

Tweet Analysis

Ashlyn Jew

5/14/2020

Text and sentiment analysis of tweets including keywords “anti-asian” and “virus” (scraped from Jan 24th to May 10th).

Tokenize text, remove irrelevant words, get count of each word

```
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1      v purrr   0.3.3
## v tibble  2.1.3      v dplyr  0.8.3
## v tidyr   1.0.2      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.4.0
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
library(tidytext)
```

```
## Warning: package 'tidytext' was built under R version 3.6.3
```

```
antiasian <- read_csv("anti-asian-3.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
```

```
## cols(
##   .default = col_character(),
##   X1 = col_double(),
##   X = col_double(),
##   user_id = col_double(),
##   tweet_id = col_double(),
##   timestamp = col_datetime(format = ""),
##   timestamp_epochs = col_double(),
##   has_media = col_logical(),
##   video_url = col_logical(),
##   likes = col_double(),
##   retweets = col_double(),
##   replies = col_double(),
##   is_replied = col_logical(),
##   is_reply_to = col_logical(),
##   parent_tweet_id = col_logical(),
##   `data$tweet_id` = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
#without stop words
```

```
my_stop_words <- stop_words %>% select(-lexicon) %>%  
  bind_rows(data.frame(word = c("asian", "anti", "virus", "https", "twitter.com", "status", "19", "pic."
```

```
## Warning in bind_rows(x, .id): binding character and factor vector,  
## coercing into character vector
```

```
ordered_text_count <- antiasian %>%  
  select(text) %>%  
  unnest_tokens(word, text) %>%  
  add_count(word) %>%  
  distinct() %>%  
  anti_join(my_stop_words) %>%  
  arrange(desc(n))
```

```
## Joining, by = "word"
```

```
ordered_text_count
```

```
## # A tibble: 4,848 x 2  
##   word      n  
##   <chr>    <int>  
## 1 chinese    354  
## 2 racism     338  
## 3 coronavirus 227  
## 4 racist     164  
## 5 people     159  
## 6 covid      139  
## 7 china      129  
## 8 corona     125  
## 9 hate       125  
## 10 americans 122  
## # ... with 4,838 more rows
```

Histogram of Words That Appear in “Anti-Asian Virus” Tweets

```
ordered_text_count %>%  
  filter(n > 110) %>%  
  mutate(rel_freq = n/sum(n)) %>%  
  mutate(word = reorder(word, rel_freq)) %>%  
  ggplot(aes(word, rel_freq)) +  
  geom_col(fill = "#6baed6") +  
  xlab(NULL) +  
  ylab("Relative Frequency of Word") +  
  ggtitle("Histogram of Words That Appear in \"Anti-Asian Virus\" Tweets") +  
  coord_flip()
```

A horizontal bar chart showing the relative frequency of the top 100 words in the corpus. The x-axis is labeled 'Relative Frequency of Word' and ranges from 0.00 to 0.15. The bars are blue and arranged in descending order of frequency. The first bar (the word 'the') has a relative frequency of approximately 0.18, and the 100th bar has a relative frequency of approximately 0.06.

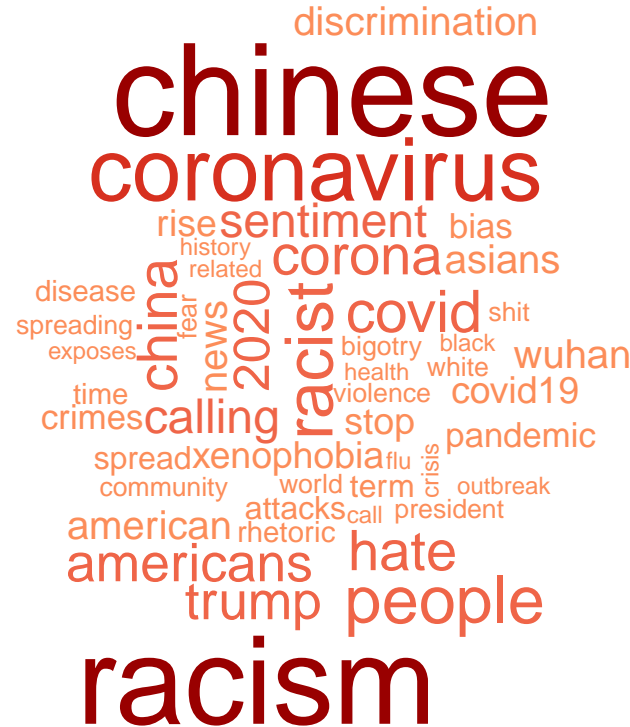
Rank	Word	Relative Frequency
1	the	0.18
2	and	0.17
3	of	0.11
4	in	0.08
5	to	0.08
6	a	0.07
7	is	0.065
8	was	0.065
9	on	0.065
10	at	0.065
11	but	0.065
12	by	0.065
13	from	0.065
14	with	0.065
15	as	0.065
16	for	0.065
17	and	0.065
18	the	0.065
19	of	0.065
20	in	0.065
21	to	0.065
22	a	0.065
23	is	0.065
24	was	0.065
25	on	0.065
26	at	0.065
27	but	0.065
28	by	0.065
29	from	0.065
30	with	0.065
31	as	0.065
32	for	0.065
33	and	0.065
34	the	0.065
35	of	0.065
36	in	0.065
37	to	0.065
38	a	0.065
39	is	0.065
40	was	0.065
41	on	0.065
42	at	0.065
43	but	0.065
44	by	0.065
45	from	0.065
46	with	0.065
47	as	0.065
48	for	0.065
49	and	0.065
50	the	0.065
51	of	0.065
52	in	0.065
53	to	0.065
54	a	0.065
55	is	0.065
56	was	0.065
57	on	0.065
58	at	0.065
59	but	0.065
60	by	0.065
61	from	0.065
62	with	0.065
63	as	0.065
64	for	0.065
65	and	0.065
66	the	0.065
67	of	0.065
68	in	0.065
69	to	0.065
70	a	0.065
71	is	0.065
72	was	0.065
73	on	0.065
74	at	0.065
75	but	0.065
76	by	0.065
77	from	0.065
78	with	0.065
79	as	0.065
80	for	0.065
81	and	0.065
82	the	0.065
83	of	0.065
84	in	0.065
85	to	0.065
86	a	0.065
87	is	0.065
88	was	0.065
89	on	0.065
90	at	0.065
91	but	0.065
92	by	0.065
93	from	0.065
94	with	0.065
95	as	0.065
96	for	0.065
97	and	0.065
98	the	0.065
99	of	0.065
100	in	0.065

```
library(wordcloud)

## Warning: package 'wordcloud' was built under R version 3.6.3

## Loading required package: RColorBrewer

wordcloud(ordered_text_count$word, freq = ordered_text_count$n, max.words = 50, colors = c
```



Assigning different sentiment values

```
library(textdata)
```

```
## Warning: package 'textdata' was built under R version 3.6.3
```

```
#assigns numeric values
```

```
afinn_sent <- get_sentiments("afinn")
```

#negative or positive

```
bing_sent <- get_sentiments("bing")
```

#emotion

```
nrc_sent <- get_sentiments("nrc")
```

#emotion words

```
ordered_text_count %>%
```

```
left_join(nrc_sent) %>%
```

```
filter(!is.na(sentiment), word != "trump")
```

```
## Joining, by = "word"
```

```
## # A tibble: 1,822 x 3
##   word          n sentiment
##   <chr>        <int> <chr>
## 1 hate          125 anger
## 2 hate          125 disgust
## 3 hate          125 fear
## 4 hate          125 negative
## 5 hate          125 sadness
## 6 discrimination    79 anger
## 7 discrimination    79 disgust
## 8 discrimination    79 fear
## 9 discrimination    79 negative
## 10 discrimination    79 sadness
## # ... with 1,812 more rows
```

```
#summary of emotion words
emotion_summary <- ordered_text_count %>%
  left_join(nrc_sent) %>%
  filter(!is.na(sentiment), word != "trump") %>%
  group_by(sentiment) %>%
  summarise(n = n())
```

```
## Joining, by = "word"
```

```
emotion_summary
```

```
## # A tibble: 10 x 2
##   sentiment      n
##   <chr>        <int>
## 1 anger          196
## 2 anticipation   114
## 3 disgust        143
## 4 fear           217
## 5 joy             80
## 6 negative       385
## 7 positive       300
## 8 sadness        151
## 9 surprise        70
## 10 trust         166
```

```
#affin
ordered_text_count %>%
  left_join(afinn_sent) %>%
  filter(!is.na(value)) %>%
  summarise(mean_value = mean(value))
```

```
## Joining, by = "word"
```

```
## # A tibble: 1 x 1
##   mean_value
##   <dbl>
## 1    -0.951
```

```
#bing
bing_summary <- ordered_text_count %>%
  left_join(bing_sent) %>%
  filter(!is.na(sentiment), word != "trump") %>%
  group_by(sentiment)
```

```
## Joining, by = "word"
```

```
bing_summary
```

```
## # A tibble: 586 x 3
## # Groups:   sentiment [2]
##   word          n sentiment
##   <chr>      <int> <chr>
## 1 racism      338 negative
## 2 racist      164 negative
## 3 hate        125 negative
## 4 discrimination 79 negative
## 5 bias         63 negative
## 6 attacks       48 negative
## 7 rhetoric      45 negative
## 8 bigotry       33 negative
## 9 fear          32 negative
## 10 shit         30 negative
## # ... with 576 more rows
```

Histogram of Tweet Sentiments

```
sum(bing_summary[bing_summary$sentiment == "negative", ]$n)
```

```
## [1] 2114
```

```
sum(bing_summary$n)
```

```
## [1] 2409
```

```
ggplot(bing_summary, aes(x = sentiment, y = n/sum(n))) +
  geom_bar(stat="identity", fill = "#7bccc4") +
  xlab("Sentiment") +
  ylab("Relative Frequency") +
  ggtitle("Histogram of Tweet Sentiments" )
```

