DATA301 Project Final Report

Summary

In this project, I research how the 2020 U.S. election is likely to impact the coronavirus pandemic in the U.S.I use a Python library, which is nltk, for parsing semantics and the VADER algorithm to help me analyze the data I get from GDELT EVENT DATABASE and GDELT DOC API. With the new U.S. President Joe Biden taking a more cautious approach to the COVID-19, I think the coronavirus pandemic in the U.S. has been better controlled.

Introduction

I used the GDELT 2.0 EVENT DATABASE to retrieve the data. GDELT collects information from media in more than 100 languages worldwide and automatically encodes it into events by computer using a specific coding system, spanning the period 1979 to the present day, which continues to be updated daily.

By setting the keyword = ("Covid-19" OR coronavirus) and the country as the United States, I made the URL on the GDELT Summary website to obtain the data related to the coronavirus pandemic before and after the U.S. election. And by modifying the URL, I got an information table containing the website, release time, and title of the news.

I decided to study whether the U.S. election had an impact on the coronavirus epidemic in the U.S. because one of the campaign strategies of the Democratic candidate Joe Biden in this U.S. election was to increase the emphasis on COVID-19 and put forward related policies to protect people and revive the economy. On the other hand, the Trump administration's contempt for COVID-19 has helped accelerate the spread of the coronavirus in the United States. So I think after the election, the change of political parties will lead to a shift in policy, which will impact the spread of the virus.

To prove my hypothesis, I decided to import a python library specifically designed to analyze contextual and textual sentiment, nltk, to help analyze the data I obtained from GDELT. This library relies on the Vader algorithm. The Vader algorithm outputs emotion scores into three types of emotions: 'Negative', 'Neutral', and 'Positive', and uses 'compound' to represent the total number of emotions. So by analyzing 'compound' we can see how people's attitudes toward something over different periods.

Experimental Design and Methods

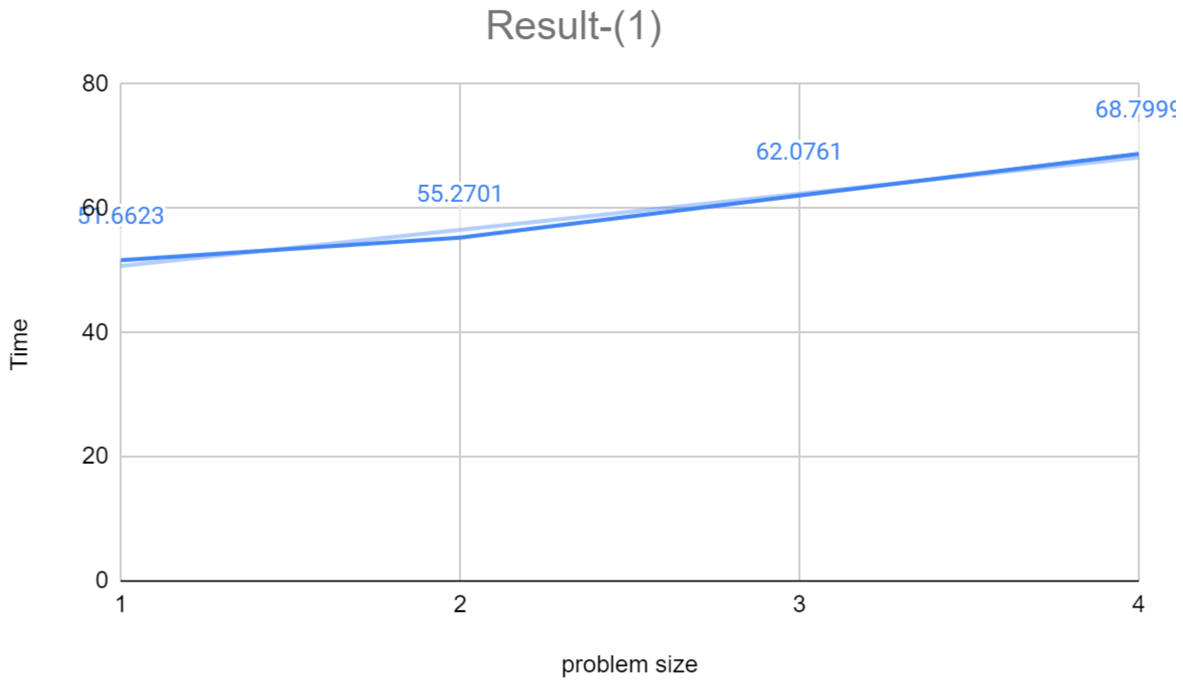
I finally decided to obtain the data of two months before and after the election respectively for analysis ("2020 Aug 3" to "2020 Nov 3") and ("2021 Jan 20" to "2021 Apr 20") because the U.S. election starts from November 3. It ends on January 20 when a new president, Biden, takes office. After retrieving the data, I found that there were many duplicates in the data, so I removed the data with the same title via 'dropDuplicates(['Title'])'.

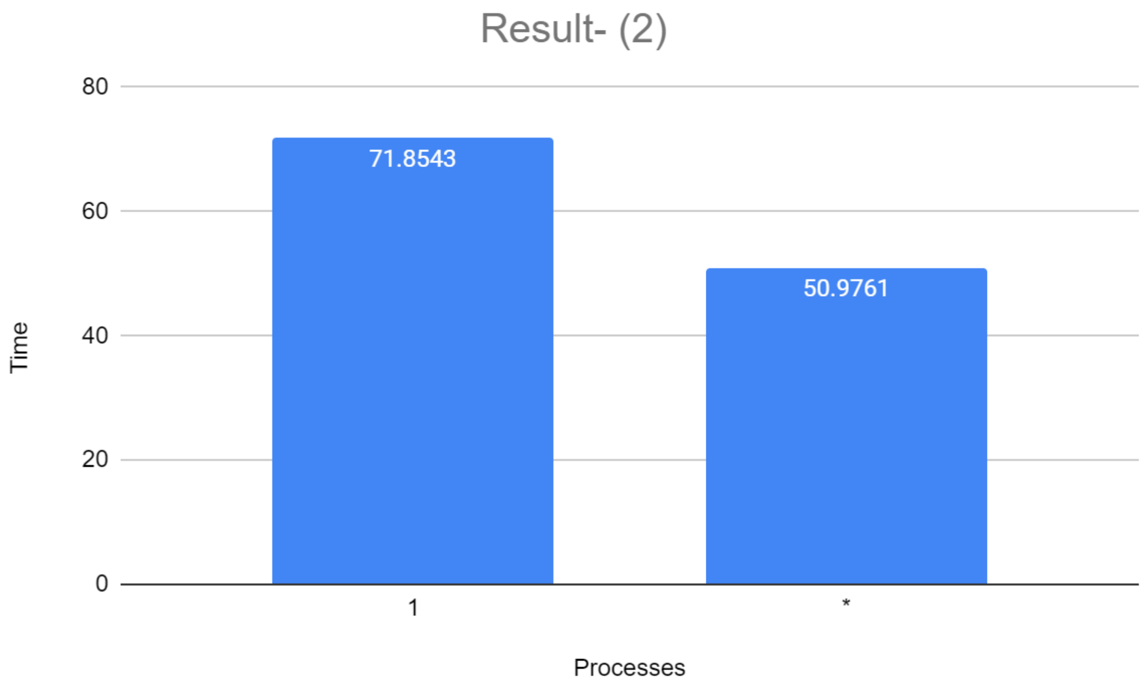
Then I calculated the above data through the POLARITY_SCORING function, which is used in the Vader algorithm to calculate compound scores. Then I define a compound greater than 0.2 to be positive, a compound less than -0.2 to be negative, and the rest to be neutral. Finally, we can obtain people's attitudes towards the coronavirus epidemic before and after the U.S. election by calculating the proportion of three different emotions.

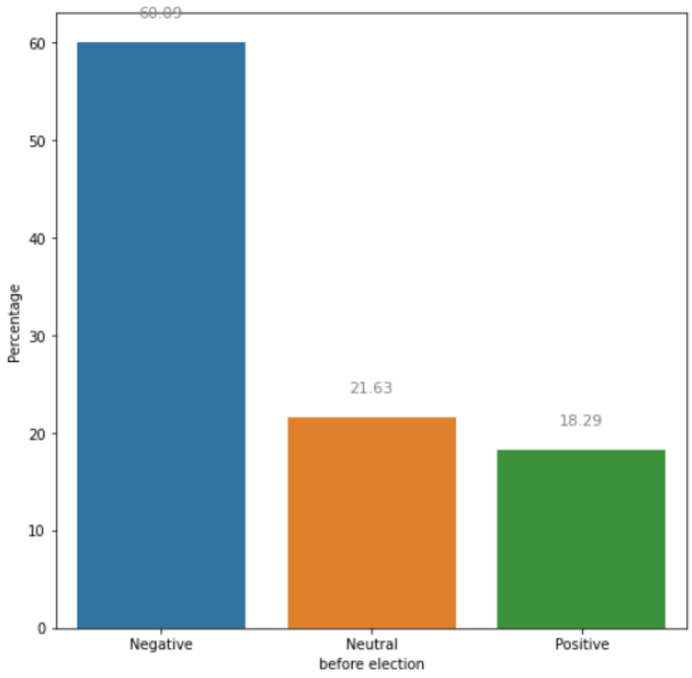
By the code below, I imported nltk with the following code to do semantic sentiment analysis and made histograms with pyplot.

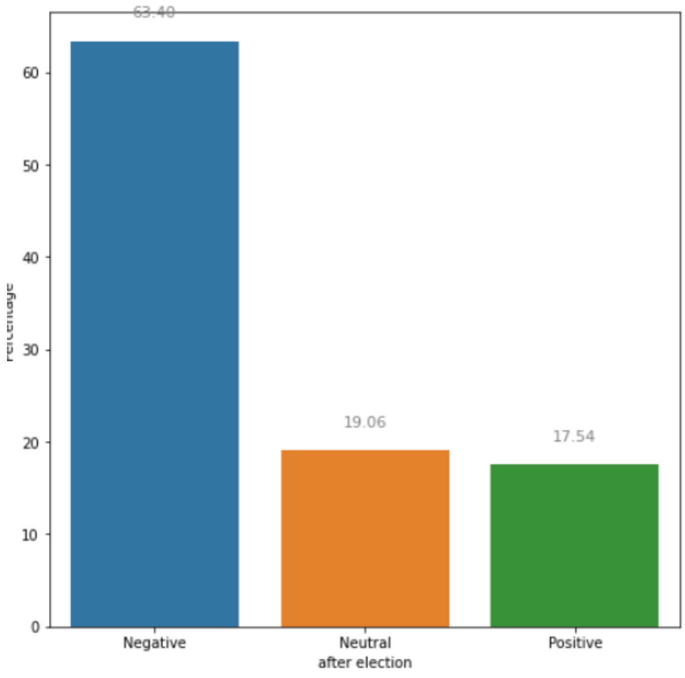
```
import nltk
nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from matplotlib import pyplot
```

Result









As can be seen from the above two graphs, in the three months before the U.S. election, people's emotions towards COVID-19 mainly were negative, accounting for 60.09%, followed by neutral feelings, accounting for 21.63%, and positive feelings, accounting for 18.29%. The general trend has not changed since Biden took office, and even the proportion of negative emotions has increased to 63.40%, compared with 19.06% for neutral feelings and 17.54% for positive emotions. So I can answer that my hypothesis is wrong and that the results of the U.S. election have not made the American people more optimistic about the coronavirus pandemic.

Conclusion

I can answer my hypothesis because of the chart above. We can see the American people's attitudes towards the spread of the virus at different periods with those two bar charts.

My research results show that although Biden and Trump have different attitudes towards the virus, it does not affect people's attitudes towards the epidemic. This means that a positive attitude can affect people less than powerful outcomes, such as economic recovery, successful vaccine development, herd immunity, etc.

Next, I might research why Biden's more positive attitude towards fighting the spread of the virus did not affect people's more positive emotions.

Critique of Design and Project

When I first designed the method, I planned to get the mentions table through GDELT and analyze the Cameo code to analyze the attitude of the American people towards COVID-19. However, when obtaining data, I often encounter the bug that the data can not be obtained at a particular moment. After trying to solve it but failing, I adopt the method of querying the keywords through the URL to get data.

In terms of algorithm, I initially considered using the TF-IDF algorithm to analyze word frequency or the cosine similarity algorithm to calculate the difference of data in two time periods. But when I found the Vader algorithm, I abandoned the previous scheme because the Vader algorithm was more suitable for my design.

Reflection

Pyspark is useful to analyze data, ensuring data processing with lightning speed.
Resilient Distributed Datasets (RDDs), it is distributed that data is distributed among the multiple nodes in a cluster.
Gdelt is monitoring recent webs, news, and broadcasts, every event with many attributes was recorded in Gdelt. It is convenient to extract and analyze them.

Reference:

https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vader_lexicon.txt

https://github.com/nltk/nltk/blob/develop/nltk/sentiment/vader.py

https://zhuanlan.zhihu.com/p/66206132

https://www.programcreek.com/python/example/100005/nltk.sentiment.vader.SentimentIntensityAnalyzer

https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

https://ourladylakes.org/588043-how-is-the-vader-compound-BSICAK