Topic 3: Sentiment Analysis I

Overview

Sentiment analysis is a tool for assessing the mood of a piece of text. For example, we can use sentiment analysis to understand public perceptions of topics in environmental policy like energy, climate, and conservation.

```
library(tidyr) #text analysis in R
library(lubridate) #working with date data
library(pdftools) #read in pdfs
library(tidyverse)
library(tidytext)
library(here)
library(LexisNexisTools) #Nexis Uni data wrangling
library(sentimentr)
library(readr)
```

Introduction to the Nexis Uni data source

```
setwd(here("dat"))
#to follow along with this example, download this .docx to your working directory:
#https://github.com/MaRo406/EDS_231-text-sentiment/blob/main/nexis_dat/Nexis_IPCC_Results.docx
#my_files <- list.files(pattern = ".docx", path = getwd(),
#                       full.names = TRUE, recursive = TRUE, ignore.case = TRUE)

dat <- lnt_read(here("dat","Nexis_IPCC_Results.docx")) #Object of class 'LNT output'


meta_df <- dat@meta
articles_df <- dat@articles
paragraphs_df <- dat@paragraphs

dat2<- data_frame(element_id = seq(1:length(meta_df$Headline)),
                  Date = meta_df$Date,
                  Headline = meta_df$Headline)
```

```
#can we create a similar graph to Figure 3A from Froelich et al.?
mytext <- get_sentences(dat2$Headline)
sent <- sentiment(mytext)

sent_df <- inner_join(dat2, sent, by = "element_id")

sentiment <- sentiment_by(sent_df$Headline)
```

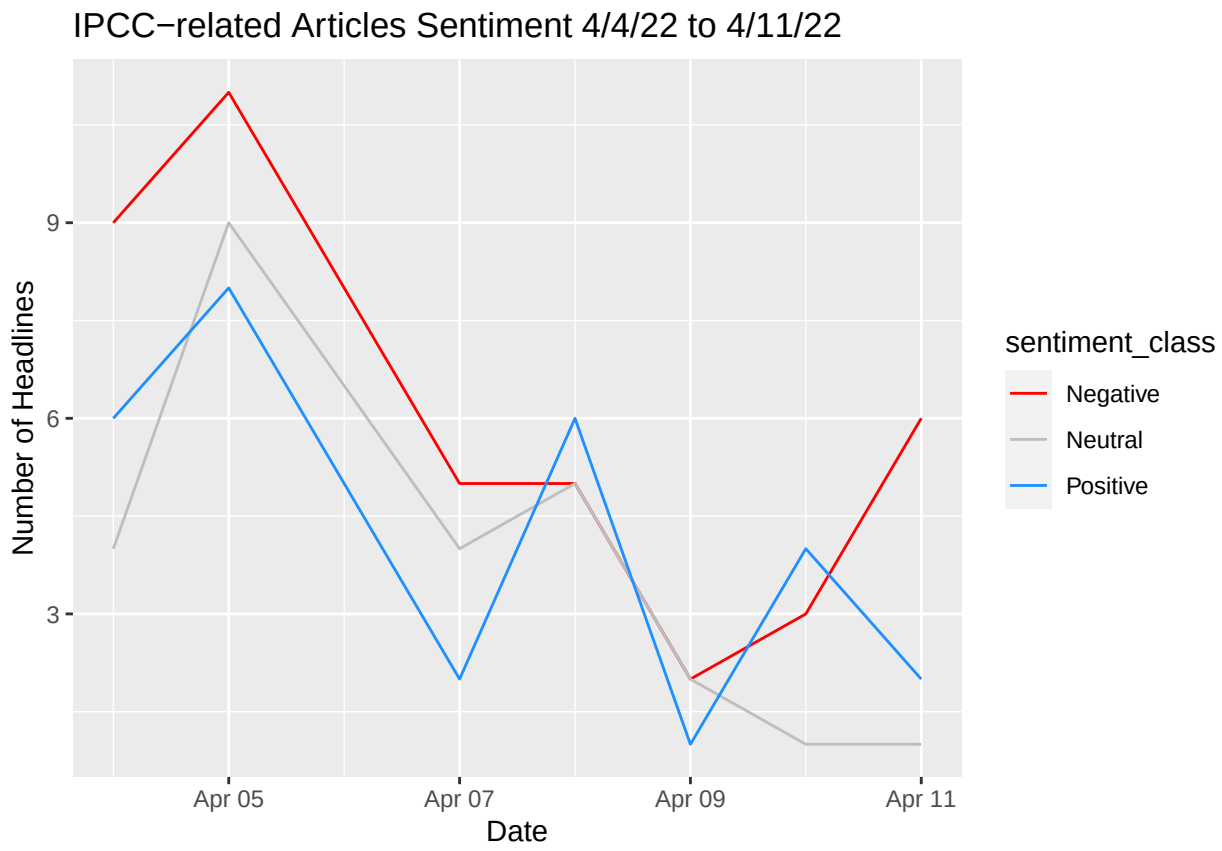
```
sent_df %>%
  arrange(sentiment)
```

```
## # A tibble: 109 x 6
##    element_id Date       Headline                sentence_id word_count sentiment
##         <int> <date>     <chr>                         <int>      <int>     <dbl>
##  1         66 2022-04-04 Scientists risk arres~           1          7    -0.756
##  2         91 2022-04-07 The 'climate change' ~           1          9    -0.75
##  3         28 2022-04-09 The Dread 1.5 Degree ~           1          6    -0.714
##  4         43 2022-04-06 India's banks unprepa~           1          7    -0.510
##  5         34 2022-04-08 Dangerous radicals ar~           1          6    -0.449
##  6         14 2022-04-04 'Now or never' to avo~           1          8    -0.442
##  7         78 2022-04-07 Statewide Gas Ban Bil~           1         10    -0.427
##  8         50 2022-04-04 Guardian: Media 'Bare~           1          8    -0.407
##  9         62 2022-04-06 Governor Youngkin's I~           1         11    -0.377
## 10          7 2022-04-05 Narrow path to avoid ~           1          8    -0.354
## # ... with 99 more rows
```

```
sent_df_aggregate <- sent_df %>%
  mutate(sentiment_class = case_when(sent_df$sentiment > 0 ~ 'Positive'
                                     ,sent_df$sentiment == 0 ~ 'Neutral'
                                     ,sent_df$sentiment < 0 ~ 'Negative')) %>%
  group_by(Date, sentiment_class) %>%
  summarize(count_sentiment=n())
```

0 Recreate Figure 1A from Froelich et al.

```
ggplot(sent_df_aggregate,
       aes(x=Date, y=count_sentiment, group=sentiment_class, color=sentiment_class)) +
  geom_line()+
  ylab("Number of Headlines")+
  scale_color_manual(values=c("red", "gray", "dodgerblue")) +
  labs(title = "IPCC-related Articles Sentiment 4/4/22 to 4/11/22")
```




```
custom_stop_words <- bind_rows(tibble(word = c("your_word"),
                                      lexicon = c("custom")),
                               stop_words)
```

1-3 Query on Nexis Uni

"Cloud seeding" was the search term used in on Nexis Uni database to grab the first 1000 articles.

4 Read in Nexis Uni data

```
my_files <- list.files(pattern = ".docx", path = here("dat","cloud_seeding"),
                       full.names = TRUE, recursive = TRUE, ignore.case = TRUE)

cloudseed_dat <- lnt_read(my_files) #Object of class 'LNT output'


cloudseed_meta_df <- cloudseed_dat@meta
cloudseed_articles_df <- cloudseed_dat@articles
cloudseed_paragraphs_df <- cloudseed_dat@paragraphs

cloudseed_dat2<- data_frame(element_id = seq(1:length(cloudseed_meta_df$Headline)),
                            Date = cloudseed_meta_df$Date,
                            Headline = cloudseed_meta_df$Headline)
```

5 Clean Nexis Uni data

```
#May be of use for assignment: using the full text from the articles
cloudseed_paragraphs_dat <- data_frame(element_id = cloudseed_paragraphs_df$Art_ID,
                                       Text  = cloudseed_paragraphs_df$Paragraph)
#
cloudseed_dat3 <- inner_join(cloudseed_dat2,cloudseed_paragraphs_dat, by = "element_id") %>%
                  janitor::clean_names()

cloudseed_dat3 <- subset(cloudseed_dat3, text != " " )
cloudseed_dat3 <- cloudseed_dat3[!grepl("POSTED", cloudseed_dat3$text,ignore.case = TRUE),]
cloudseed_dat3 <- cloudseed_dat3[!grepl("GRAPHIC", cloudseed_dat3$text,ignore.case = TRUE),]
cloudseed_dat3 <- cloudseed_dat3[!grepl(":", cloudseed_dat3$text),]
cloudseed_dat3 <- cloudseed_dat3[!grepl("LINK TO", cloudseed_dat3$text,ignore.case = TRUE),]
cloudseed_dat3 <- cloudseed_dat3[grepl("[a-zA-Z]", cloudseed_dat3$text),]
```

```
bing_sent <- get_sentiments('bing') #grab the bing sentiment lexicon from tidytext
#head(bing_sent, n = 20)
```

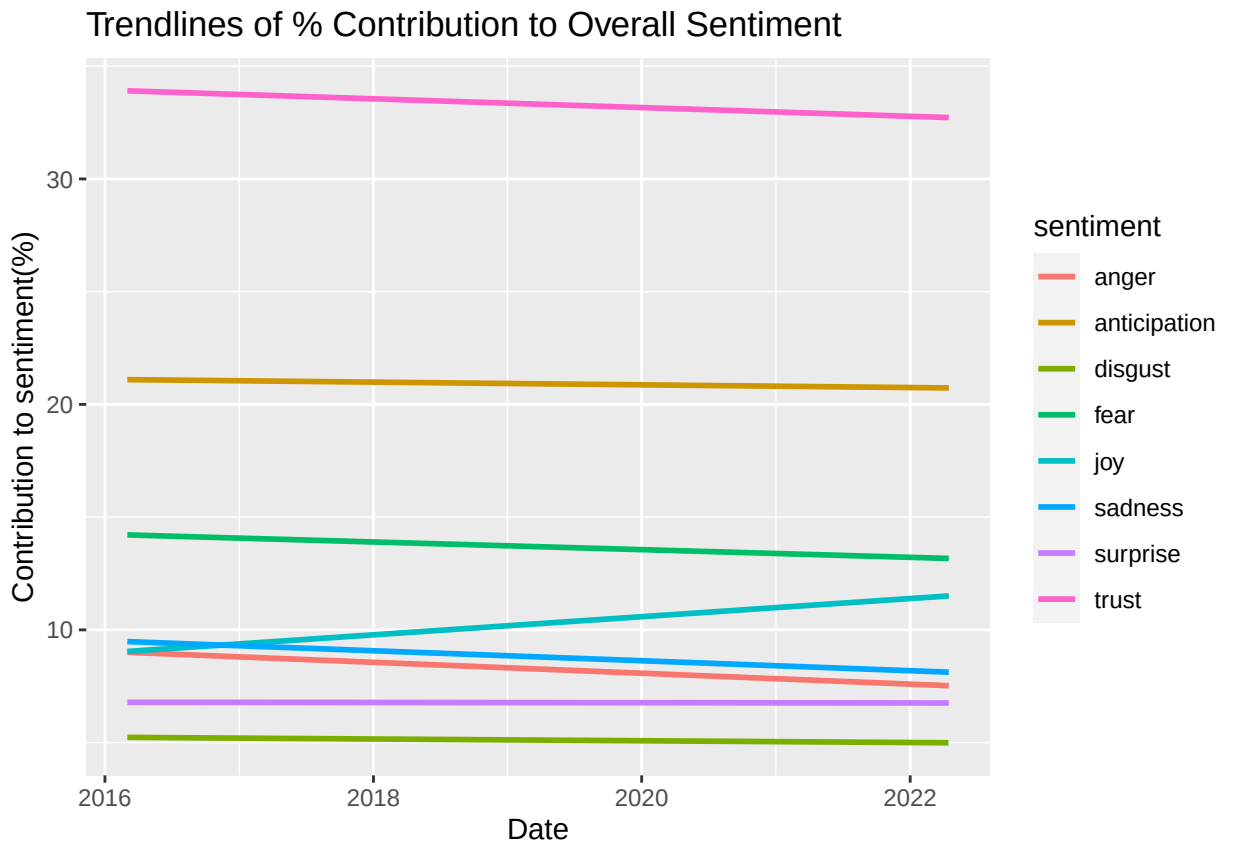
```
#unnest to word-level tokens, remove stop words, and join sentiment words
cloudseed_text_words <- cloudseed_dat3  %>%
  unnest_tokens(output = word, input = text, token = 'words')

cloudseed_sent_words <- cloudseed_text_words %>% #break text into individual words
  anti_join(stop_words, by = 'word') %>% #returns only the rows without stop words
  inner_join(bing_sent, by = 'word') #joins and retains only sentiment words
```

6 Explore dataset

```
nrc_sent <- get_sentiments('nrc') %>%
            filter(!sentiment %in% c("positive","negative")) #requires downloading a large dataset via

nrc_fear <- get_sentiments("nrc") %>%
  filter(sentiment == "fear")

#most common words by sentiment
cloudseed_fear_words <- cloudseed_dat3  %>%
  unnest_tokens(output = word, input = text, token = 'words') %>%
  inner_join(nrc_fear) %>%
  count(word, sort = TRUE)

cloudseed_word_counts <- cloudseed_text_words %>%
  inner_join(nrc_sent) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()
```

7 Trends in Sentiment over time

```
cloudseed_sent_counts <- cloudseed_text_words %>%
        inner_join(nrc_sent) %>%
        group_by(date) %>%
        count(sentiment, sort = TRUE) %>%
        mutate(sentwords_per_day = sum(n)) %>%
        mutate(pct_contribution = ((n/sentwords_per_day)*100))

cloudseed_sent_counts %>%
  group_by(date) %>%
  ggplot(aes(date, pct_contribution, group=sentiment, color=sentiment))  +
  geom_smooth( method="lm", se=F)  +
  labs(x = "Date",
       y = "Contribution to sentiment(%)",
       title = "Trendlines of % Contribution to Overall Sentiment")
```



The concept of cloud seeding has always been very controversial. In recent years, it seems that the anticipation for the potential of the technology has stagnated. There seems to be a lot of trust (the dominating sentiment through all the years) in the technology but it is slowly waning as well. The decrease in sadness and increase in joy shows potential shifts in attitudes towards cloud seeding. Perhaps, we need to go further back in time to explore more drastic changes in sentiment.