Assignment 3 - Sentiment Analysis II

Marie Rivers

4/26/2022

```
library(quanteda)
library(quanteda.sentiment)
library(quanteda.textstats)
library(tidyverse)
library(tidytext)
library(lubridate)
library(wordcloud) #visualization of common words in the data set
library(reshape2)
library(sentimentr)
library(kableExtra)
```

This assignment uses tweet data for the term 'IPCC'

```
raw_tweets <- read.csv("https://raw.githubusercontent.com/MaRo406/EDS_231-text-sentiment/main/dat/IPCC_

dat<- raw_tweets[,c(4,6)] # Extract Date and Title fields

tweets <- tibble(text = dat$Title,
                  id = seq(1:length(dat$Title)),
                 date = as.Date(dat$Date,'%m/%d/%y'))

#clean up the URLs from the tweets (people linking to news articles and such)
tweets$text <- gsub("http[^[:space:]]*", "",tweets$text) # substitute http links with nothing
tweets$text <- str_to_lower(tweets$text)
```

1. Think about how to further clean a twitter data set. Let's assume that the mentions of twitter accounts is not useful to us. Remove them from the text field of the tweets tibble.

```
tweets_clean <- tweets %>%
  mutate(text_clean = text)  # keeping a column of the original text as a check

tweets_clean$text_clean <- gsub("@[^[:space:]]*", "", tweets_clean$text_clean)
head(tweets_clean)
```

```
## # A tibble: 6 x 4
##   text                                                      id date       text_clean
##   <chr>                                                  <int> <date>     <chr>
## 1 "thank you, followers, for the great photo sugges~     1 2022-04-01 "thank yo~
## 2 "greenpeace: the real solution to the climate cri~     2 2022-04-01 "greenpea~
## 3 "governments have a responsibility to ensure that~     3 2022-04-01 "governme~
## 4 "next week, the ipcc will publish a new report de~     4 2022-04-01 "next wee~
## 5 "live stream of virtual ipcc press conference rel~     5 2022-04-01 "live str~
## 6 "attention journalists: the deadline for embargoe~     6 2022-04-01 "attentio~
```

2. Compare the ten most common terms in the tweets per day. Do you notice anything interesting?

```
#tokenize tweets to individual words
words <- tweets_clean %>%
  select(id, date, text_clean) %>%
  unnest_tokens(output = word, input = text_clean, token = "words") %>%
  anti_join(stop_words, by = "word")
```

```
words_count <- words %>%
  count(date, word)

top_ten_per_day <- words_count %>%
  group_by(date) %>%
  top_n(10, n)
```

```
top_ten_table = aggregate(top_ten_per_day$word, list(top_ten_per_day$date), paste, collapse=", ") %>%
  rename(Date = Group.1) %>%
  rename(top_words = x) %>%
  kable(col.names = c("Date", "Top 10 Words")) %>%
  kable_paper(full_width = TRUE) %>%
  row_spec(c(0), background = "lightgray")
```

```
## Warning in latex_new_row_builder(target_row, table_info, bold, italic,
## monospace, : Setting full_width = TRUE will turn the table into a tabu
## environment where colors are not really easily configable with this package.
## Please consider turn off full_width.
```

```
top_ten_table
```

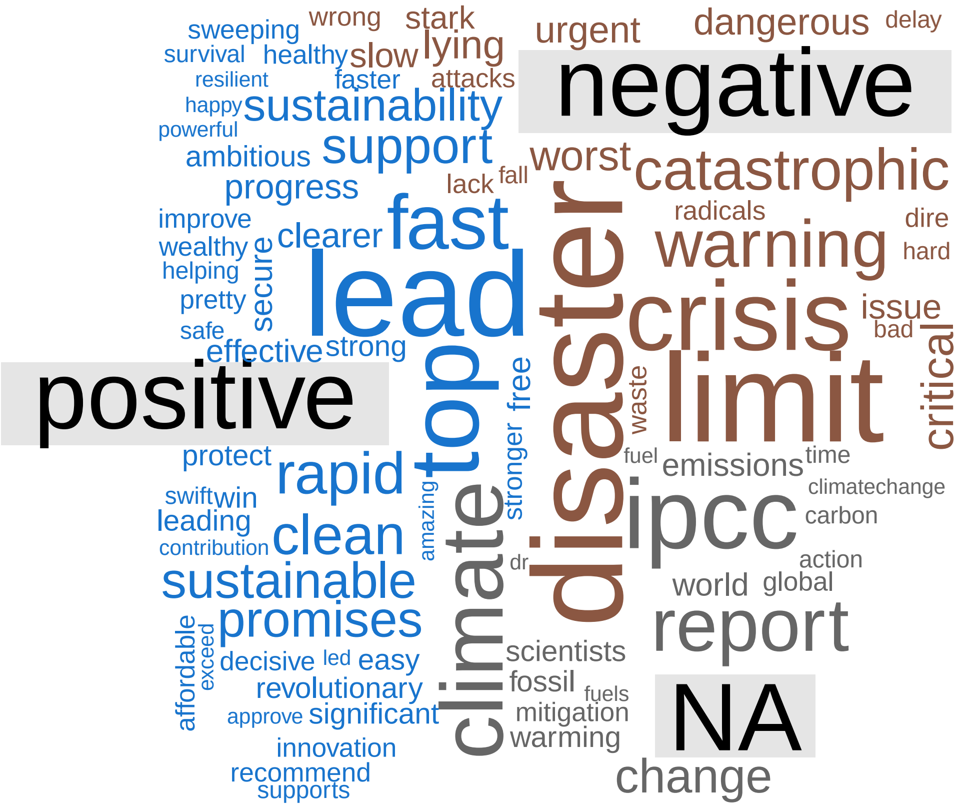
Date	Top 10 Words
2022-04-01	carbon, change, climate, climatereport, fossil, ipcc, monday, rapid, read, report, upcoming
2022-04-02	04, 2022, carbon, change, climate, emissions, gt, ipcc, monday, report, scenarios
2022-04-03	aitt, authors, climate, dasgupta, dipak, dr, fossil, hosted, ipcc, joyashree, lead, lifespaces, mahindra, mitigation, purushottam, reminder, report, roy, scientists, set, space, sunita, teri, twitter, unpack
2022-04-04	change, climate, emissions, fossil, ipcc, limit, report, scientists, warming, world
2022-04-05	action, change, climate, emissions, fossil, global, ipcc, report, warming, world
2022-04-06	change, climate, crisis, emissions, fossil, ipcc, listen, oil, report, scientists, world
2022-04-07	change, climate, climatechange, emissions, energy, global, ipcc, report, time, world
2022-04-08	action, carbon, change, climate, climatechange, emissions, global, ipcc, released, report, warming, world
2022-04-09	carbon, change, climate, emissions, fossil, fuels, global, ipcc, it’s, oil, report, warming, world
2022-04-10	change, climate, emissions, fossil, fuel, global, ipcc, report, time, warming

3. Adjust the wordcloud in the “wordcloud” chunk by coloring the positive and negative words so they are identifiable.

```
#load sentiment lexicon
bing_sent <- get_sentiments('bing')

words_sent <- words %>%
  left_join(bing_sent, by = "word") %>%
  left_join(
    tribble(
      ~sentiment, ~sent_score,
      "positive", 1,
      "negative", -1),
    by = "sentiment")

wordcloud_sent <- words_sent %>%
  count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("lightsalmon4", "dodgerblue3", "gray40"),
                   max.words = 100)
```



4. Let's say we are interested in the most prominent entities in the Twitter discussion. Which are the top 10 most tagged accounts in the data set. Hint: the "explore_hashtags" chunk is a good starting point.

```r
corpus <- corpus(dat$Title) #enter quanteda
# corpus is a collection of documents (ie tweets) with metadata
```

```r
tagged_tweets <- tokens(corpus, remove_punct = TRUE) %>%
               tokens_keep(pattern = "@*")
dfm_tagged<- dfm(tagged_tweets)

tstat_freq <- textstat_frequency(dfm_tagged, n = 10)

#tidytext gives us tools to convert to tidy from non-tidy formats
tagged_tib<- tidy(dfm_tagged)

tagged_tib %>%
   count(term) %>%
   with(wordcloud(term, n, max.words = 10))
```

Tag	Count
@antonioguterres	16
@conversationedu	10
@ipcc	9
@ipcc_ch	131
@logicalindians	38
@nytimes	14
@potus	13
@un	12
@yahoo	14
@youtube	11

@conversationedu
@ipcc_ch
@logicalindians
@yahoo
@potus
@ipcc
@antonioguterres
@un
@nytimes
@youtube

```
top_ten_tags <- tagged_tib %>%
  count(term) %>%
  top_n(10, n) %>%
  kable(col.names = c("Tag", "Count")) %>%
  kable_paper(full_width = FALSE) %>%
  row_spec(c(0), background = "lightgray")
top_ten_tags
```

5. The Twitter data download comes with a variable called "Sentiment" that must be calculated by Brandwatch. Use your own method to assign each tweet a polarity score (Positive, Negative, Neutral) and compare your classification to Brandwatch's (hint: you'll need to revisit the "raw_tweets" data frame).

```
dat2<- raw_tweets[,c(4, 6, 10)] # Extract Date, Title, and Sentiment fields

tweets2 <- tibble(text = dat2$Title,
                  element_id = seq(1:length(dat2$Title)),
                 date = as.Date(dat2$Date,'%m/%d/%y'),
                 sent_brandwatch = dat2$Sentiment)
```

```
sent_method2 <- sentiment_by(tweets2$text)

tweets2 <- inner_join(tweets2, sent_method2, by = "element_id") %>%
  mutate(sent_method2 = case_when(
```

Comparison	Count	Percent
both negative	184	7.6
both neutral	241	10.0
both positive	18	0.7
brandwatch negative, method 2 neutral	4	0.2
brandwatch negative, method 2 positive	62	2.6
brandwatch neutral, method 2 negative	739	30.7
brandwatch neutral, method 2 positive	1162	48.2
brandwatch positive, method 2 negative	1	0.0

```
    ave_sentiment < 0 ~ "negative",
    ave_sentiment > 0 ~ "positive",
    ave_sentiment == 0 ~ "neutral"))
```

```
sent_method_comparison <- tweets2 %>%
  mutate(sent_comparison = case_when(
    sent_brandwatch == "positive" & sent_method2 == "positive" ~ "both positive",
    sent_brandwatch == "negative" & sent_method2 == "negative" ~ "both negative",
    sent_brandwatch == "neutral" & sent_method2 == "neutral" ~ "both neutral",
    sent_brandwatch == "positive" & sent_method2 == "negative" ~ "brandwatch positive, method 2 negativ
    sent_brandwatch == "positive" & sent_method2 == "neutral" ~ "brandwatch positive, method 2 neutral"
    sent_brandwatch == "neutral" & sent_method2 == "positive" ~ "brandwatch neutral, method 2 positive"
    sent_brandwatch == "neutral" & sent_method2 == "negative" ~ "brandwatch neutral, method 2 negative"
    sent_brandwatch == "negative" & sent_method2 == "positive" ~ "brandwatch negative, method 2 positiv
    sent_brandwatch == "negative" & sent_method2 == "neutral" ~ "brandwatch negative, method 2 neutral")

sent_method_comparison_counts <- sent_method_comparison %>%
  count(sent_comparison)
```

```
sent_method_comparison_counts2 <- sent_method_comparison %>%
  group_by(sent_brandwatch, sent_method2) %>%
  summarise(count = n())
```

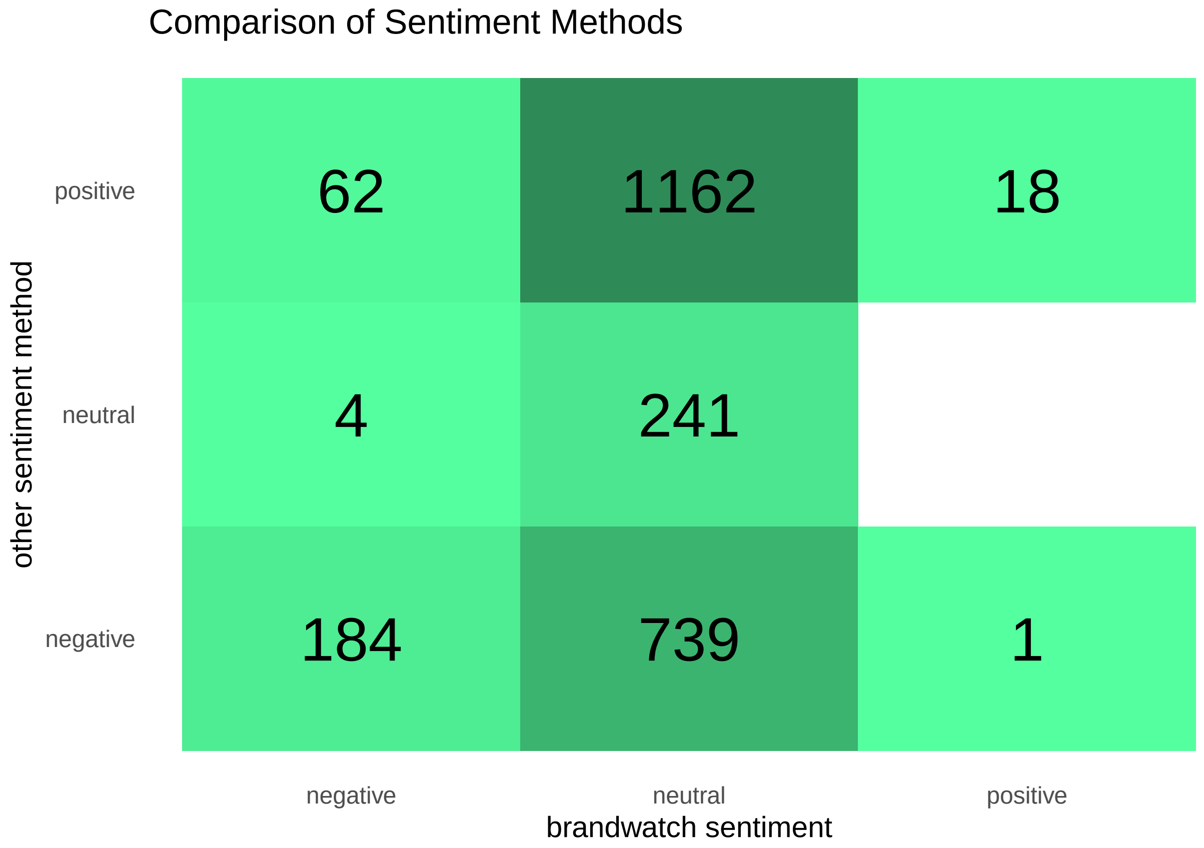
```
## ‘summarise()‘ has grouped output by ’sent_brandwatch’. You can override using
## the ‘.groups‘ argument.
```

```
n_tweets <- nrow(tweets2)
sent_method_comparison_table <- sent_method_comparison %>%
  count(sent_comparison) %>%
  mutate(percent = round((n / n_tweets) * 100, 1)) %>%
  kable(col.names = c("Comparison", "Count", "Percent")) %>%
  kable_paper(full_width = FALSE) %>%
  row_spec(c(0), background = "lightgray")
sent_method_comparison_table
```

```
both_neg <- sent_method_comparison_counts$n[sent_method_comparison_counts$sent_comparison == "both nega
both_pos <- sent_method_comparison_counts$n[sent_method_comparison_counts$sent_comparison == "both posi
both_neutral <- sent_method_comparison_counts$n[sent_method_comparison_counts$sent_comparison == "both n
bw_neu_meth2_pos <- sent_method_comparison_counts$n[sent_method_comparison_counts$sent_comparison == "b
bw_neu_meth2_neg <- sent_method_comparison_counts$n[sent_method_comparison_counts$sent_comparison == "b
opposite <- (sent_method_comparison_counts$n[sent_method_comparison_counts$sent_comparison == "brandwat
```

There were 184 tweets where both methods assigned a negative sentiment, 18 tweets where both methods assigned a positive sentiment, and 241 tweets where both methods assigned a neutral sentiment. The greatest disagreements were when brandwatch assigned a neutral sentiment but the other method assigned a positive sentiment (1162 tweets) or a negative sentiment (739 tweets). There were 63 tweets where the two methods assigned completely opposite sentiments.

```
ggplot(data = sent_method_comparison_counts2, aes(x = sent_brandwatch, y = sent_method2)) +
  geom_tile(aes(fill = count), show.legend = FALSE) +
  geom_text(aes(label = count), color = "black", size = 8) +
  theme_minimal() +
  theme(panel.grid.major = element_blank()) +
  scale_fill_gradientn(colors = c("seagreen1", "seagreen4")) +
  labs(title = "Comparison of Sentiment Methods",
       x = "brandwatch sentiment",
       y = "other sentiment method")
```


Comparison of Sentiment Methods
other sentiment method
positive
neutral
negative
negative
neutral
positive
brandwatch sentiment