eds231_text_sentiment_analysis

Charles Hendrickson

04/19/2022

```
# Load packages
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)
library(textreadr) #Read in .docx content
library(janitor)
library(textdata)
```

Part 1

Using the "IPCC" Nexis Uni data set from the class presentation and the pseudo code we discussed, recreate Figure 1A from Froelich et al. (Date x # of 1) positive, 2) negative, 3) neutral headlines):

Introduction to the Nexis Uni data source

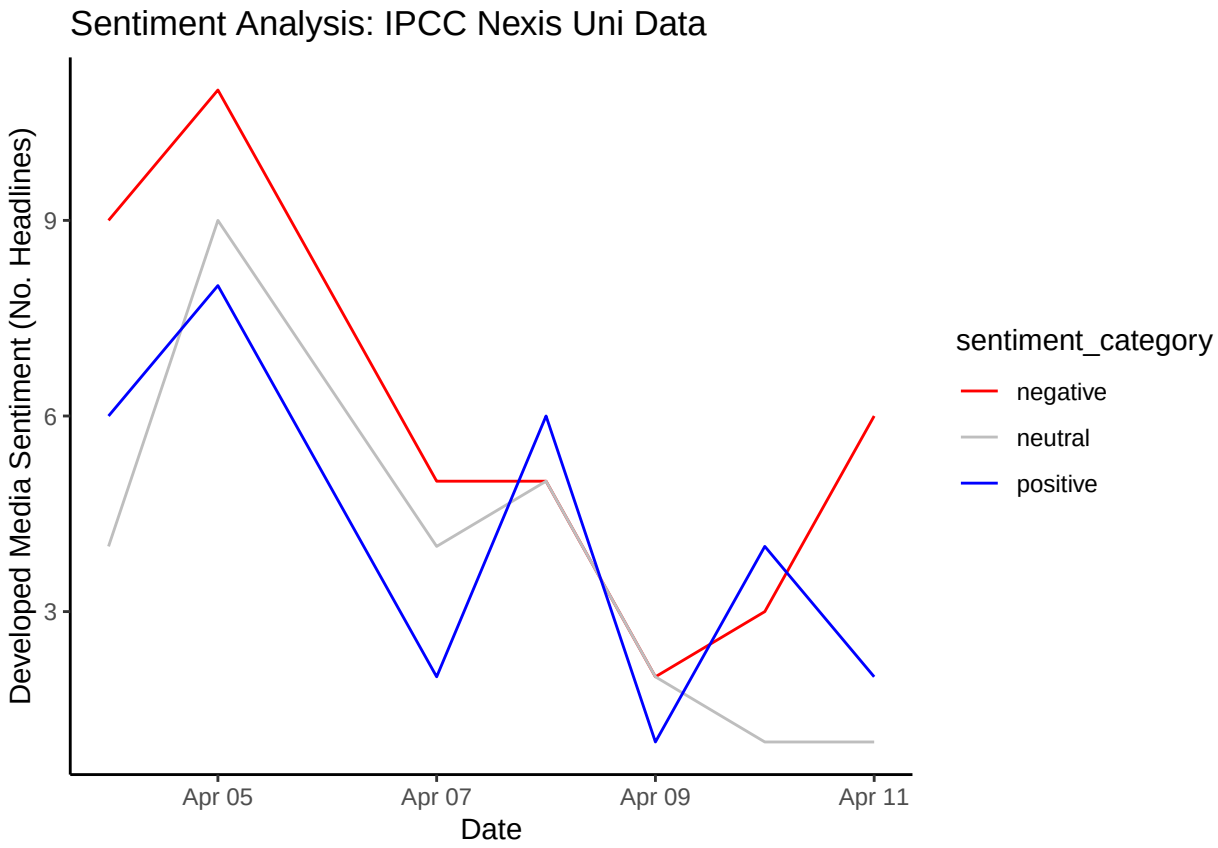
```
#to follow along with this example, download this .docx to your working directory:
#https://github.com/MaRo406/EDS_231-text-sentiment/blob/main/nexis_dat/Nexis_IPCC_Results.docx

# Specify URL where file is stored
#IPCC_url <- "https://github.com/MaRo406/EDS_231-text-sentiment/blob/main/nexis_dat/Nexis_IPCC_Results.

# Specify destination where file should be saved
#destfile <- "data/IPCC.docx"

# Download file to destination
#download.file(IPCC_url, destfile)


#my_files <- list.files(pattern = ".docx", path = getwd(),
                       #full.names = TRUE, recursive = TRUE, ignore.case = TRUE)

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


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

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

#May be of use for assignment: using the full text from the articles
# paragraphs_dat <- data_frame(element_id = paragraphs_df$Art_ID, Text  = paragraphs_df$Paragraph)
#
# dat3 <- inner_join(dat2,paragraphs_dat, by = "element_id")

#can we create a similar graph to Figure 1A 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
```

```r
# Make new columns for 'positive', 'negative', and 'neutral' sentiments

sent_new <- sent_df %>%
  mutate(sentiment_category = case_when(sentiment > 0 ~ "positive",
                                        sentiment < 0 ~ "negative",
                                        sentiment == 0 ~ "neutral")) %>%
           group_by(Date, sentiment_category) %>%
           summarise(headline_count = n()) #number of headlines per date and sentiment category
```

```r
# Plot sentiments
ggplot(data = sent_new, aes(x = Date, y = headline_count, color = sentiment_category)) +
  geom_line() +
  scale_color_manual(values = c("red", "grey", "blue")) +
  labs(title = "Sentiment Analysis: IPCC Nexis Uni Data",
       x = "Date",
       y = "Developed Media Sentiment (No. Headlines)") +
  theme_classic() +
  theme(legend.background = element_blank())
```



Part 2

Access the Nexis Uni database through the UCSB library: https://www.library.ucsb.edu/research/db/211

Choose a key search term or terms to define a set of articles.

Use your search term along with appropriate filters to obtain and download a batch of at least 100 full text search results (.docx).

Read your Nexis article document into RStudio.

This time use the full text of the articles for the analysis. First clean any artifacts of the data collection process (hint: this type of thing should be removed: "Apr 04, 2022(Biofuels Digest: http://www.biofuelsdigest.com/ Delivered by Newstex"))

Explore your data a bit and try to replicate some of the analyses above presented in class if you'd like (not necessary).

Plot the amount of emotion words (the 8 from nrc) as a percentage of all the emotion words used each day (aggregate text from articles published on the same day). How does the distribution of emotion words change over time? Can you think of any reason this would be the case?

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

#read in data
dat <- lnt_read(my_files) #Object of class 'LNT output'

#view meta data, articles, and paragraphs
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$

#May be of use for assignment: using the full text from the articles
paragraphs_dat <- data_frame(element_id = paragraphs_df$Art_ID, Text  = paragraphs_df$Paragraph)

dat3 <- inner_join(dat2,paragraphs_dat, by = "element_id") %>% clean_names() #Clean the variable names
```

First clean any artifacts of the data collection process (hint: this type of thing should be removed: "Apr 04, 2022(Biofuels Digest: http://www.biofuelsdigest.com/ Delivered by Newstex"))

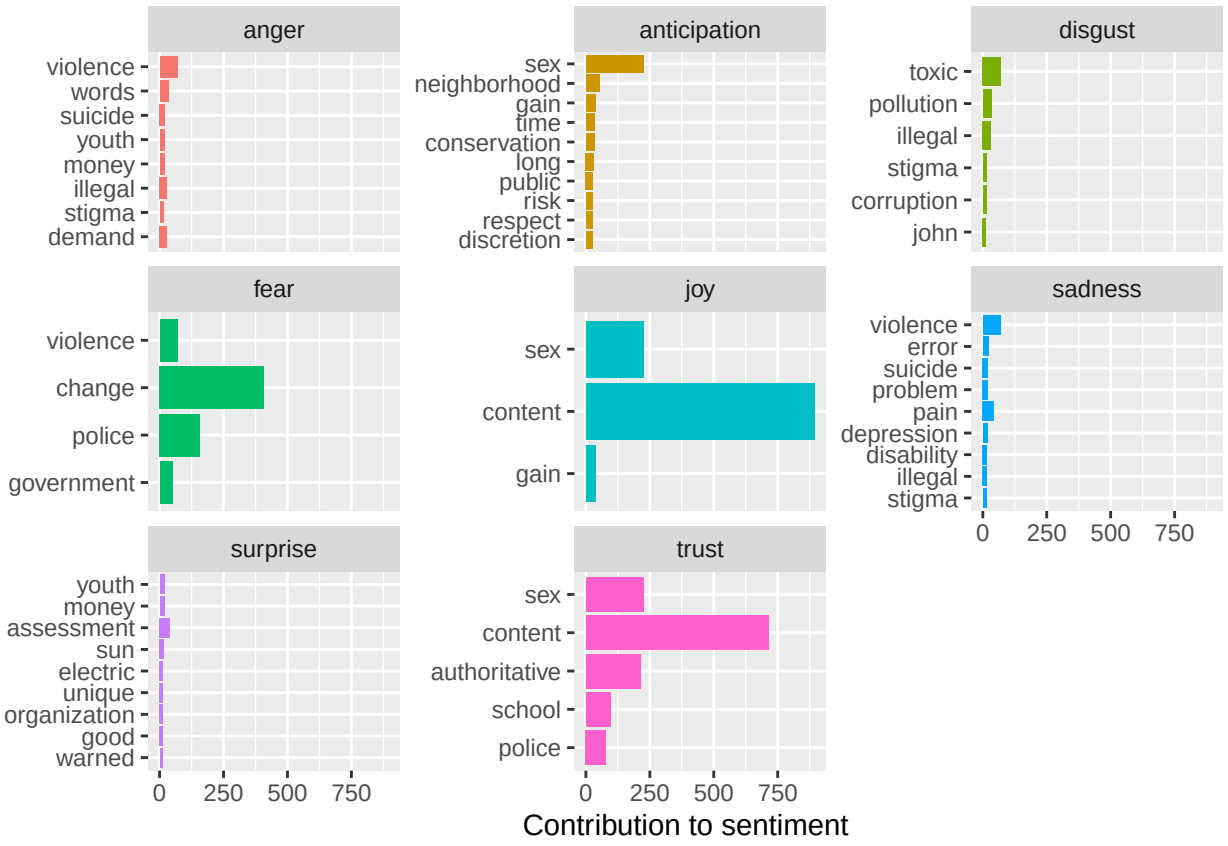
```
#Remove lines that contain 'http' from text rows
dat3_clean <- dat3[!grepl("http", dat3$text),] #grepl finds a pattern in string or string vector

#Remove lines that contain 'n#' from text rows
patterns1 <- c("n6", "n7", "n8", "n9", "n10")

dat3_clean <- dat3_clean[!grepl(paste(patterns1, collapse = "|"), dat3_clean$text),] #grepl finds a pat
```

Plot the amount of emotion words (the 8 from nrc) as a percentage of all the emotion words used each day (aggregate text from articles published on the same day).

```
nrc_sent <- get_sentiments('nrc') #requires downloading a large dataset via prompt
```

```
nrc_word_counts <- dat3_clean %>%
  unnest_tokens(output = word, input = text, token = 'words') %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, date, sort = TRUE) %>%
  ungroup()
```

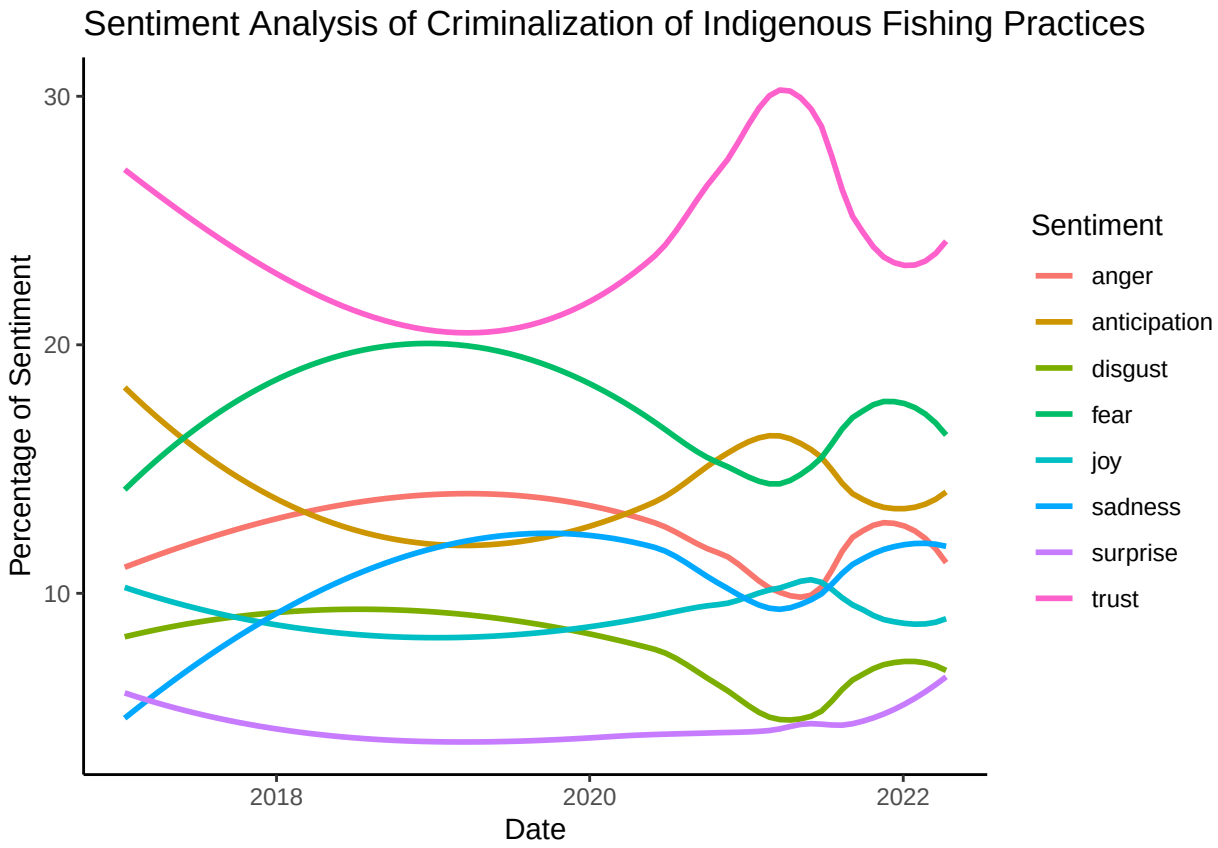
```
nrc_word_counts %>%
  group_by(sentiment) %>%
  filter(sentiment != "positive" & sentiment != "negative") %>%
  slice_max(n, n = 10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(x = "Contribution to sentiment",
       y = NULL)
```




```
nrc_word_counts_new <- nrc_word_counts %>%
  filter(sentiment != "positive" & sentiment != "negative") %>%
  group_by(date, sentiment)%>%
  count(sentiment) %>%
  ungroup() %>%
  group_by(date) %>%
  mutate(total_n = sum(n), percentage = 100*(n/total_n))
```

Percentage of Sentiment Words

```
ggplot(data = nrc_word_counts_new, aes(x = date, y = percentage, color = sentiment)) +
  geom_smooth(se = FALSE) +
  theme_classic() +
  labs(x = "Date",
       y = "Percentage of Sentiment",
       color = "Sentiment",
       title = "Sentiment Analysis of Criminalization of Indigenous Fishing Practices")
```



I was expecting anger, sadness, fear, and disgust to continually and steadily increase in the percentage of sentiment because the criminalization of indigenous fishing practices is much a sensitive and controversial topic. However, these sentiments peaked around 2019 and then steadily decreased around 2020 to a local minimum in mid 2021. They then increased again in 2022.

Interestingly, the sentiments of trust, anticipation, and joy had the exact opposite percentages. Instead, they declined around 2019 and then steadily increased around 2020 to a local maximum in mid 2021. They then decreased again in 2022. 'Surprise' peaked around 2016, slowly decreased and then increased again in 2022.

A possible explanation for why generally negative sentiments peaked in 2019 and then decreased to a minimum in 2021 could be a change in fishing management or law enforcement policies in 2021, which possibly led to the decriminalized indigenous fishing practices. This makes sense because negative sentiments about this subject in the media would likely peak before enough support could be built to change policies in 2021, at which point negative sentiments would decrease to a minimum and positive sentiments about the policy change would increase.