Final Project

Hongyang Yang

4/28/2020

Analysis of Relation between NYTimes Coronavirus Article & Public Opinion Polls

GitHub: https://github.com/timothyyang96/Data-Wrangling

Introduction

While the broadsheet decline with the popularity of social media such as Facebook and Twitter, and even influencers on Instagram or Youtube have millions of followers, whether the mainstream media could impact on the public remains a question. Therefore, the project is to make a study on the New York Times articles of Coronavirus and the public opinion polls to have a rough idea of if there is any relation between a broadsheet and its influence on the general public.

Sentiment analysis is extracting the perception, which indicates a positive, negative, or neutral sentiment, of people towards a particular issue, brand, scheme, etc., from textual data. It has a wide range of applications, from brand monitoring, product review analysis to policymaking.

For this project, the method of sentiment analysis and a simple count of newspaper coverage on the relevant topic will be implemented on the articles after using several sets of query words to retrieve search results.

Dataset

The datasets contain two parts: one is NYT search result whose schema is cleaned into four columns: `created_time`, `headline`, `web_url`, and `content`; the other is poll data collected from FiveThirtyEight, a website focusing on opinion poll analysis whose GitHub repositories have such data aggregation, and hence there are raw data for direct use.

For NYT articles, the keywords are limited to four prefixes of `coronavirus`, `Covid-19`, `pandemic`, and `epidemic`, and three suffixes of `Trump`, `infection`, and `economy`. The latter three are the most popular topics of the epidemic for U.S.A people and even worldwide. Thus, there are 12 kinds of combinations. On account of NYT developer API limiting each request no more than 2000 times, the date range should be set in three months from February to May, the search type in an article one, and the result order in a relevance sorting.

```
################################################################################
####            function - search news article with API                     ####
nytime = function (keyword) {
  searchQ = URLencode(keyword)
  url = paste('http://api.nytimes.com/svc/search/v2/articlesearch.json?q=',searchQ,
              '&begin_date=20200201&end_date=20200430&sort=relevance&fq=document_type:"article"&api-key=
  #get the total number of search results
  initialsearch = fromJSON(url,flatten = T)
  maxPages = round((initialsearch$response$meta$hits / 10) - 1)

  #try with the max page limit at 200
  maxPages = ifelse(maxPages >= 199, 199, maxPages)

  #creat a empty data frame
  df = data.frame(created_time=character(), headline=character(),web_url = character())

  #save search results into data frame
  for(i in 0:maxPages){
    #get the search results of each page
    tryCatch({
      nytSearch = fromJSON(paste0(url, "&page=", i), flatten = T)
      temp = data.frame(created_time = nytSearch$response$docs$pub_date,
                        headline = nytSearch$response$docs$headline.main,
                        web_url = nytSearch$response$docs$web_url)
      df=rbind(df,temp)
      Sys.sleep(6) #sleep for 6 second
    }, error = function(e) return(NA_character_))

  }
  return(df)
}
```

```
# retrieve full body
body_text <- function(url_list) {
  content_list = rep(NA, NROW(url_list))
  for(i in 1:NROW(url_list)){
    content_list[i] = tryCatch({
    xml2::read_html(url_list[i]) %>% rvest::html_nodes(".StoryBodyCompanionColumn p") %>%
  rvest::html_text() %>% paste(collapse = "\n")
    }, error = function(e) return(NA_character_))
  }

  return(content_list)
}
```

The next step is to clean and wrangle the search result, including removing rows with duplicates and missing values, converting the columns of `created_time` and `content` into the formats of **Date** and **Character**. The table below is the NYTimes search result of Trump and there are 2606 pieces in total. The Infection one has 2937 results while the Economy one has 2867 results.

```
## # A tibble: 6 x 4
##   created_time headline             web_url             content
##   <date>       <chr>                <chr>               <chr>
## 1 2020-03-17   Citing Coronavirus,~ https://www.nytime~ "WASHINGTON - The ~
## 2 2020-04-25   Coronavirus and the~ https://www.nytime~ "In an interview w~
## 3 2020-04-29   Coronavirus Casts U~ https://www.nytime~ "WASHINGTON - Two ~
## 4 2020-04-27   In Kayleigh McEnany~ https://www.nytime~ "WASHINGTON - Kayl~
## 5 2020-04-28   Trump Vows More Cor~ https://www.nytime~ "WASHINGTON - Pres~
## 6 2020-04-30   Trump’s Disinfectan~ https://www.nytime~ "SAN FRANCISCO - M~
```

```
## # A tibble: 6 x 3
##   created_time headline                       web_url
##   <date>       <chr>                          <chr>
## 1 2020-03-15   A Complete List of Trump’s At~ https://www.nytimes.com/2020~
## 2 2020-03-06   As Coronavirus Spreads, How Y~ https://www.nytimes.com/2020~
## 3 2020-03-04   Nursing Homes Are Starkly Vul~ https://www.nytimes.com/2020~
## 4 2020-03-02   Outbreak Strikes Seattle Area~ https://www.nytimes.com/2020~
## 5 2020-02-27   Iran Vice President Is One of~ https://www.nytimes.com/2020~
## 6 2020-02-27   Germany Tries to Solve a Coro~ https://www.nytimes.com/2020~
## # A tibble: 6 x 3
##   created_time headline                          web_url
##   <date>       <chr>                             <chr>
## 1 2020-03-19   Witnessing the Birth of the ~ https://www.nytimes.com/2020/~
## 2 2020-03-26   The Coronavirus Economy: Whe~ https://www.nytimes.com/2020/~
## 3 2020-04-29   Wall Street Rallies as the F~ https://www.nytimes.com/2020/~
## 4 2020-04-29   Fed Suggests Tough Road Ahea~ https://www.nytimes.com/2020/~
## 5 2020-04-29   Fed Weighs Next Steps to For~ https://www.nytimes.com/2020/~
## 6 2020-04-26   There's Really Only One Way ~ https://www.nytimes.com/2020/~
```

Meanwhile, for the opinion poll data, FiveThirtyEight collects the whole package of pollsters and sample information and calculates the approval and disapproval rates for the subject of Trump, Infection, and Economy. The raw data is presented as follows. The columns such as `startdate`, `enddate`, `approve_adjusted`, and `disapprove_adjusted` would be selected for further processing. The final table should look like a key-value pair, {`average_date`: [`approve_adjusted`, `disapprove_adjusted`]}. Moreover, for the concern-over-economy-or-infection poll data, there are also columns of `startdate` and `enddate`, while four degrees from `very`, `somewhat`, `not_very`, to `not_at_all` should also be included in the processed table.

```
## # A tibble: 6 x 19
##   subject modeldate party startdate enddate pollster grade samplesize
##   <chr>   <chr>     <chr> <chr>     <chr>   <chr>    <chr>      <dbl>
## 1 Trump   5/4/2020  D     2/2/2020  2/4/20~ YouGov   B-           523
## 2 Trump   5/4/2020  D     2/7/2020  2/9/20~ Morning~ B/C          817
## 3 Trump   5/4/2020  D     2/9/2020  2/11/2~ YouGov   B-           510
## 4 Trump   5/4/2020  D     2/16/2020 2/18/2~ YouGov   B-           529
## 5 Trump   5/4/2020  D     2/23/2020 2/25/2~ YouGov   B-           525
## 6 Trump   5/4/2020  D     2/24/2020 2/26/2~ Morning~ B/C          786
## # ... with 11 more variables: population <chr>, weight <dbl>,
## #   influence <dbl>, multiversions <chr>, tracking <lgl>, approve <dbl>,
## #   disapprove <dbl>, approve_adjusted <dbl>, disapprove_adjusted <dbl>,
## #   timestamp <chr>, url <chr>
```

```
## # A tibble: 6 x 23
##   subject modeldate party startdate enddate pollster grade samplesize
##   <chr>   <chr>     <chr> <chr>     <chr>   <chr>    <chr>      <dbl>
## 1 concer~ 5/4/2020  all   1/27/2020 1/29/2~ Morning~ B/C         2202
## 2 concer~ 5/4/2020  all   1/31/2020 2/2/20~ Morning~ B/C         2202
## 3 concer~ 5/4/2020  all   2/7/2020  2/9/20~ Morning~ B/C         2200
## 4 concer~ 5/4/2020  all   2/13/2020 2/18/2~ Kaiser ~ <NA>        1207
## 5 concer~ 5/4/2020  all   2/24/2020 2/26/2~ Morning~ B/C         2200
## 6 concer~ 5/4/2020  all   2/27/2020 2/27/2~ SurveyM~ <NA>        1051
## # ... with 15 more variables: population <chr>, weight <dbl>,
## #   influence <dbl>, multiversions <lgl>, tracking <lgl>, very <dbl>,
## #   somewhat <dbl>, not_very <dbl>, not_at_all <dbl>, very_adjusted <dbl>,
## #   somewhat_adjusted <dbl>, not_very_adjusted <dbl>,
## #   not_at_all_adjusted <dbl>, timestamp <chr>, url <chr>
```

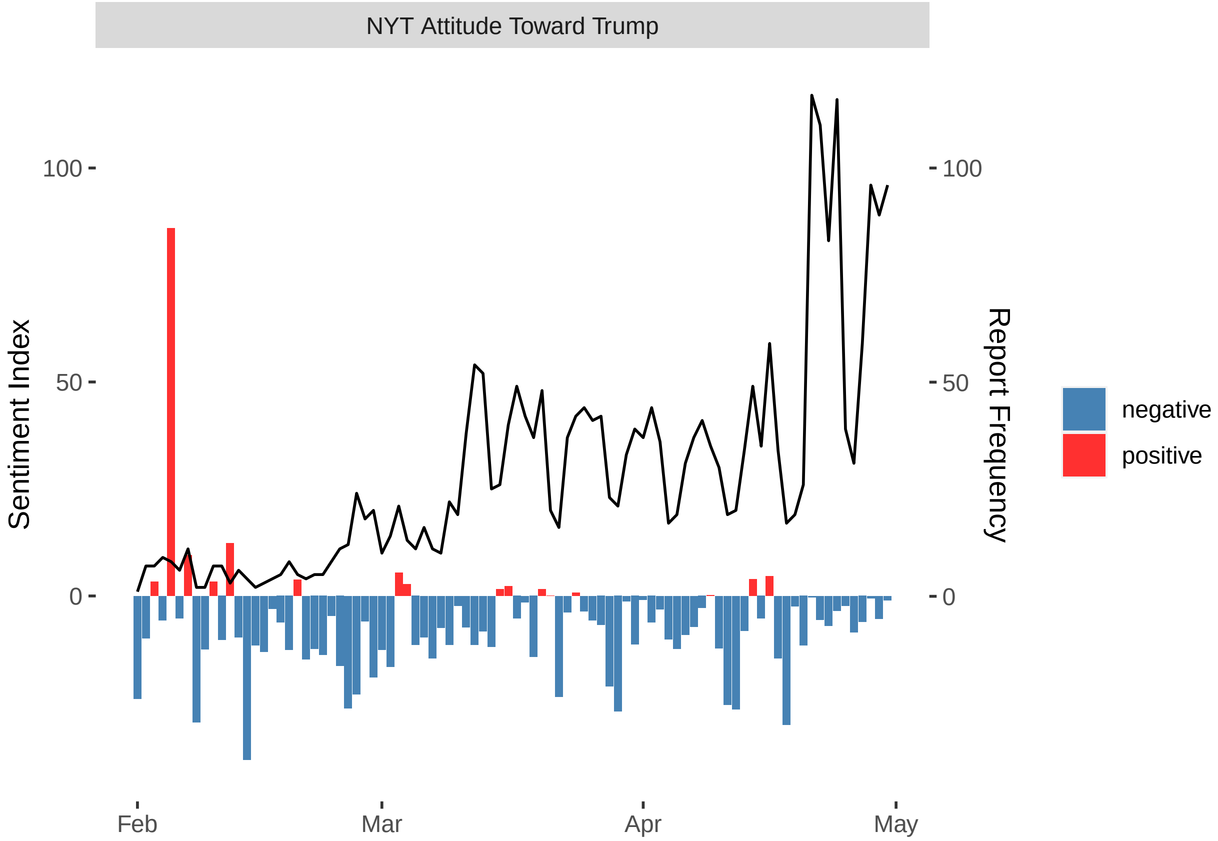
Data Analysis Strategy

Here comes the basic sentiment analysis of context with `textdata`. For this project, the lexicon picked is `AFINN` which assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment.

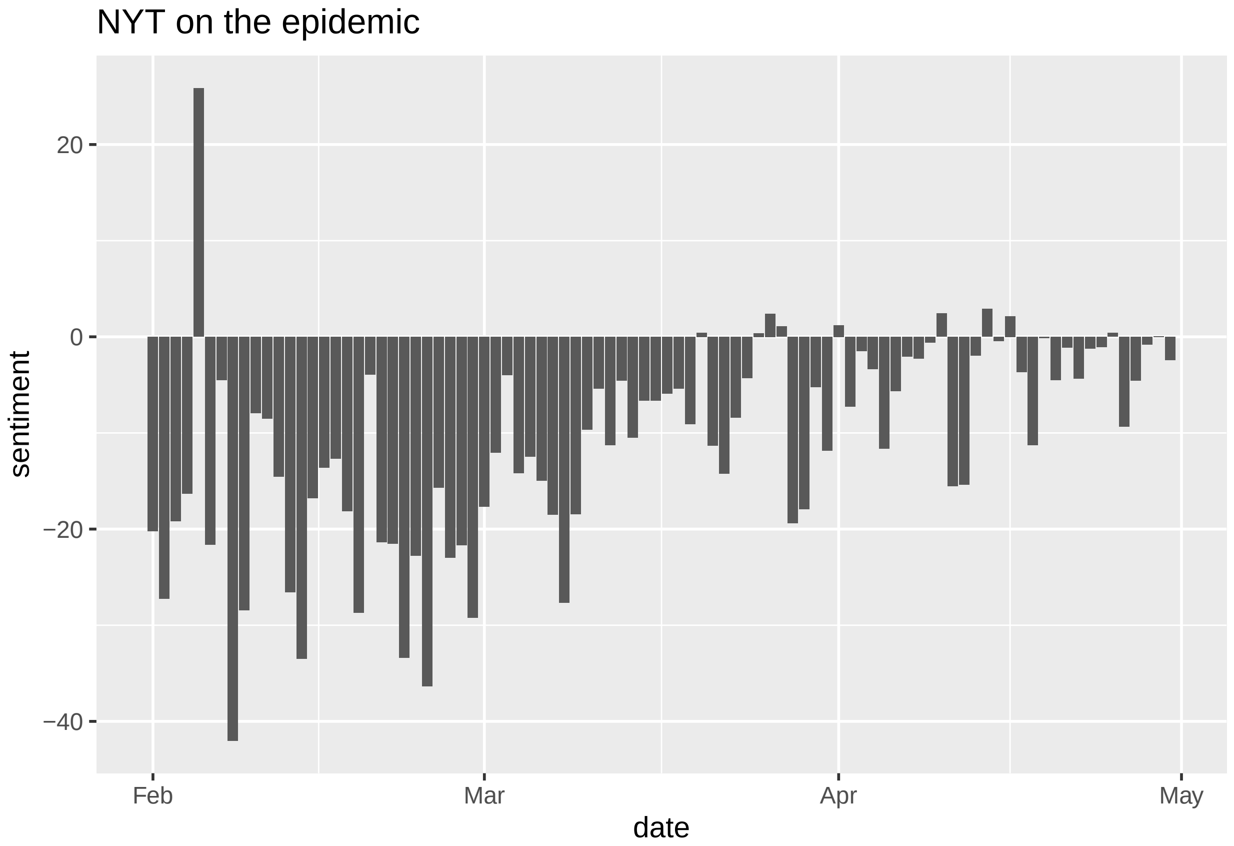
```
melted_corpus = function(complete_corpus) {
  sentiment_by_day = complete_corpus %>%
    select(web_url, content) %>%
    unnest_tokens(word, content) %>%
    anti_join(get_stopwords(), by = "word") %>%
    inner_join(lexicon_afinn(), by = "word") %>%
    group_by(web_url) %>%
    summarize(sentiment = sum(value)) %>%
    left_join(complete_corpus, by = "web_url") %>%
    select(created_time, sentiment) %>%
    group_by(date = created_time) %>%
    summarize(sentiment = mean(sentiment), n = n())

  return(sentiment_by_day)
}
```

The processed corpus using `Afinn` is the content of the epidemic and Trump. For comparison, the time series of sentiment analysis with the only epidemic is also attached as follows. Since the epidemic related articles might be already classified as negative, the NYTimes attitude toward Trump's dealing with the epidemic could have been a little more positive than the result shows. Besides, the count of daily coverage is to show the extent of NYTimes focusing on the topic for reference.

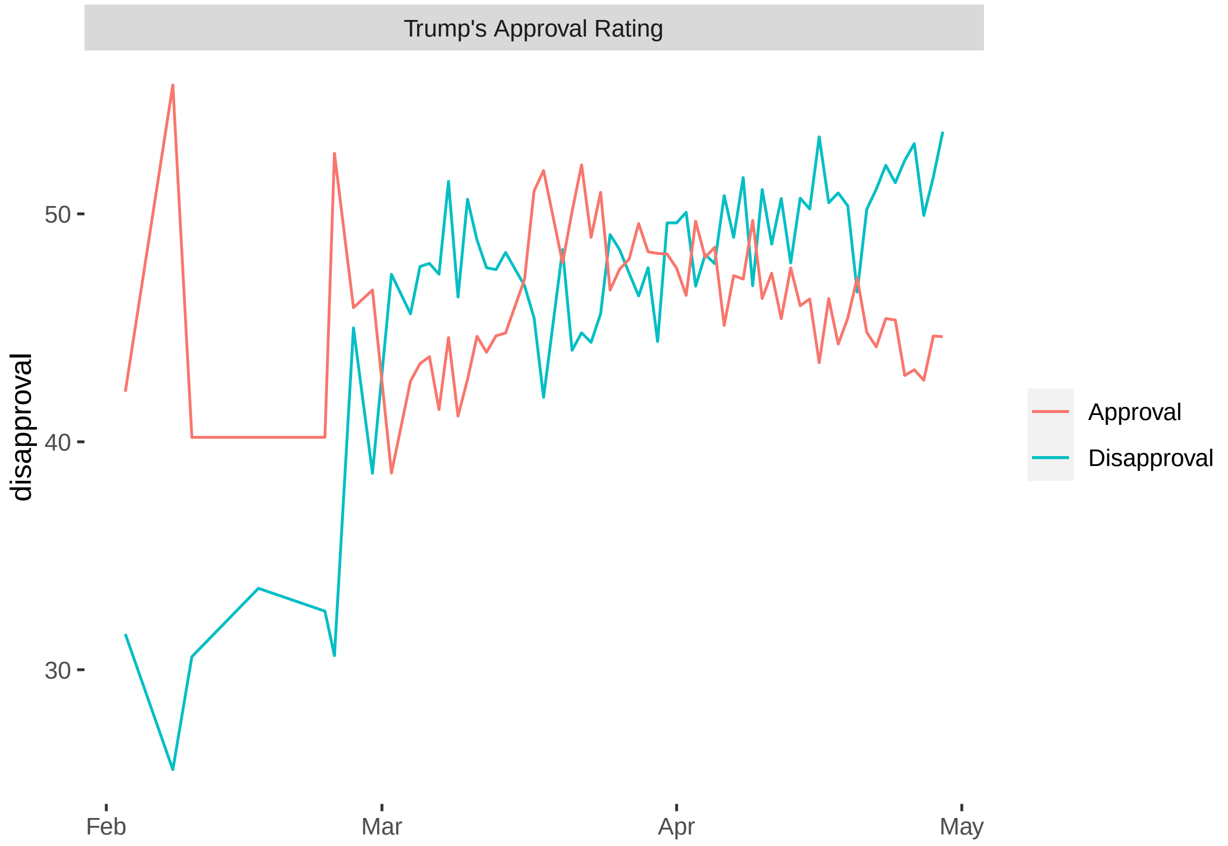
According to the chart below, the negative attitude of the NYTimes toward Trump dealing with the epidemic is unevenly distributed, however, it is not at all surprising that the reports from the media rarely hold a positive attitude in such an obvious way due to a similar or even more negative position on the word of epidemic only.



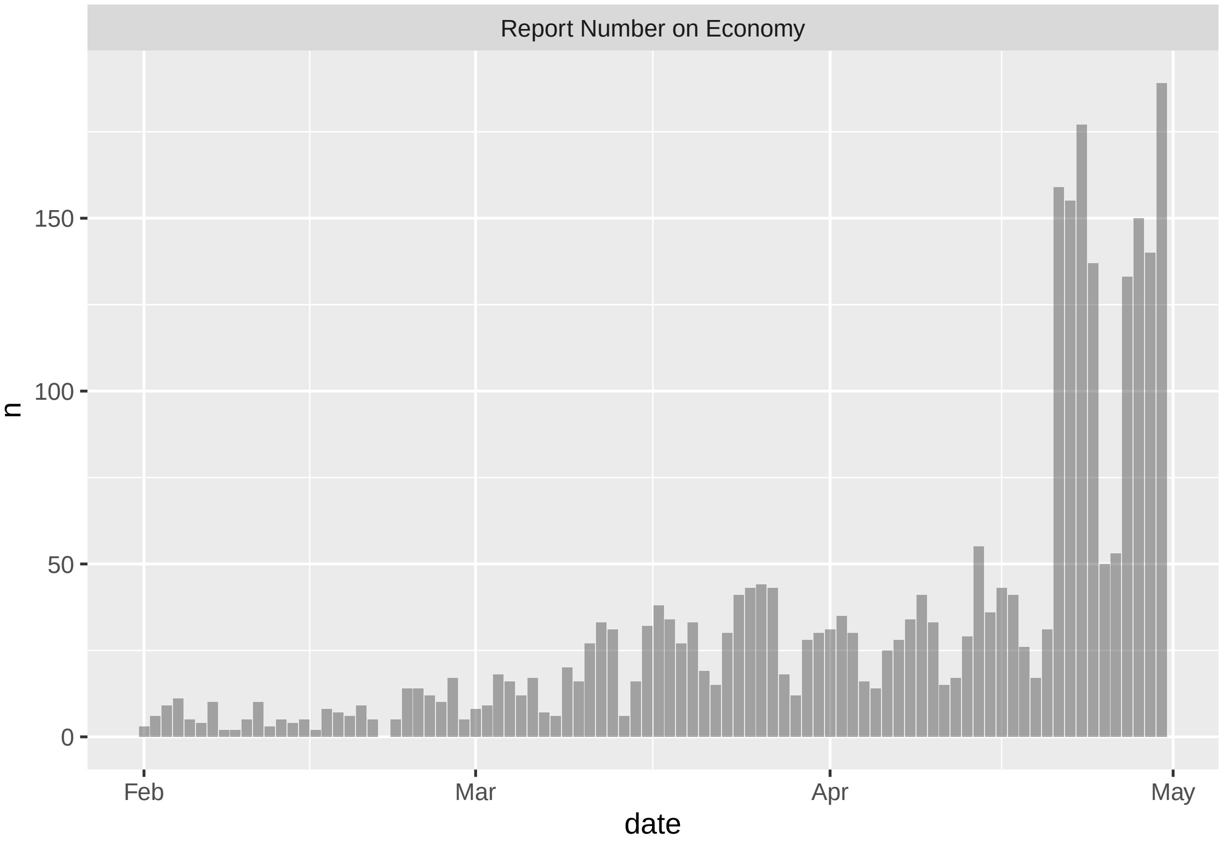
```
df %>% melted_corpus() %>%
  ggplot(aes(x = date, y = sentiment)) +
  geom_bar(stat = "identity", position = "identity") +
  ggtitle("NYT on the epidemic")
```

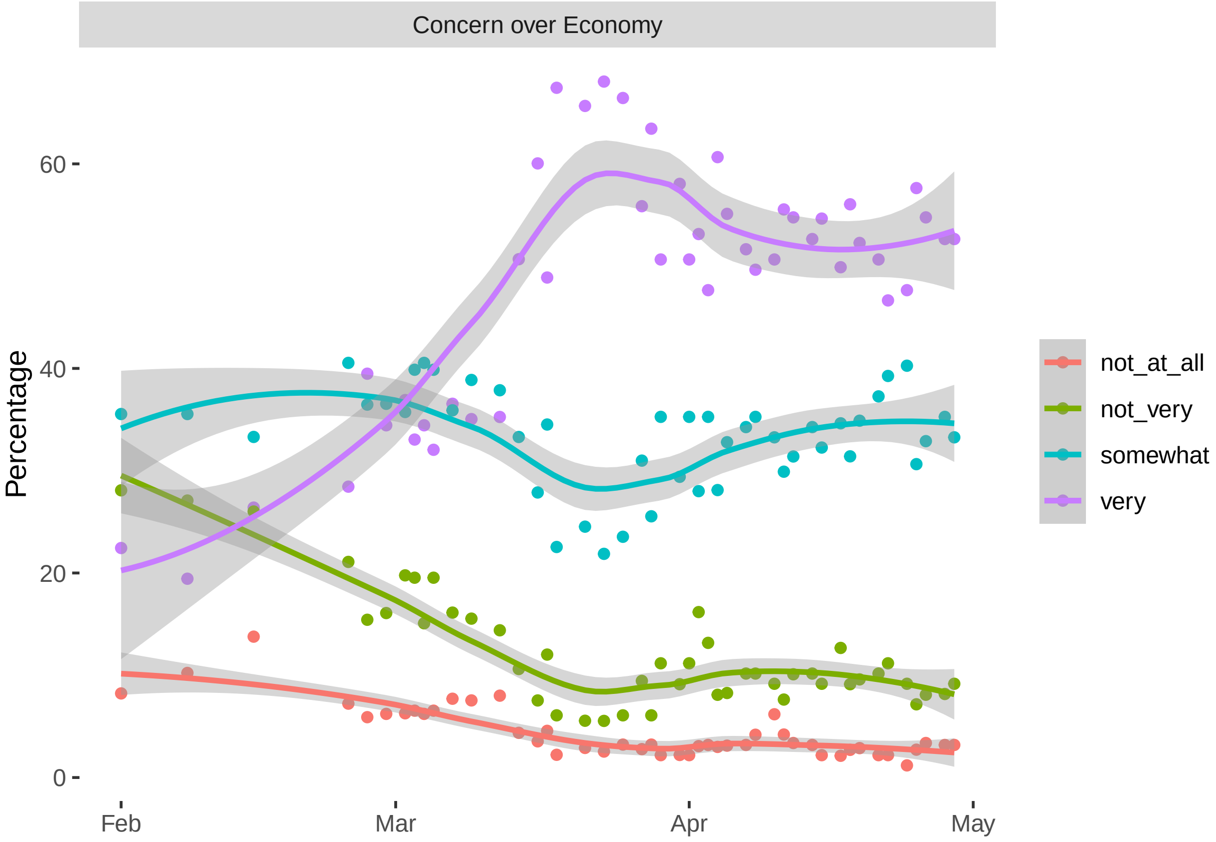


The graph of Trump's approval rating from February to May presents a searing picture, starting with the first month's figures, in which the support leads the opposition by a large margin, yet such figures don't make sense, given that there was no epidemic in the United States at that time; by the second month, the support lags first and then begins to lead the opposition; and by April, the support keeps slipping and the opposition accounts for almost half the voices.

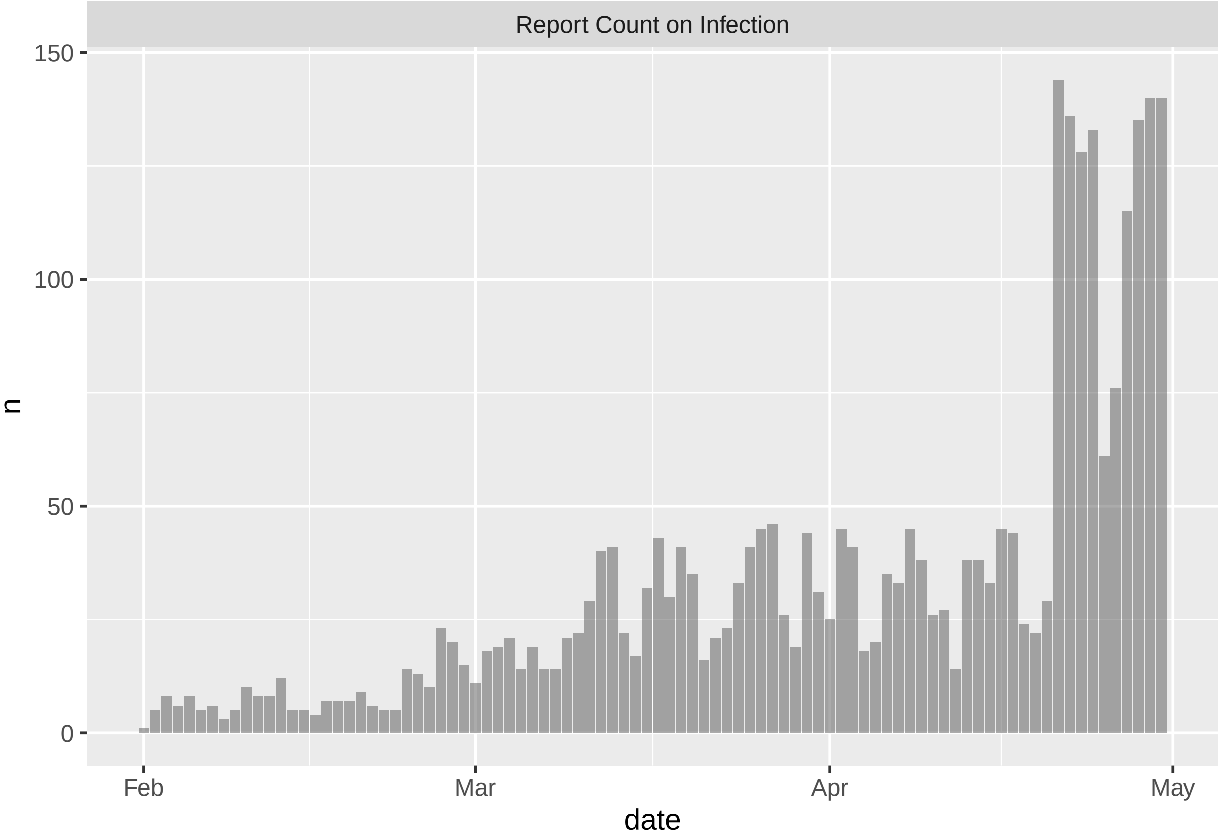


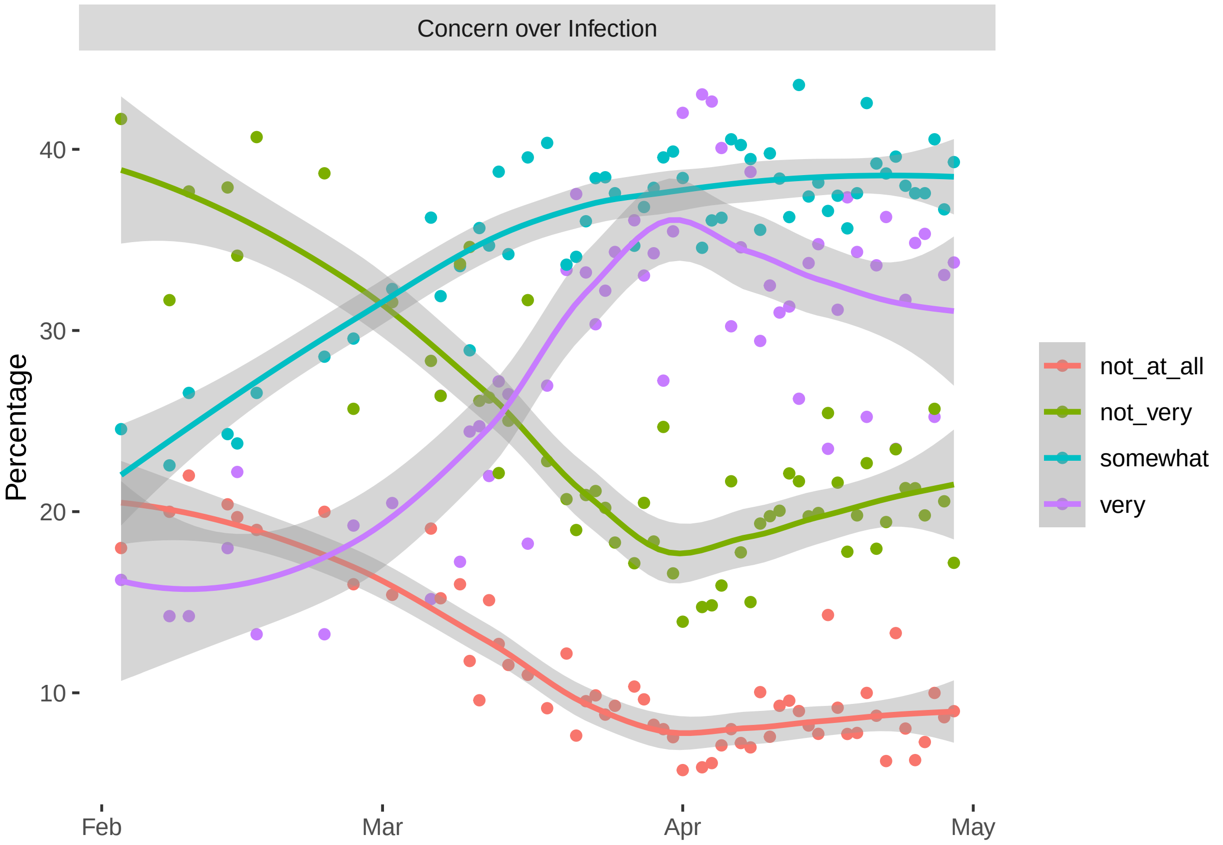
On the question of whether the American public is more worried about the infection or the economy, the following polling chart gives a very clear answer. During the most anxious period of late March, which was also the time of the U.S. outbreak, people were far more concerned about the economy than they were about being infected, with 60 percent expressing great concern about the economic situation, while only 40 percent were very nervous about whether they might be infected.


Report Number on Economy
n
150
100
50
0
Feb
Mar
Apr
May
date


Concern over Economy
Percentage
60
40
20
0
Feb
Mar
Apr
May
not_at_all
not_very
somewhat
very

The New York Times, meanwhile, showed a similar trend in its coverage of either the infection or the economy, which is when media attention to both topics peaked in mid to late April.





Looking at the four graphs above, it is difficult to conclude that the New York Times' topic setting has an impact on the source of public anxiety. However, the negative sentiment about Trump's handling of the epidemic was somewhat linked to the gradual climb of the opposition in April, which was also the month with the most intense media coverage.

Conclusion and Future Work

Challenge

1. Since the NYT Developer API limits the number of requests per keyword to no more than 2,000, and thus uses multiple prefixes with suffixes to get the raw data. Even so, the processed data have such a flaw of fewer than 3,000 pieces each that cannot be ignored. Despite limiting the sort method to relevance when searching, the articles searched still shows a much higher volume in recent time than before mid-April, which may lead to that it is hard to discern whether the media shape public opinion. Also, since both the NYT reports and the opinion data are time series, it needs more work to process the raw data to find a correlation between the two.
2. The sentiment analysis of the articles in this study used the `afinn` lexicon and as a result, it appears that most of the articles exhibited negative attitudes and although the epidemic search result itself was singled out for comparison, it was not possible to derive media-specific attitudes and variations in the president's handling of the epidemic in a simple 1+1=2 manner. The first is to compare other lexicon results with the `afinn`'s, and the second is to compare media coverage of the president himself in the same period to see if the media has similar attitudes toward other issues of the president.

Improvement

1. On the basis of the polls, it is possible to add data mining on related topics on Twitter, which can directly show public opinion.
2. In terms of methodology, it is possible to use advanced NLP approaches, including Naive Bayes, N-grams, and Neural Networks, to increase the accuracy of sentiment analysis about the New York Times reports, while having the data from tweets on Twitter to make the relation between the mainstream media and public opinion more clearly drawn.