LOUISIANA STATE UNIVERSITY
DIVISION OF COMPUTER SCIENCE AND ENGINEERING (CSE)

Sentiment Analysis of Covid-19 Tweets of Louisiana and Washington and Topic Modeling Using Apache Spark

Nurjahan
Email: nurja1@lsu.edu
Date: 11 October, 2024
CSC 7740

October, 2024

Contents

List of Tables

List of Figures

1 Introduction

Due to technological advancements, people are becoming more interactive on social media platforms and proving their opinions. The generated data of the users are unstructured which is growing at an unprecedented rate. Nowadays, business analysts are considering these data to understand the reviews of their products which helps them to reorganize and rethink products, thus can help to flourish business profits. Meanwhile, customers are considering these reviews to buy new products. Similarly, by analyzing the tweet patterns and emotions, it becomes easier to understand how a policy or pandemic affects the sentiments of the people. Covid-19 was first identified in Wuhan, China in December 2019 and the World Health Organization declared this deadly coronavirus pandemic on March 11, 2020 [1]. People passed their time among uncertainty during these period. By analyzing the Twitter data, it is possible to understand people's fear, thinking about COVID-19, trust about the policies, their reaction patterns, and many more. The natural language processing field involves processing the Twitter text to extract the meanings and emotions from the text.

1.1 Motivation

Generally, the data from online media contains lots of information including misinformation. During the pandemic time, most people stayed home, had to work from home, restriction from roaming public places. People become more active in online when staying at home and can not participate in physical activities. The frustration, fear, or various mental health problems can be understood by analyzing their online activities. After analyzing the data about that crucial time about community people, the decision maker or the health professionals can take several preventive measures. Sometimes, the community people can be misled by wrong information and can show mistrust of the healthcare system or state's policy. We were interested in how people's emotions reacted during these periods.

1.2 Project Description

In this work, we have collected a total of 2769978 raw tweets to analyze the sentiments of the people. Among these tweets, 314534 are covid-19 related tweets. We have analyzed the sentiment score of Twitter users based on gender, age, and period. We have also deployed the unsupervised machine learning technique Latent Dirichlet allocation (LDA) to find the most used topics from the text data. To analyze such a huge number of data, we work on the distributed environment using Spark.

2
System Design

The main tasks of this project have been demonstrated in figure 2.1. The tasks can be divided into two broad categories: sentiment analysis and topic modeling. The process of topic modeling has been shown in Figure 3.1.

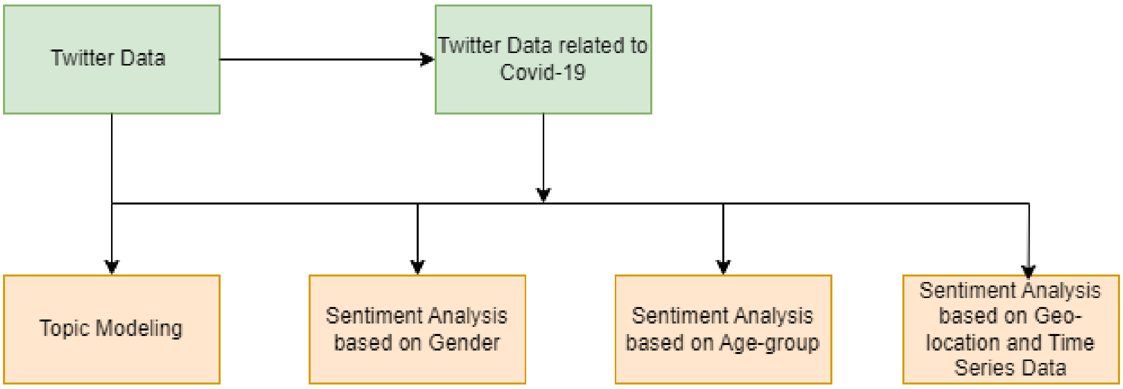

Twitter Data
Twitter Data related to Covid-19
Topic Modeling
Sentiment Analysis based on Gender
Sentiment Analysis based on Age-group
Sentiment Analysis based on Geo-location and Time Series Data

Figure 2.1: Analysis of Twitter Data from Different Perspectives

2.1 Chosen Big Data Frameworks

To do this project, Apache Spark framework [2] has been utilized for parallel and distributing computing.To utilize spark famework, we have utilized pyspark library of python. Along with that I have done topic modeling using pyspark's MLlib library [3]. Besides, sql functions is being used to aggregate data based on different groups.

2.2 Chosen Datasets

In this project, I have collected the Twitter data of Louisiana and Washington. Mainly, the data includes 48 hours before and after of stay at home order and mask-mandate order of two different states. The dataset is not publicly available [1]. The total number of rows of the dataset is 2769978 and the number of attribute is 16. However, the Twitter data related to Covid-19 is 314534. The attribute of the dataset is mentioned in Table 2.1

Table 2.1: Schema of Twitter Data and Data Characteristics

Column Name	**Data Type**	**Nullable**	**Null value count**
TweetID	long	Yes	0
UserID	long	Yes	0
Username	string	Yes	0
Timestamp	string	Yes	0
Text	string	Yes	0
State	string	Yes	0
County	string	Yes	42246
City	string	Yes	42246
Sentiment	double	Yes	0
COVID-related	long	Yes	0
AgeGroup	string	Yes	0
Age_Confidence	string	Yes	0
Gender	string	Yes	0
Gender_Confidence	string	Yes	0
Org_Confidence	string	Yes	0
Retweet	long	Yes	0

[1] https://lsumail2-my.sharepoint.com/:u:/r/personal/nurja1_lsu_edu/Documents/Nurjahan_CSC7740/COVID19MisinformationPaper/themes_of_misinfo_project_tweets/sqlite_db/tweets_db_2023-03-12_17-57-49.db?csf=1&web=1&e=RX4s51

3

Detailed Description of Components

The whole task is subdivided into sentiment analysis and topic modeling.

3.0.1 Each Component Description

3.0.1.1 Sentiment Analysis

We have analyzed the average sentiment score based on the age and age group of the two states and reported the results. Furthermore, we also analyzed the overall sentiment based on the announcement time and showed the results.

3.0.1.2 Topic Modeling

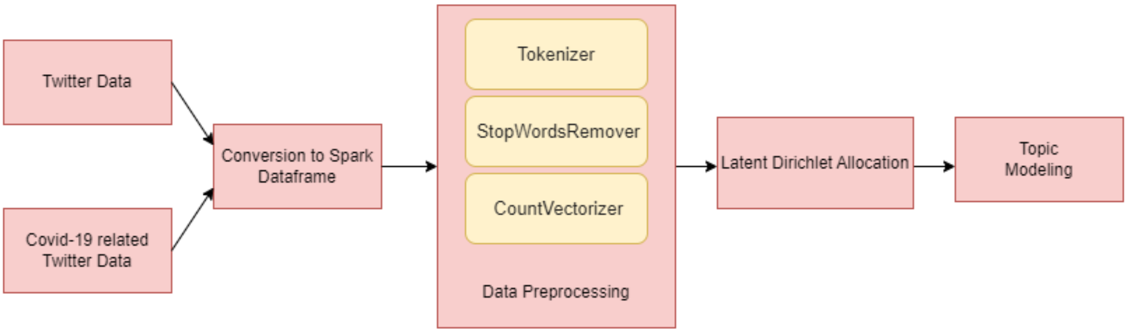
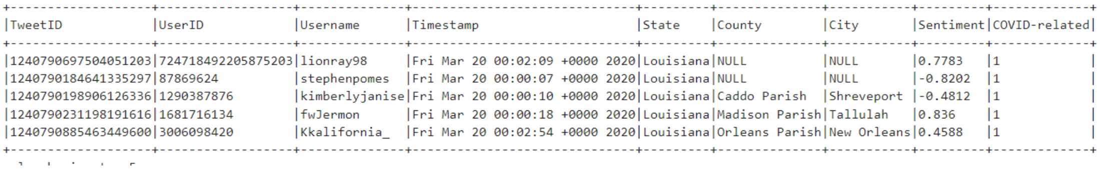

Twitter Data
Covid-19 related Twitter Data
Conversion to Spark Dataframe
Tokenizer
StopWordsRemover
CountVectorizer
Data Preprocessing
Latent Dirichlet Allocation
Topic Modeling

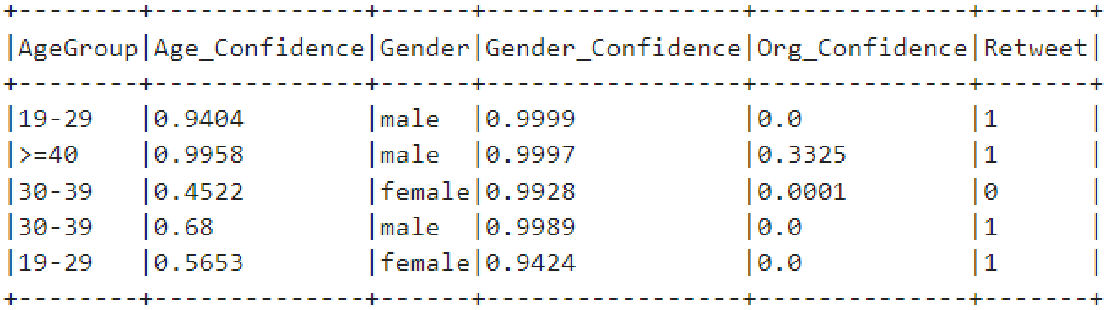
Figure 3.1: Topic Modeling Using LDA

Data Loading: First we have read the db file of the Twitter data. Then we converted the db file to csv format. We have read the csv file using pandas dataframe and then converted to spark dataframe. The data sample has been shown on the figures of 3.2, 3.3 and 3.4.



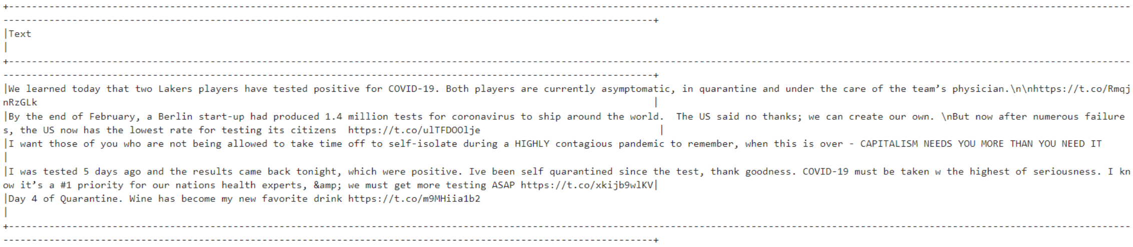
```
+-------------------+-------------------+-------------+------------------------------+---------+--------------+------------+---------+------------+
|TweetID            |UserID             |Username     |Timestamp                     |State    |County        |City        |Sentiment|COVID-related|
+-------------------+-------------------+-------------+------------------------------+---------+--------------+------------+---------+------------+
|1240790697504051203|724718492205875203|lionray98    |Fri Mar 20 00:02:09 +0000 2020|Louisiana|NULL          |NULL        |0.7783   |1           |
|1240790184641335297|87869624           |stephenpomes |Fri Mar 20 00:00:07 +0000 2020|Louisiana|NULL          |NULL        |-0.8202  |1           |
|1240790198906126336|1290387876         |kimberlyjanise|Fri Mar 20 00:00:10 +0000 2020|Louisiana|Caddo Parish  |Shreveport  |-0.4812  |1           |
|1240790231198191616|1681716134         |fwJermon     |Fri Mar 20 00:00:18 +0000 2020|Louisiana|Madison Parish|Tallulah    |0.836    |1           |
|1240790885463449600|3006098420         |Kkalifornia_ |Fri Mar 20 00:02:54 +0000 2020|Louisiana|Orleans Parish|New Orleans|0.4588   |1           |
+-------------------+-------------------+-------------+------------------------------+---------+--------------+------------+---------+------------+
```

Figure 3.2: Data Sample (Part-01)



```
+--------+--------------+------+-----------------+--------------+-------+
|AgeGroup|Age_Confidence|Gender|Gender_Confidence|Org_Confidence|Retweet|
+--------+--------------+------+-----------------+--------------+-------+
|19-29   |0.9404        |male  |0.9999           |0.0           |1      |
|>=40    |0.9958        |male  |0.9997           |0.3325        |1      |
|30-39   |0.4522        |female|0.9928           |0.0001        |0      |
|30-39   |0.68          |male  |0.9989           |0.0           |1      |
|19-29   |0.5653        |female|0.9424           |0.0           |1      |
+--------+--------------+------+-----------------+--------------+-------+
```

Figure 3.3: Data Sample (Part-02)



```
+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------+
|Text
|
+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------+
|We learned today that two Lakers players have tested positive for COVID-19. Both players are currently asymptomatic, in quarantine and under the care of the team’s physician.\n\nhttps://t.co/Rmqj
nRzGLk                                                                                                                          |
|By the end of February, a Berlin start-up had produced 1.4 million tests for coronavirus to ship around the world.  The US said no thanks; we can create our own. \nBut now after numerous failure
s, the US now has the lowest rate for testing its citizens  https://t.co/ulTFDOOlje                                              |
|I want those of you who are not being allowed to take time off to self-isolate during a HIGHLY contagious pandemic to remember, when this is over - CAPITALISM NEEDS YOU MORE THAN YOU NEED IT
|
|I was tested 5 days ago and the results came back tonight, which were positive. Ive been self quarantined since the test, thank goodness. COVID-19 must be taken w the highest of seriousness. I kn
ow it’s a #1 priority for our nations health experts, &amp; we must get more testing ASAP https://t.co/xkijb9wlKV|
|Day 4 of Quarantine. Wine has become my new favorite drink https://t.co/m9MHiia1b2
|
+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------+
```

Figure 3.4: Data Sample (Part-03)

Data Preprocessing: We have checked the dataset for null values and found 42246 null values both in county and city values. We have preprocessed the text field using tokenizer, stop word remover, count vectorizer to make it suitable for the model.

Model: We applied unsupervised algorithm LDA [4] for modeling the topics.

4
Evaluation & Test

4.1 Software Manual

We have utilized Jupyter notebook on our local machine. We have used pyspark version 3.5.3 which is updated version. But this version does not work with updated version of python 3.12. Thats why we first downgraded the python version and perform the computing.

4.1.1 Test Environment

We have performed the analysis on Windows PC, version 11, Processor 13th Gen Intel(R) Core(TM) i7-13620H, 2400 Mhz, 10 Core(s), 16 Logical Processor(s) and installed Physical Memory (RAM) 16.0 GB.

4.1.2 Test Results

The figure 4.1, 4.2 and 4.3 shows the Twitter count based on states, gender and age groups presented in the dataset.

The figure 4.4 shows the sentiment score distribution of all covid-19 related tweets. It can be concluded that most of of the tweets are neural.

We have demonstrated the sentiment analysis based on time result on figure 4.5, 4.6, 4.7 and 4.8. The figure 4.5 shows the overall sentiment before the stay-at-home order and after the order in Louisiana. The red line depicts the announcement time of stay at home order. The figure shows that most of the users shows positive sentiment score and the pattern is almost similar.

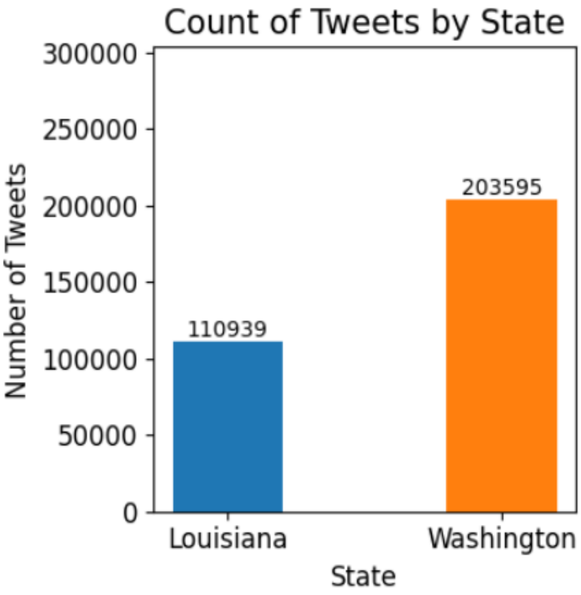


Figure 4.1: Number of Tweets based on States

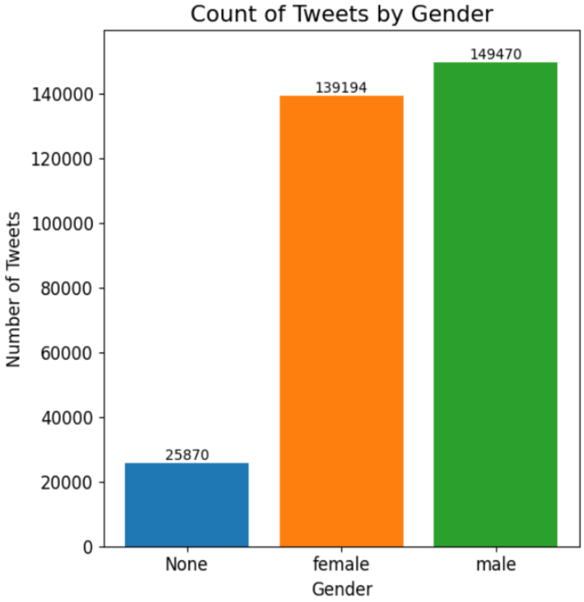


Figure 4.2: Number of Tweets based on Gender

The figure 4.6 shows the overall sentiment before the mask mandate order and after the order in Louisiana. The red line depicts the announcement time of mask mandate order. The figure shows that most of the users shows negative sentiment score unlike the time of stay at home time.

The figure 4.7 shows the overall sentiment before the stay-at-home order and after the order in Washington. The red line depicts the announcement time of stay at home order. The figure shows that most of the users shows negative sentiment score before the order which changes after the announcement time.

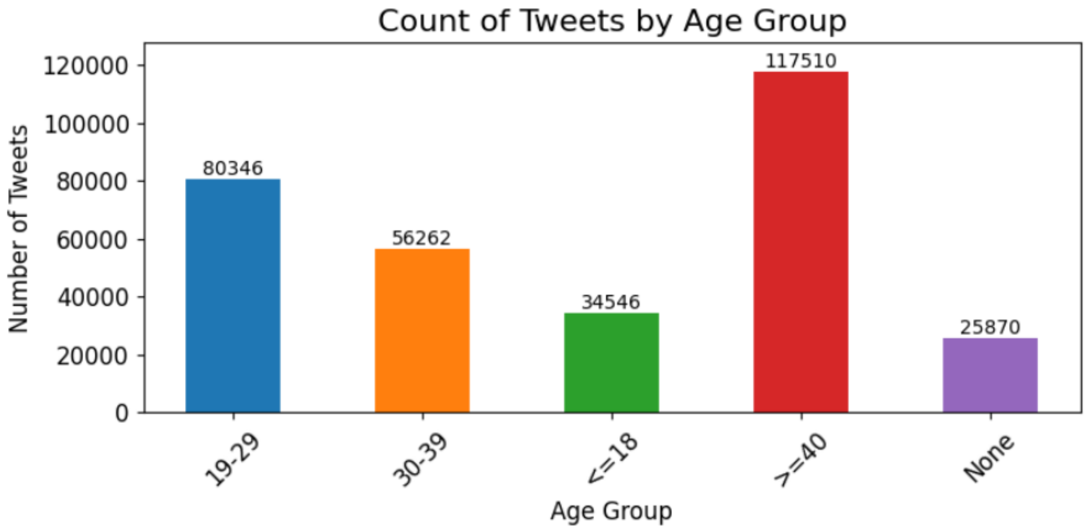


Figure 4.3: Number of Tweets based on Age-group

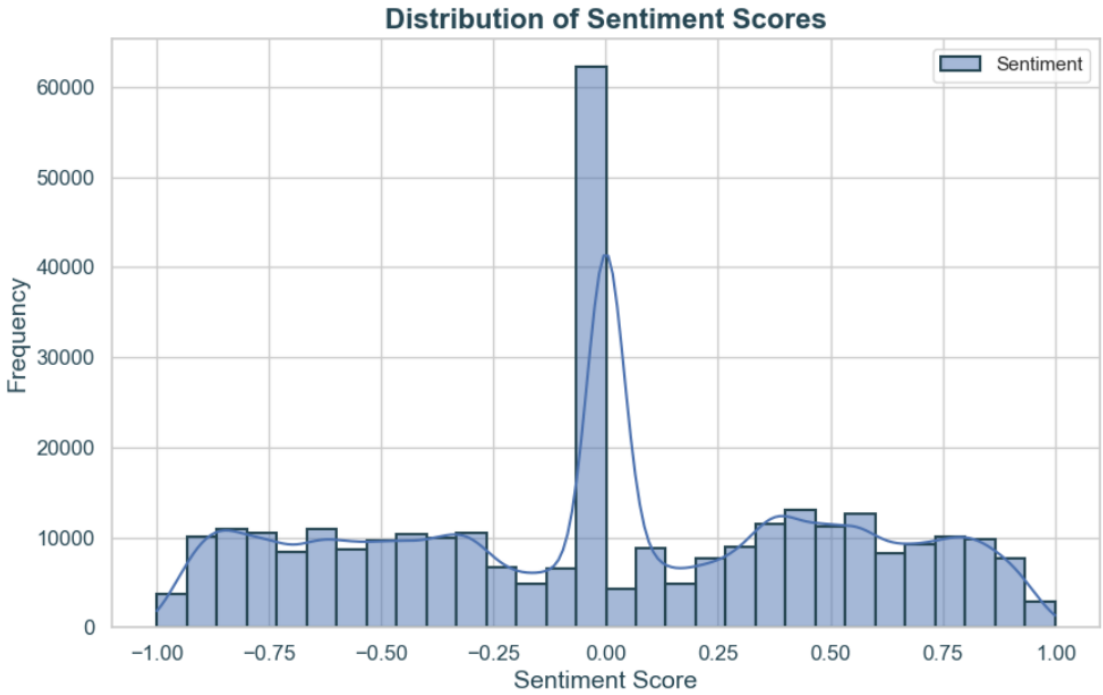


Figure 4.4: Distribution of Sentiment Score

The figure 4.8 shows the overall sentiment before the mask mandate order and after the order in Washington. The red line depicts the announcement time of mask mandate order. The figure shows that most of the users shows negative sentiment score during this time.

The figure of 4.9, 4.10, 4.11, 4.12 shows the average sentiment of two states during the periods of stay home and mask mandate order based on gender and age group. The pattern depicted that the avergae sentiment during mask mandate time is mostly negative

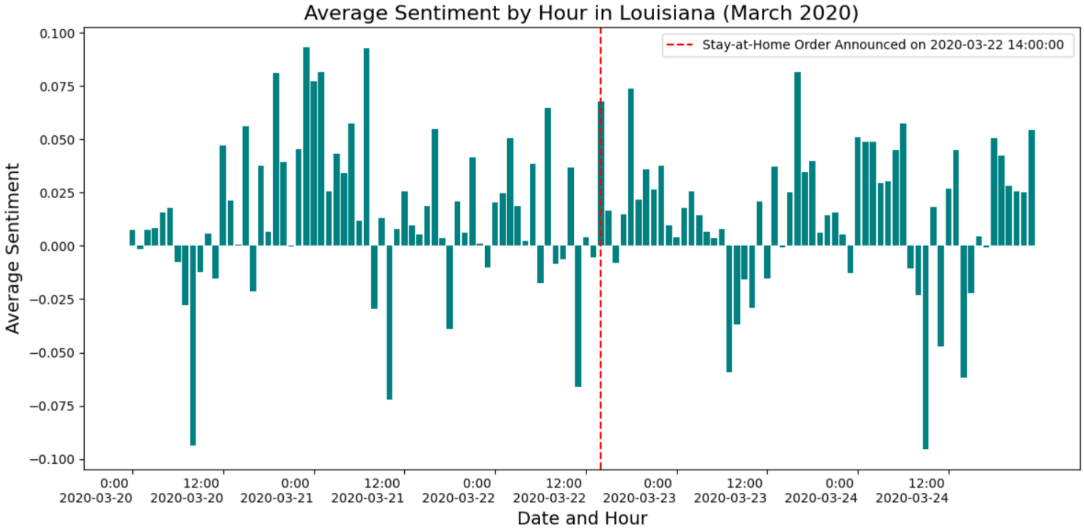


Figure 4.5: Sentiment Analysis Result during Stay-at-home Order in Louisiana

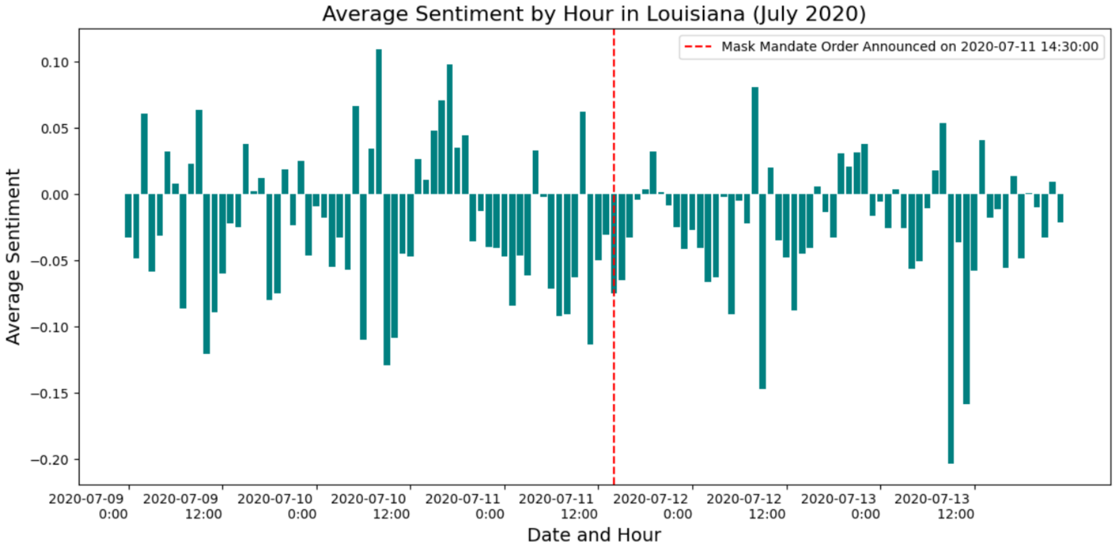


Figure 4.6: Sentiment Analysis Result during Mask-mandate Order in Louisiana

in both states and the people of age more than forty shows negative sentiment across all periods. The figure 4.13 shows that overall the female greater than forty mostly posed negative average sentiment.

The figures 4.14 and 4.15 corresponds the topic modeling result of covid-19 related tweets and total overall tweets respectively.

From the result of figure 4.14 we can see that five topics are related to discussion about trust and the responses of institution, the use of social media to spread news about policies like mask, personal experiences about quarantine, political moves of trump administration, mixture of political and personal experience respectively.

The topic modeling based on all Tweets 4.15 reveals political and social issues like black rights, personal opinion of individual, emotions of personal happiness or sadness,

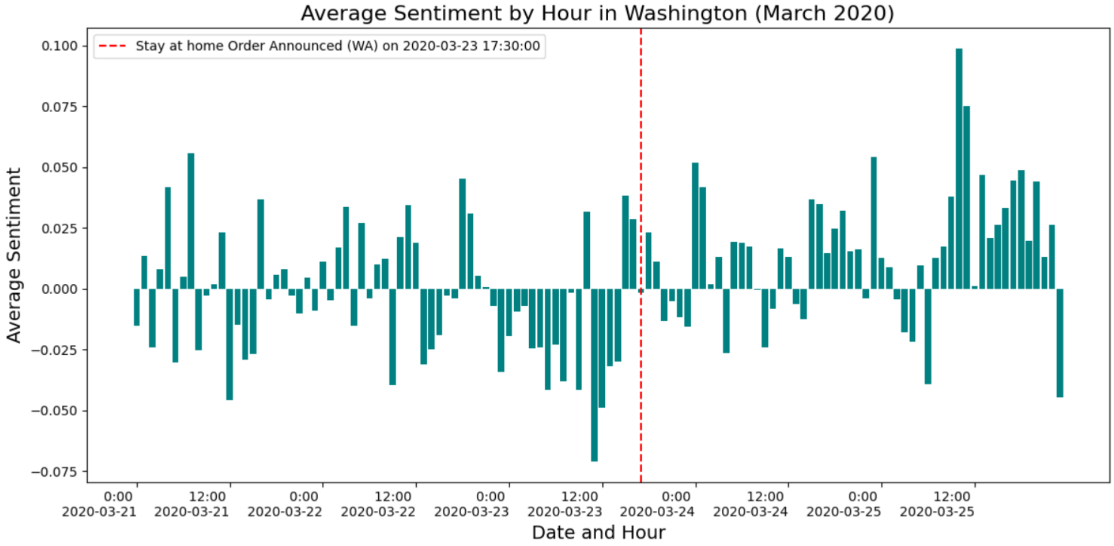


Figure 4.7: Sentiment Analysis Result during Stay-at-home Order in Washington

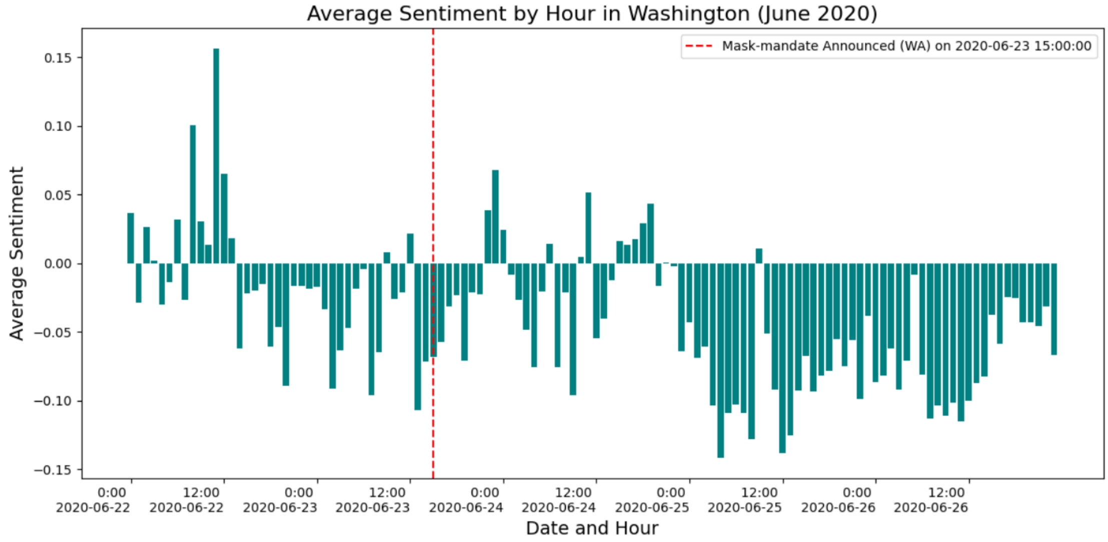


Figure 4.8: Sentiment Analysis Result during Mask-mandate Order in Washington

combination of emotional and action oriented words, neutral emotional respectively.

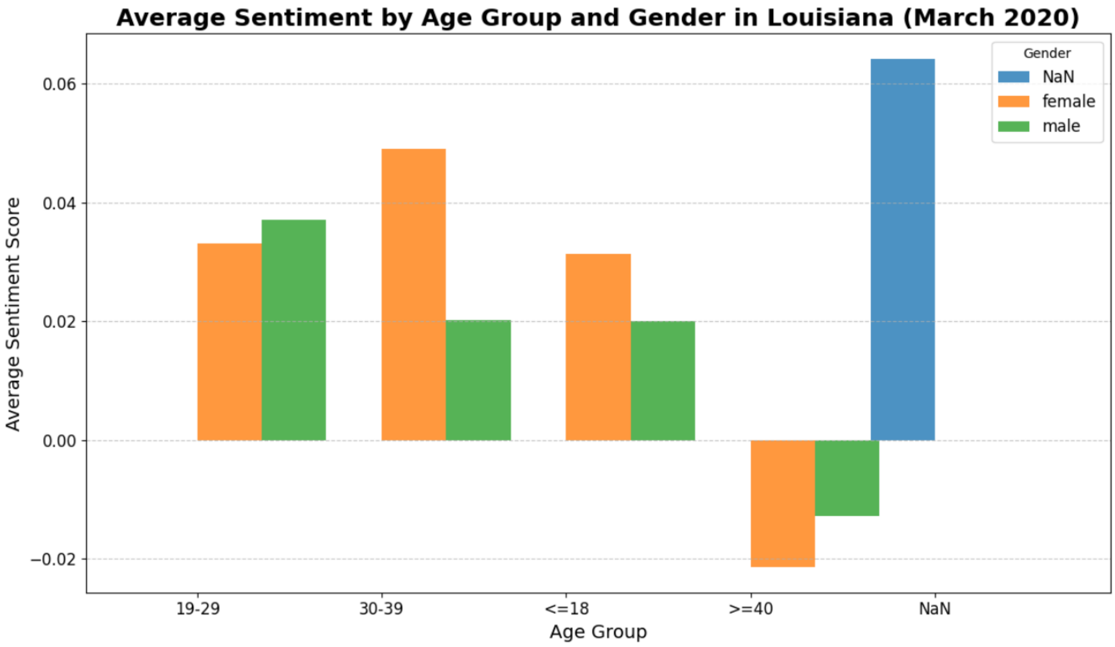


Figure 4.9: Sentiment Analysis Result based on Gender and Age during stay home order in LA

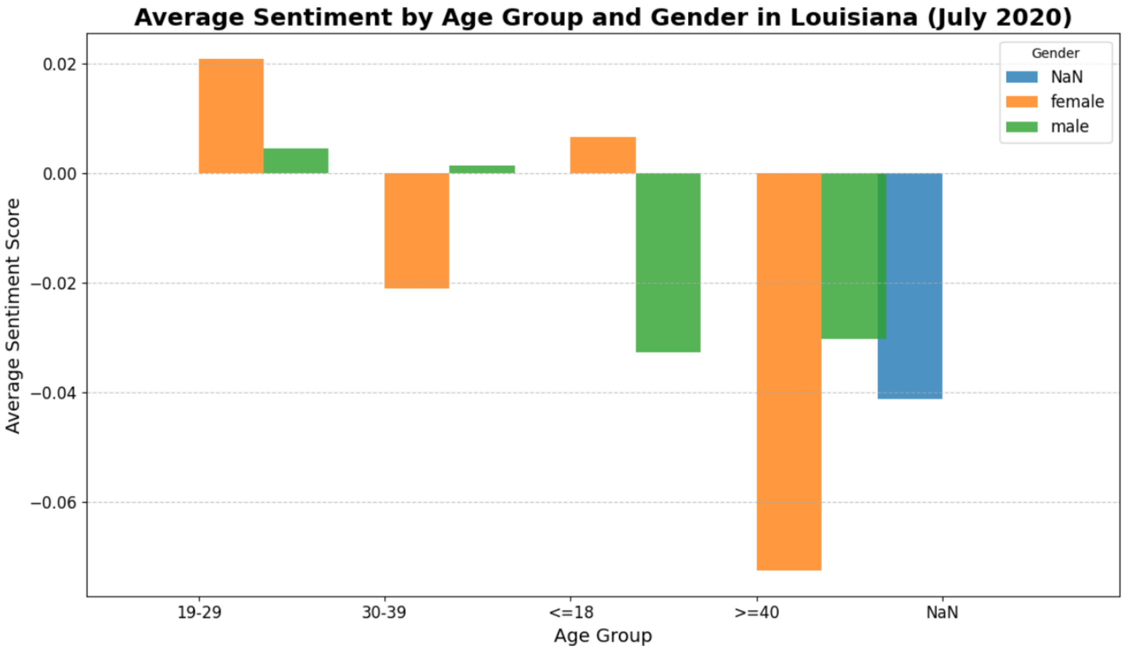


Figure 4.10: Sentiment Analysis Result based on Gender and Age during mask mandate order in LA

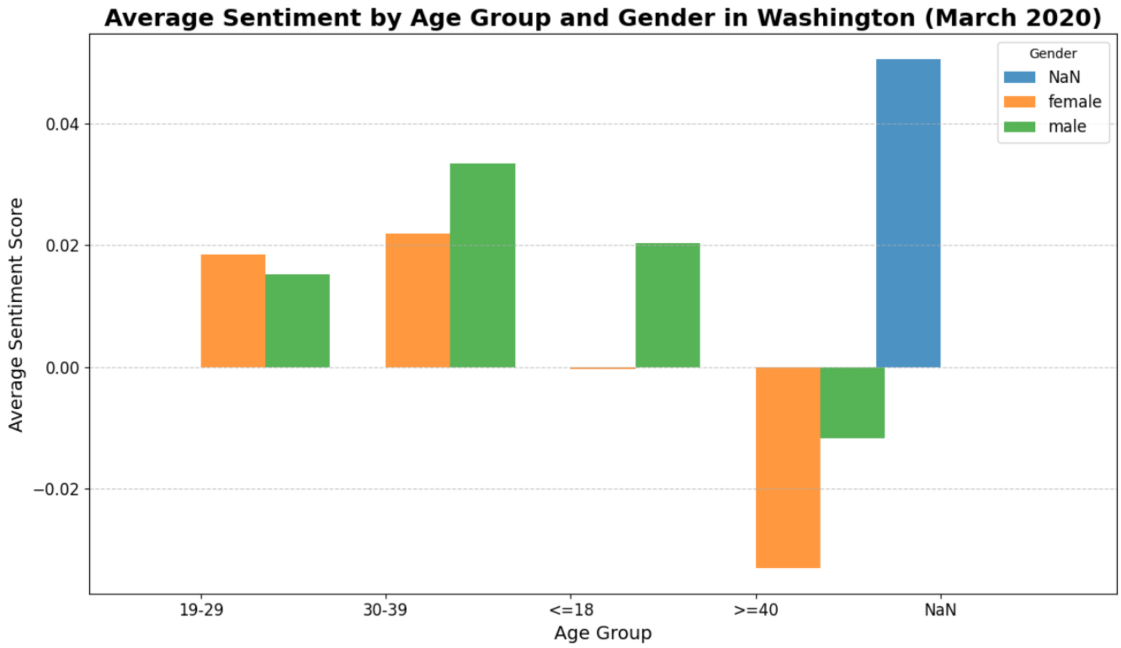


Figure 4.11: Sentiment Analysis Result based on Gender and Age during stay home order in WA

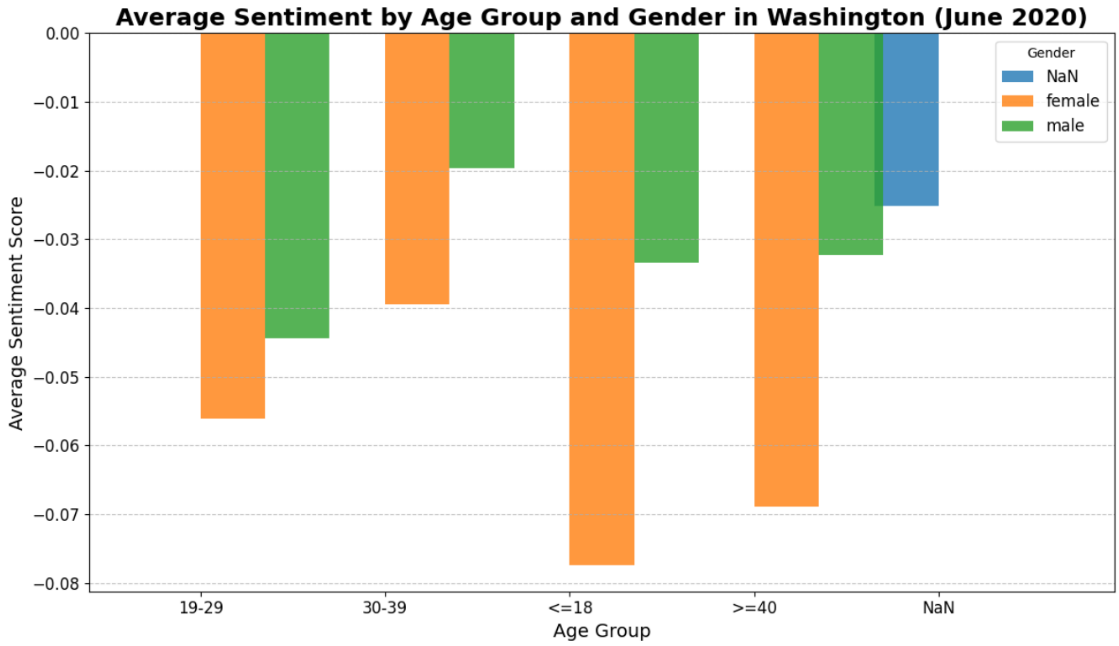


Figure 4.12: Sentiment Analysis Result based on Gender and Age during mask mandate order in WA

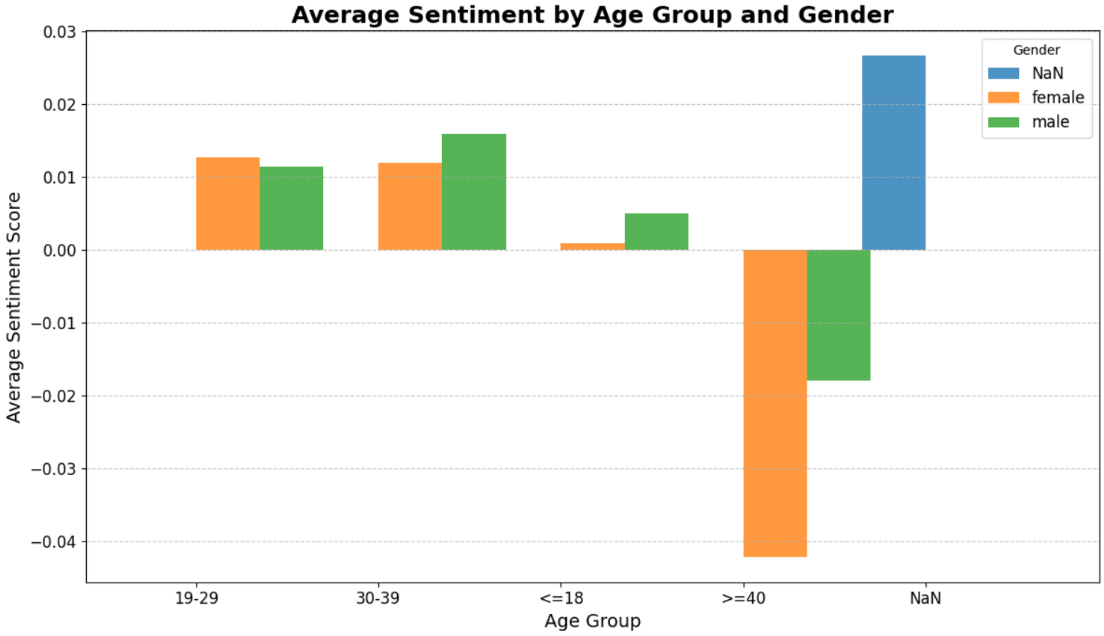
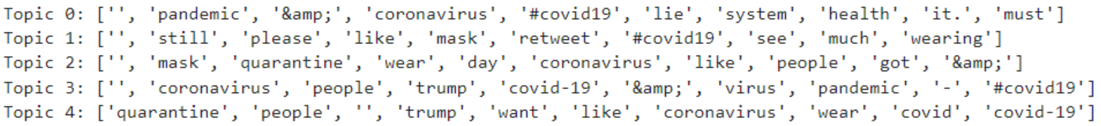
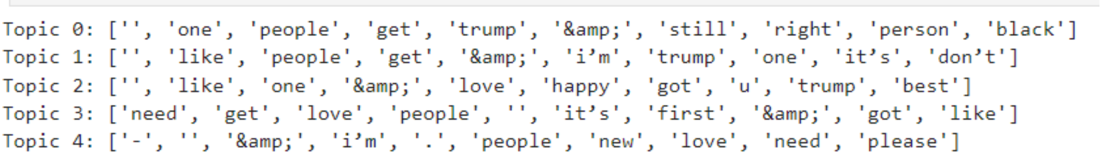


Figure 4.13: Sentiment Analysis Result based on Gender and Age



```
Topic 0: ['', 'pandemic', '&amp;', 'coronavirus', '#covid19', 'lie', 'system', 'health', 'it.', 'must']
Topic 1: ['', 'still', 'please', 'like', 'mask', 'retweet', '#covid19', 'see', 'much', 'wearing']
Topic 2: ['', 'mask', 'quarantine', 'wear', 'day', 'coronavirus', 'like', 'people', 'got', '&amp;']
Topic 3: ['', 'coronavirus', 'people', 'trump', 'covid-19', '&amp;', 'virus', 'pandemic', '-', '#covid19']
Topic 4: ['quarantine', 'people', '', 'trump', 'want', 'like', 'coronavirus', 'wear', 'covid', 'covid-19']
```

Figure 4.14: Topic Modeling based on Covid-19 Related Tweets



```
Topic 0: ['', 'one', 'people', 'get', 'trump', '&amp;', 'still', 'right', 'person', 'black']
Topic 1: ['', 'like', 'people', 'get', '&amp;', 'i’m', 'trump', 'one', 'it’s', 'don’t']
Topic 2: ['', 'like', 'one', '&amp;', 'love', 'happy', 'got', 'u', 'trump', 'best']
Topic 3: ['need', 'get', 'love', 'people', '', 'it’s', 'first', '&amp;', 'got', 'like']
Topic 4: ['-', '', '&amp;', 'i’m', '.', 'people', 'new', 'love', 'need', 'please']
```

Figure 4.15: Topic Modeling based on all the Tweets

5

Conclusion

Social media becomes a part and parcel for our daily life nowadays. A large amounts of social media data is being generated in every moments. People are providing their opinions, views, likings, dislikings about any issues, policies, products everyday using various social media platforms such as Twitter, Facebook, Instragram, Tiktok and many more. Analyzing these unstructured natural language text has opened up new research dimensions to find out interesting patters, people's positive and negative emotions over a event. In this paper, We have worked on twitter data of two different states of USA, Washington and Louisiana. The data has been collected from Twitter and it contains data during the periods of stay-at-home order and mask mandate order of this states. The analyzed data showed several latent patterns.

Appendix

Table 5.1: File Names and Line Counts

File Name	# Lines
exploratory_data_analysis.ipynb	419
sentiment_analysis.ipynb	477
time_series.ipynb	461
topic_modeling.ipynb	68
topic_modeling_all_tweets.ipynb	106

Table 5.2: File Names and Number of Spark Operations

File Name	# Spark Operations
exploratory_data_analysis.ipynb	61
sentiment_analysis.ipynb	42
time_series.ipynb	66
topic_modeling.ipynb	10
topic_modeling_all_tweets.ipynb	12

Bibliography

[1] D. Cucinotta and M. Vanelli, "Who declares covid-19 a pandemic," *Acta bio medica: Atenei parmensis*, vol. 91, no. 1, p. 157, 2020.

[2] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin *et al.*, "Apache spark: a unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.

[3] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar, "Mllib: Machine learning in apache spark," *Journal of Machine Learning Research*, vol. 17, no. 34, pp. 1–7, 2016. [Online]. Available: http://jmlr.org/papers/v17/15-237.html

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.