 2020-03-20, the number of tweets arrived peak among the 30 days. The time series of the tweets will be presented and analyzed in the following sections.
4. The sentiments is distributed almost evenly in the coronavirus tweets datset we have.
5. For **Hash** and **Mention**, there are 15,398 unique hashtags and 10,323 unique mentions in our original tweets. One of the interesting finding is for all the tweets, the user "realdonaldtrump" was mentioned most, with a frequency of 254. Other frequently mentioned users are tesco, sainsburys, borisjohnson, amazon, narendramodi, asda. Either political characters or supermarket companies.

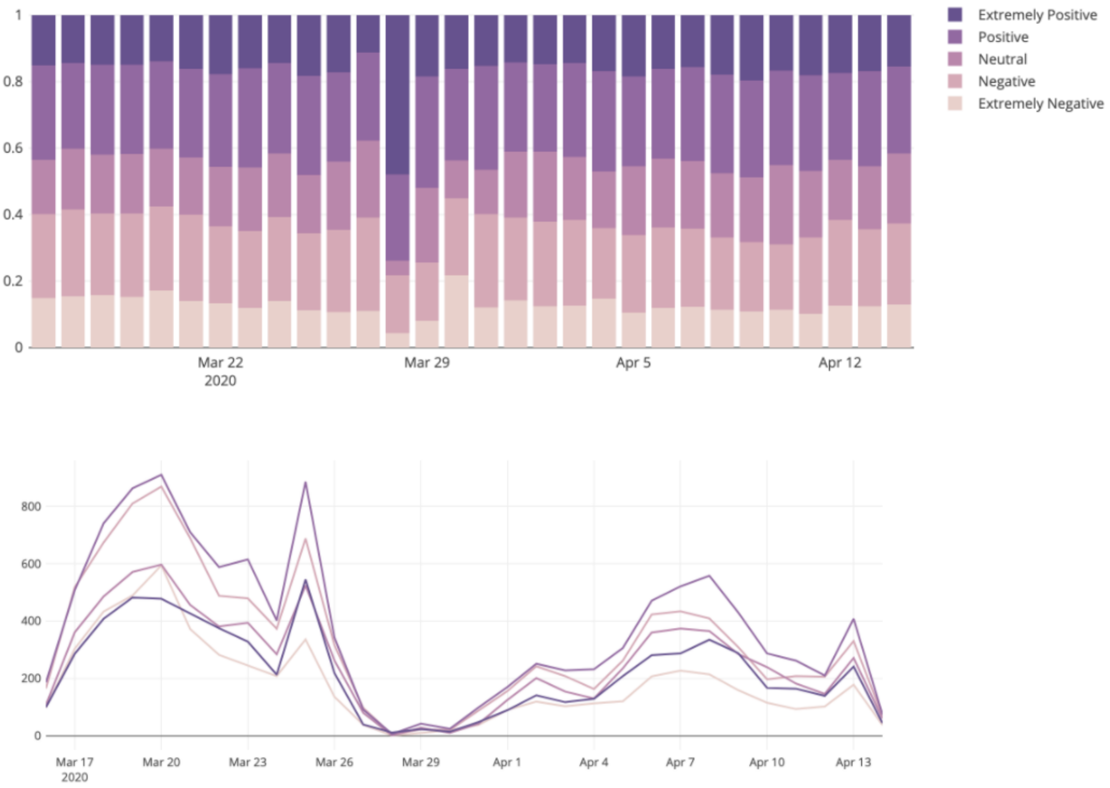


Figure 2: *The sentiment distribution change over time*

2.2 Sentiment Analysis

2.2.1 Sentiment distribution over time

To dig deeper on the sentiment distribution time series, I summarized the data and made two visualizations as above.

There are some interesting findings for the sentiment distribution over time

· The sentiment distribution of the tweets are very stable, the proportion of positive, neutral and negative sentiments are very similar(around 25% each), while overall the proportion of positive sentiments is higher.
· The # of tweets on March is much higher than the # of tweets in April. This might because, for March, it was the time when COVID-19 became a worldwide pandemic. As a unknown type of disease, the discussion around coronavirus peaked in March and in April as people got to know more about the unknown disease and how to prevent it, the number of tweets around coronavirus decreased during April.
· One surprising finding from the bar plot and the line chart was, in 28th, Mar to 1st, Apr, the number of tweets about coronavirus decreased rapidly and the positive tweets dominated at these days. Firstly, I think there might be breaking news during that week that led to this abnormal distribution. However, I tried to search on the news on CNN, Fox during that week, didn't find any notable news. Another guess is there might be something wrong with the data collection pipeline that caused the unbalanced data collection.

2.2.2 Word Frequency analysis

To figure out what topics people care most when they talk about coronavirus, we conducted a word frequency analysis.

1. Conducted the word frequency analysis for the original tweets and visualize in a wordcloud as follow[3](right figure).

2. We also summarized how many times each hashtag appears in tweets, filtering out the top 20 mentioned hashtags[3](left figure).Can also see a interactive version in my flask app.

From the visualization, we can know that main topics in coronavirus tweets are grocery/panicbuying related, lockdown/stayathome policy related, or the COVID-19 itself.

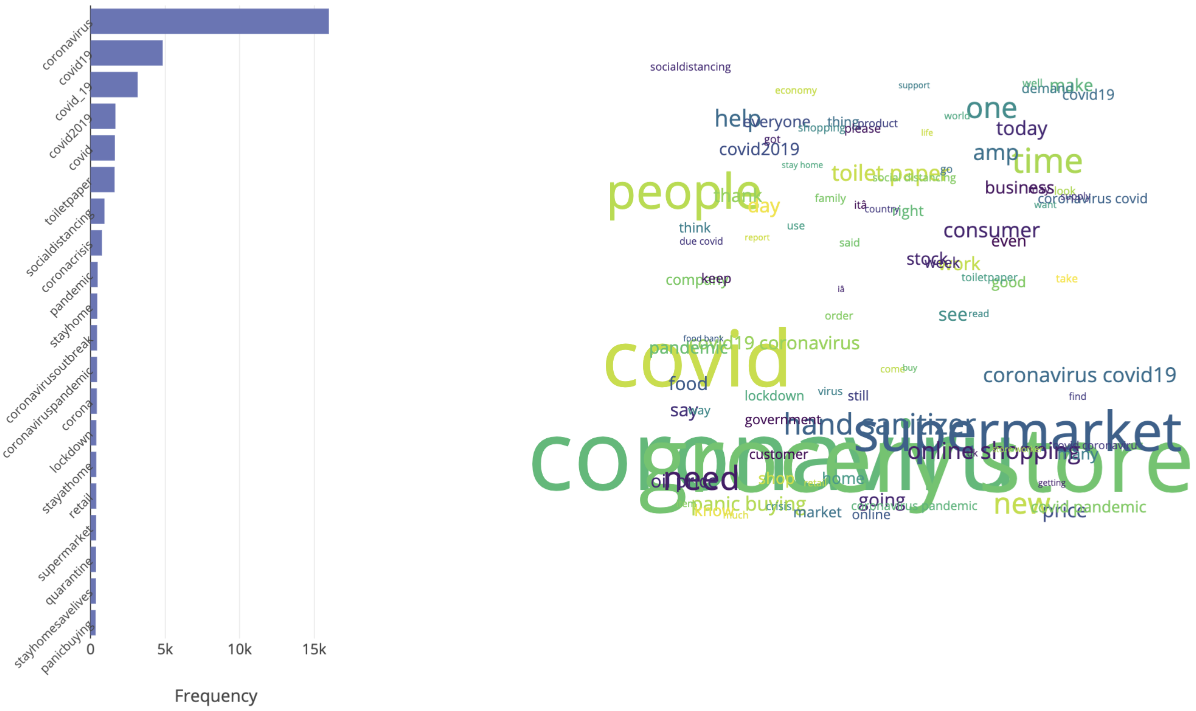


Figure 3: *The wordcloud and top 20 popular hashtags*

3 Coronavirus tweets Modelling

3.1 Topic Modelling

To dig deeper on When people talk about the COVID-19, what topics they care about, beside the simple frequency analysis, we also applied a topic modelling algorithm——Latent Dirichlet Allocation(LDA)[4]. This is a Statistical model for discovering the abstract ”topics” that occur in a collection of documents. This methods make use of the idea of unsupervised machine learning and the Bayesian statistics, built up a model to explain the components of topics in a corpus(i.e. 20% from topic A+ 50% from topic B +30% from the topic C). We took the number of topics as 5 here.

3.2 Interesting Findings from LDA

We expected to find more interesting insights from the modelling result. However, result from this algorithm doesn't give much extra information than the word frequency analysis. This might due to two factors, training sample quality (the tweets are too short to be seen as a corpus) and the uncertain characteristics of this kind of unsupervised learning algorithm. But we still visualized the 5 topics we clustered in Figure 4.

We can see the central words for each topic in the flask app ($\lambda = 0.6$). In conclusion,

- For the topic 1, the focus is more about the economy effect of the pandemic, with frequent mentioned words like supermarket, oil, price, economy, company, crisis and scam, etc.
- The topic 3 and topic 2 are very close to each other by distance mapping, they are more related to people lives during pandemic, with key words like worker, store, food, online, shopping, etc.
- The topic 4 is about the things that can be used to prevent COVID-19, with a key word list as hand, sanitizer, glove, mask and toiletpaper(certainly can't prevent COVID, but it's important during pandemic), etc.
- Topic 5 is not as meaningful as other topics, the key words contain some colloquial word, like uh, nah. It's also relatively distant from other topics.

3.3 Text Classification

I applied Bert[5][1] for sequence classification model to this text classification task. In the beginning, I used the original sentiment label(i.e. ”Neutral” , ”Positive”, ”Extremely Negative”, ”Negative”, ”Extremely Positive”). However, the final validation accuracy is only 81%. I realized the labelling of Extremely and not Extremely is very objective, even for human being, it's hard to tell which is Extremely Negative, which is Negative. For this data analysis, a more important thing is to tell negative sentiment from positive sentiment. Thus, I simplified the label as ”Neutral” , ”Positive”, ”Negative” by combining the Extremely negative/positive to negative/positive. The final parameter setting is in Table 2.

3.4 Classification Model Performance

The model reaches an accuracy of 91% on the validation set. It can be used to classify the new tweets data, which is implemented in the flask app. This is also a way to **validate** the model, whether it's still robust for the new data.

Figure 5 is the confusion matrix and the loss during training, we can see 2 epochs is enough for training.

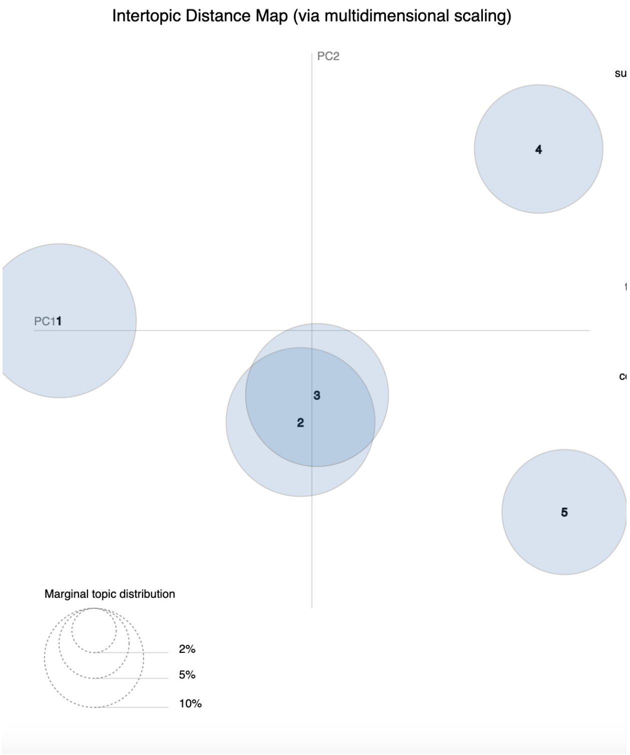


Figure 4: *The topic modelling results*

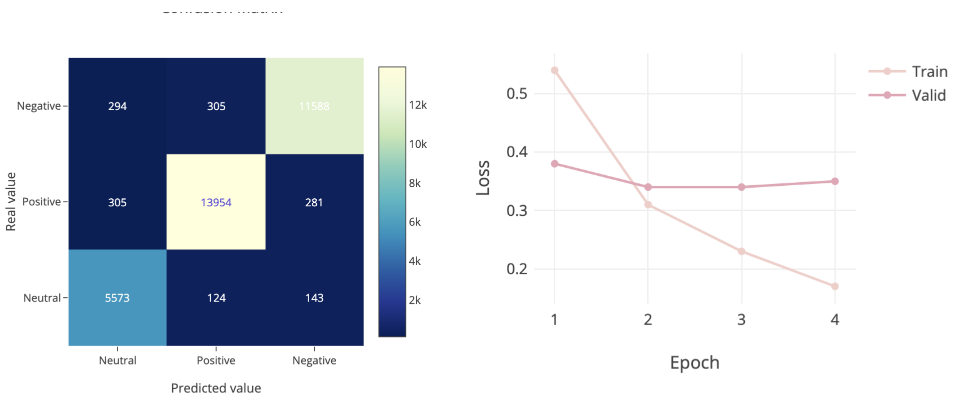
Parameter	Value
Optimizer	Adam
Learning Rate	2×10^{-5}
Epoch	4
Batch	64
Train:Validate	$9:1$

Table 2: *The parameter setting*

4 API and web front-end

4.1 Backend API server

The framework of the API is showed at Figure 6.



(a) *confusion matrix* **(b)** *loss by epoch*

Figure 5: *Model Performance*

- **First**, we create a Flask app and set configurations[3].
- **Second**, use the decorator define a bunch of functions, render serveral pages from the **html** template. Among the app route list here, the sentiment page is generated when user inputs the new information to the form in model page. and in eda page, it will request the data generated by the /api/data, /bar, /wordcloud for visualization
- **Third** In the main function, import the original tweet .csv data and the pretrained Bert model and put in the data cleaning pipeline for later use. The reason why we don't use the cleaned data, is we want to transfer this project to new dataset in the future this design may make it easier to transfer.
- **Fourth** Run the app in localhost/ other servers.

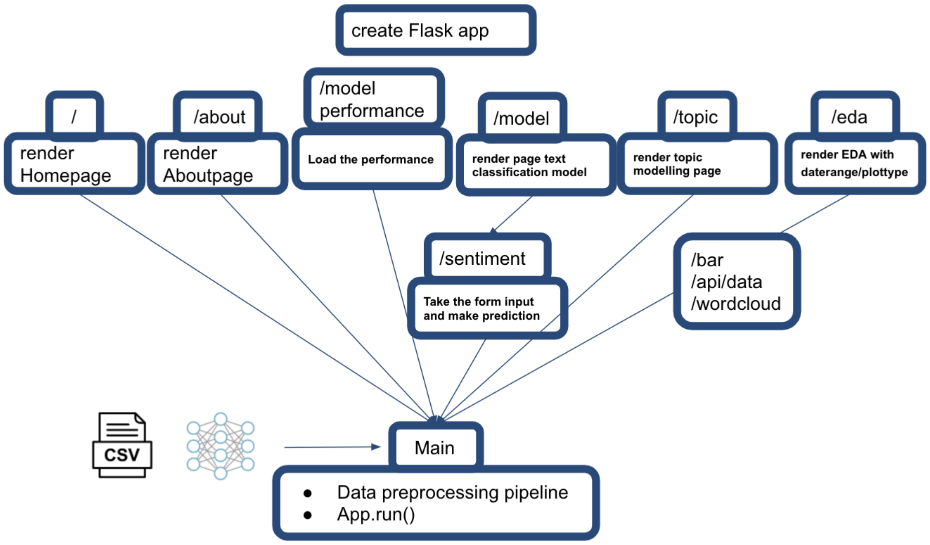

create Flask app
/
render Homepage
/about
render Aboutpage
/model performance
Load the performance
/model
render page text classification model
/topic
render topic modelling page
/eda
render EDA with daterange/plottype
/sentiment
Take the form input and make prediction
/bar
/api/data
/wordcloud
CSV
Main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

Figure 6: *API server framework*

4.2 Web Frontend

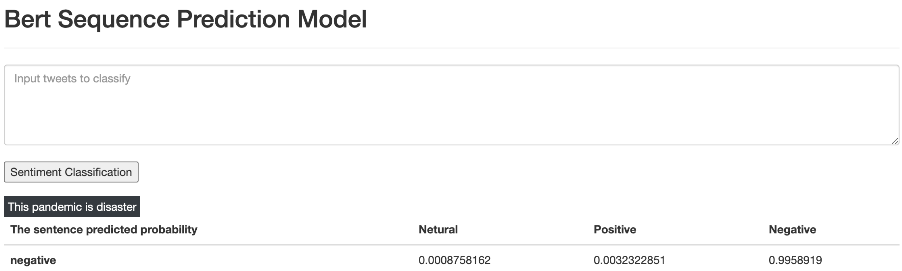
The web frontend has 6 pages[2].

1. The **Homepage** and **About** page is simply a navigation and introduction to the whole projects(with some data decription a background).

2. The **EDA** page contains the summary statistics of the word frequency and the sentiment distribution, it looks like Figure 2,3 in appearance. You will be able to select the date range and the plot type(bar/scatter) for timeseries of sentiment. Hovering on the plot, you will be able to see the # of tweets of different sentiments. When you click on a day in the the sentiment distribution graph(either bar/scatter). A word cloud and top20 hashtag bar plot will be generated below. This feature enables you to analyze the topics change overtime.

3.The **Sentiment prediction** page contains a prediction model, you will be able to input some sentences/tweets, click on sentiment classification to analyze the sentiment of it. See the Figure 7 below:

4. Under **More** tab, there are two more pages one is **Sentiment Analysis** page, containing information of text classification model training(parameter setting/ confusion matrix/loss) just as Figure 5. Another page **LDA** is the result of topic modelling, user can specify the λ and hover on topic to see the central words on the right panel.



Bert Sequence Prediction Model

This pandemic is disaster

The sentence predicted probability	Netural	Positive	Negative
negative	0.0008758162	0.0032322851	0.9958919

Figure 7: *The demo of sentiment prediction page*

Note: This flask app is also partially deployed on bis634.herokuapp.com. Bert model is too large to deploy on heroku and the wordcloud model takes more than 30 seconds to respond, will cause a time out, other features work fine.

5 Discussion

5.1 Findings and Surprises

- People's topics in March were around supermarket/toiletpaper while in April we focus on the social distancing more

- People like to mention @realdonaldtrump when they send COVID-19 related tweets.

- The topics among the coronavirus tweets can be classified as economy, lives, necessity for COVID-19 prevention, coronavirus and other random topics.

- The sentiment distribution at the end of March is very different and # of tweet was 1/10 of 20th, March.

5.2 Difficulties

- Location variable is too dirty and not applicable Tweets are different from other text. Need to deal with the url/mentions/hashtags
- Find the reason for the abnormal data is extremely difficult.
- API Implementation: Need to load large data set, take times how to make the whole pipeline more efficient is challenging. The visualization How to pass parameter to the html and get the data from the user input are very difficult to figure out in logic.

5.3 Future Work

- We can conduct sentence embedding and dimension reduction to evaluate the similarity of two tweets.
- We can improve the API and the frontend to make it more efficient.
- We need to dig more insights from this tweets dataset and generalize this work to other fields.

References

[1] *Bert pretrained*. URL: https://mccormickml.com/2019/07/22/BERT-fine-tuning/.

[2] *boostrapping style*. URL: https://getbootstrap.com/.

[3] *Flask tutorial*. URL: https://flask-appbuilder.readthedocs.io/en/latest/intro.htmll (visited on 09/30/2020).

[4] *Topic model*. URL: https://www.wikiwand.com/en/Topic_model.

[5] *transformer github*. URL: https://github.com/huggingface/transformers/tree/master/examples.

[6] Wikipedia. *Covid wikipedia*. 2019. URL: https://www.wikiwand.com/en/Coronavirus_disease_2019 (visited on 09/30/2010).