TAD-Replication-01-EDA

For this replication exercise, we chose the paper “Measuring Emotion in Parliamentary Debates with Automated Textual Analysis” by Rheault et al.(2016).

All codes are published in: <https://github.com/lrheault/emotion/tree/master>.

Before the main replication exericse, we first want to inspect a sample of the dataset (which spans an entire century) and see if we can choose an optimal subset of data for replication.

The key files we’ll be using for this step are:

1. **lexicon-polarity.csv** - Domain-specific lexicon used by the authors which we’ll later compare with other sentiment lexicons
2. **lexicon-generator.R** - The scripts used by the authors to generate their domain-specific lexicon
3. **emotion-main-models.R** - Contains their main analysis code

## 1. Inspect the dataset

# Load required libraries  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(lubridate)  
library(stringr)  
library(scales)

##   
## Attaching package: 'scales'  
##   
## The following object is masked from 'package:purrr':  
##   
## discard  
##   
## The following object is masked from 'package:readr':  
##   
## col\_factor

library(tidytext)  
  
# Read first 5000 rows  
df <- read.csv("uk\_hansard.csv", sep="\t", nrows=5000)  
  
print(head(df))

## speech\_id  
## 1 uk.org.publicwhip/debate/1919-02-04a.3.1  
## 2 uk.org.publicwhip/debate/1919-02-04a.5.0  
## 3 uk.org.publicwhip/debate/1919-02-04a.8.0  
## 4 uk.org.publicwhip/debate/1919-02-04a.11.0  
## 5 uk.org.publicwhip/debate/1919-02-04a.11.3  
## 6 uk.org.publicwhip/debate/1919-02-04a.12.0  
## name  
## 1 Colonel F. B. MILDMAY (addressing himself to the Clerk, who, standing up, pointed to him, and then sat down)  
## 2 Sir HENRY DALZIEL  
## 3 Mr. LOWTHER (who was received with general cheers)  
## 4 Mr. SPEAKER-ELECT (standing on the upper step to the Chair)  
## 5 Mr. BONAR LAW (Leader of the House)  
## 6 Sir DONALD MACLEAN  
## person\_id  
## 1 uk.org.publicwhip/person/15282  
## 2 unknown  
## 3 unknown  
## 4 unknown  
## 5 uk.org.publicwhip/person/20693  
## 6 uk.org.publicwhip/person/21115  
## text  
## 1 Sir Courtenay Ilbert, in accordance with the Gracious message just received from the Throne, I beg to move, "That the Right Honourable James William Lowther do take the Chair of this House as Speaker."  
## 2 Sir Courtenay Ilbert, I am fully conscious that it is to my Parliamentary seniority, and to that alone, I am permitted to exercise the high privilege of seconding the Motion so ably and so gracefully proposed by my right hon. Friend the Member for the Totnes Division (Colonel Mildmay), to whom, on his safe return to this House after his arduous and perilous military duty, we offer a warm and a cordial welcome. If I may be pardoned for a further personal reference, I would say that it is not altogether infelicitous or inappropriate that the great-grandson of the great Earl Grey of Reform renown should, apart from his personal claim, be the first to address the Members of a new and reformed Parliament.  
## 3 Sir Courtenay Ilbert,—In accordance with the ancient custom of this House, I rise at this moment to submit myself to the will of the House, and in doing so I should like, first of all, to tender to the two right hon. Gentlemen who have proposed and seconded my nomination to the Chair my very heartiest thanks for the extraordinary kind and eulogistic terms in which they have submitted my name. The right hon. Gentleman the Member for the Totnes Division (Colonel Mildmay) is an old college friend of mine. His presence in this Chamber reminds me from time to time of how "the noiseless step of time steals swiftly by," for it is about forty years ago since we were at Cambridge together. My right hon. Friend has won his laurels in many fields. He is one of those gallant men who left the House of Commons to do service in France and Flanders. There he has won distinction, for which the House will ever be grateful to him. His modesty in debate is the only fault which I have to find, and it is surpassed by the gallantry which he has shown both in South Africa and in the Great War. My right hon. Friend behind me (Sir H. Dalziel), who was kind enough to second my nomination, has won distinction in other fields. He has preferred the Senate to the Field. He has won distinction in our Debates, in which he has taken a frequent and most important part. The weight of his observations, the attractiveness of his speech, always secure for him a full House. He is one of those who believe that "the pen is mightier than the sword." Yet he has always carried a lance, and his lance has been the freest of free lances, couched at one time in this direction and at another time in that. Both right hon. Gentlemen have referred to my long service in this House, both at the Table and at the Chair. I would like at once—I am afraid that on such an occasion as this I have to deal with personal matters—to say that if the House be willing and prepared to call me to be the occupant of the Chair, I shall be glad again to place my services at its disposal; and in order that all doubt on the point may be set at rest, I should like to add that I am prepared,\nif no unforeseen circumstance should arise, to remain there for a year or eighteen months, or possibly two years.  
## 4 Before taking the Chair, once more as Speaker of this House, I should like to express to it my deep sense of the high honour which the House has done me for the fifth time.  
## 5 Mr. Speaker-Elect,—In the unavoidable absence of the Prime Minister, it is my privilege to congratulate you on the honour—the highest in their power to bestow—which has again been conferred upon you by your fellow-Members in selecting you, to use the old-time phraseology of the Royal Message which we have just heard, as "a proper person to preside over our deliberations." I remember well the first occasion when you were elected as Speaker in this House. The duties of your position are so responsible and so difficult—they require a combination of so many rare qualities—that, however great may have been the success of a Speaker-Elect in other fields, there is always some, if not anxieties, some uncertainties, as to what his success will be in his new sphere. Before you were chosen for this office you had long filled the post of Chairman of Committees. In that position you had so won the confidence of the House of Commons that in your case there was little, if any, of the uncertainty to which I refer. I express, I think, the opinion of every Member of the House who has served under your Speakership when I say that the high expectations which were then entertained as to your success have been more than justified by the event. As you have pointed out, the position of Speaker would be impossible unless he could rely always on the unswerving support of the House of Commons; but the wholeheartedness and enthusiasm with which that support will be given must depend always upon the personality of the occupant of the Chair, and must depend on the extent to which he succeeds in securing and maintaining not only the respect, but the good will of the Members of the House of Commons. That respect\nand good will you secured at the beginning, and every succeeding Session has made that feeling more marked on the part of Members of the House.  
## 6 I should like to say, Sir, before I endeavour to express in such way as I can the congratulations of those with whom I am associated, that I rise to catch your eye, Mr. Speaker-Elect, without my right hon. Friend the Leader of the Labour party (Mr. Adamson) also rising, by understanding and agreement with him. I am an older Member of the House, and senior to him in the great office, of Privy Councillor of His Majesty the King. Let me say this, that imperfect as my qualifications must\nbe for following the Leader of the House, I have one qualification which even he does not possess, and that is the fact that for seven years I had the very great privilege of being associated with you in the conduct of the business of this House, in a minor capacity, in both the Chair at the Table and the Chair which you now so worthily occupy.

#str(df)  
  
# the first 5000 rows seem to only contain speeches  
# up till the date 1919-03-05  
# print(df[2500,])  
# print(df[4999,])

#### Observations:

The data contains these features:

* **speech\_id**: Contains date information (e.g., 1919-02-04) and unique identifiers
* **name**: Speaker’s name with formal titles
* **person\_id**: Unique identifier for speakers (some marked as “unknown”)
* **text**: The actual speech content

### 

### Exploratory Data Analysis

df <- df %>%  
 mutate(  
 # Extract date from speech\_id  
 date = as.Date(str\_extract(speech\_id, "\\d{4}-\\d{2}-\\d{2}")),  
 # Calculate text statistics  
 word\_count = str\_count(text, "\\w+"),  
 char\_count = str\_length(text),  
 avg\_word\_length = char\_count / word\_count,  
 known\_speaker = person\_id != "unknown"  
 )  
  
# 1. Time-based Analysis (within 1919)  
daily\_counts <- df %>%  
 count(date) %>%  
 arrange(date)  
  
p1 <- ggplot(daily\_counts, aes(x=date, y=n)) +  
 geom\_line() +  
 theme\_minimal() +  
 labs(title="Number of Speeches per Day in Early 1919",  
 x="Date", y="Number of Speeches") +  
 scale\_x\_date(date\_breaks="1 week", date\_labels="%b %d")  
  
# 2. Speaker Activity Analysis  
speaker\_stats <- df %>%  
 group\_by(name) %>%  
 summarise(  
 speeches = n(),  
 avg\_length = mean(word\_count),  
 total\_words = sum(word\_count),  
 dates\_active = n\_distinct(date)  
 ) %>%  
 arrange(desc(speeches))  
  
# 3. Speech Length Analysis by Speaker Type  
length\_by\_known <- df %>%  
 group\_by(known\_speaker) %>%  
 summarise(  
 avg\_length = mean(word\_count),  
 med\_length = median(word\_count),  
 n = n()  
 )  
  
# 4. Text Analysis  
  
# Remove common parliamentary terms and standard stop words  
custom\_stop\_words <- c("hon", "right", "house", "gentleman", "member", "minister")  
word\_frequencies <- df %>%  
 unnest\_tokens(word, text) %>%  
 anti\_join(stop\_words) %>%  
 filter(!word %in% custom\_stop\_words) %>%  
 count(word, sort = TRUE)

## Joining with `by = join\_by(word)`

# 5. Daily Patterns  
df <- df %>%  
 mutate(  
 day\_of\_week = wday(date, label=TRUE),  
 week = week(date)  
 )  
  
# Print summary statistics  
cat("=== Dataset Overview (Early 1919) ===\n")

## === Dataset Overview (Early 1919) ===

cat("Date range:", min(df$date), "to", max(df$date), "\n")

## Date range: -18594 to -18565

cat("Number of speeches:", nrow(df), "\n")

## Number of speeches: 5000

cat("Number of unique speakers:", length(unique(df$name)), "\n")

## Number of unique speakers: 651

cat("Proportion of known speakers:", mean(df$known\_speaker), "\n")

## Proportion of known speakers: 0.8388

cat("\n=== Top Speakers by Number of Speeches ===\n")

##   
## === Top Speakers by Number of Speeches ===

print(head(speaker\_stats, 10))

## # A tibble: 10 × 5  
## name speeches avg\_length total\_words dates\_active  
## <chr> <int> <dbl> <int> <int>  
## 1 Mr. BONAR LAW 336 62.9 21133 14  
## 2 Mr. CHURCHILL 206 60.0 12367 9  
## 3 Captain GUEST 139 43.7 6070 8  
## 4 Colonel WEDGWOOD 121 63.5 7684 12  
## 5 Mr. DEVLIN 116 48.7 5645 11  
## 6 Mr. BALDWIN 110 59.4 6529 12  
## 7 Mr. SPEAKER 96 34.6 3321 17  
## 8 Sir J. CRAIG 84 60.3 5066 10  
## 9 Mr. HOGGE 83 107. 8865 14  
## 10 Mr. SHORTT 83 64.8 5379 8

cat("\n=== Speech Length Statistics ===\n")

##   
## === Speech Length Statistics ===

cat("Overall:\n")

## Overall:

print(summary(df$word\_count))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 18.00 37.00 78.38 76.00 1247.00

cat("\nBy Speaker Type:\n")

##   
## By Speaker Type:

print(length\_by\_known)

## # A tibble: 2 × 4  
## known\_speaker avg\_length med\_length n  
## <lgl> <dbl> <dbl> <int>  
## 1 FALSE 61.6 36 806  
## 2 TRUE 81.6 37.5 4194

cat("\n=== Most Common Substantive Words ===\n")

##   
## === Most Common Substantive Words ===

print(head(word\_frequencies, 20))

## word n  
## 1 government 1505  
## 2 question 1231  
## 3 friend 1039  
## 4 war 951  
## 5 committee 902  
## 6 time 772  
## 7 bill 694  
## 8 country 640  
## 9 matter 573  
## 10 secretary 556  
## 11 amendment 543  
## 12 army 442  
## 13 view 429  
## 14 aware 427  
## 15 officers 424  
## 16 day 406  
## 17 regard 403  
## 18 service 403  
## 19 board 381  
## 20 public 372

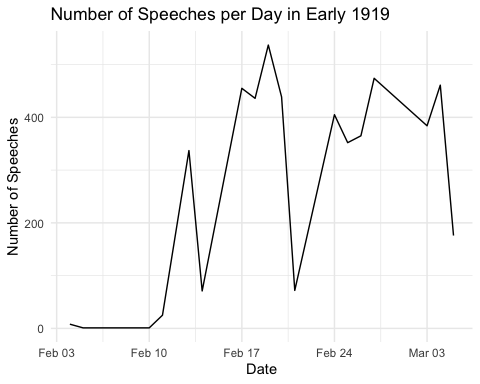
cat("\n=== Daily Activity Patterns ===\n")

##   
## === Daily Activity Patterns ===

print(table(df$day\_of\_week))

##   
## Sun Mon Tue Wed Thu Fri Sat   
## 0 1245 1282 1079 1251 143 0

# Visualization for Daily speech counts  
p1



#### Observations:

1. **Weekly Structure** in early 1919 shows that the busiest days were Monday (1245 speeches), Tuesday (1282 speeches), Wednesday (1079 speeches), and Thursday (1251 speeches). There was minimal activity on Fridays (143 speeches) and no recorded speeches on weekends, in alignment with the traditional parliamentary workweek.
2. **Speaker Analysis** reveals the most active parliamentarians: Mr. BONAR LAW, who was Leader of the House, was by far the most prolific speaker with 336 speeches and over 21,000 total words. His average speech length of 62.9 words suggests a style of concise responses. Mr. CHURCHILL follows with 206 speeches, maintaining a similar average length of 60 words. Interestingly, Mr. HOGGE stands out with the highest average speech length (106.8 words) despite having fewer speeches (83), indicating a more detailed speaking style.
3. **Speech Length Distribution** shows considerable variation. While the **median speech length is 37 words, the mean is 78.38 words**, indicating a right-skewed distribution with some very long speeches pulling the average up. The maximum speech length of 1,247 words represents extensive parliamentary addresses, while the minimum of 1 word likely represents procedural responses or interjections.
4. **Speaker Status Analysis** reveals an important distinction between known and unknown speakers. Known speakers (83.88% of speeches) tend to make longer contributions, with an average length of 81.61 words compared to 61.57 words for unknown speakers. However, the median lengths are more similar (37.5 vs 36.0), suggesting that **the difference is driven by some very long speeches from known speakers.**
5. **Most Common Words** reflect the political context of 1919. The prominence of terms like “government” (1505 occurrences), “war” (951), and “army” (442) indicates the **post-WWI context**. The frequent use of “friend” (1039 occurrences) likely reflects parliamentary courtesy in referring to other members. Terms like “committee”, “bill”, and “amendment” highlight the legislative process, while “question” suggests active parliamentary scrutiny. These patterns shows a parliamentary discourse **dominated by post-war governance issues**.

## 

## 2. Subseting for a smaller dataset (1973-1977)

For our replication, we’ll be targeting only the period of 1973-1977, which was identified as a critical period in the paper. It encompassed (a) the major recession of 1973-1975, (b) unprecedented spike in the Misery Index, (c) a peak in labor disputes, and (d) notable negativity in parliamentary debates.

Extending to 1977 (two years after the recession) would capture the entire period of recession and its aftermath, allowing us to see both the negative shock and potential recovery in sentiment.

### 2.1 Exploratory Data Analysis for subsetted data

# Read the data with explicit parsing  
hansard <- read.csv("uk\_hansard\_1973\_1977.csv",   
 sep = "\t", # tab delimiter  
 quote = "\"", # handle quotes properly  
 stringsAsFactors = FALSE,  
 encoding = "UTF-8")  
  
# head(hansard)  
# person\_id has lots of N/A, other fields remain the same as prior  
  
# Clean and verify speech\_id format  
hansard <- hansard %>%  
 mutate(  
 # Extract year using regex pattern matching the speech\_id format  
 year = str\_extract(speech\_id, "(?<=debate/)[0-9]{4}")  
 )  
  
# Basic overview  
cat("\n=== Dataset Overview: ===\n")

##   
## === Dataset Overview: ===

print(paste("Number of speeches:", nrow(hansard)))

## [1] "Number of speeches: 286057"

print(paste("Number of columns:", ncol(hansard)))

## [1] "Number of columns: 5"

# Year distribution  
year\_counts <- table(hansard$year)  
cat("\n=== Speeches per year: ===\n")

##   
## === Speeches per year: ===

print(year\_counts)

##   
## 1973 1974 1975 1976 1977   
## 55621 44082 58868 68718 58768

#### Observations:

1. 1976 had the highest number of speeches (68,718). 1974 had the lowest (44,082).
2. While the median speech length is 348 characters, the mean is much higher at 1,068 characters, indicating **a right-skewed distribution with some very long speeches.** The longest speech contains 45,687 characters, while some speeches have zero length, which may require investigation as potential data quality issues.

# Speaker analysis  
speaker\_counts <- hansard %>%  
 count(name, sort = TRUE) %>%  
 head(10)  
print("Top 10 most frequent speakers:")

## [1] "Top 10 most frequent speakers:"

print(speaker\_counts)

## name n  
## 1 Mr. Speaker 6999  
## 2 The Prime Minister 6220  
## 3 Mr. Deputy Speaker 4734  
## 4 Mr. Foot 3220  
## 5 Mr. Short 2400  
## 6 Mr. Dalyell 1788  
## 7 Mr. Skinner 1778  
## 8 Mr. Shore 1630  
## 9 Mr. Rees 1564  
## 10 Mr. Benn 1493

# Text length analysis  
hansard <- hansard %>%  
 mutate(text\_length = nchar(as.character(text)))  
  
print("Text length summary:")

## [1] "Text length summary:"

print(summary(hansard$text\_length))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 169 348 1068 745 45687

# Check for debate sessions per day  
daily\_counts <- hansard %>%  
 mutate(date = str\_extract(speech\_id, "[0-9]{4}-[0-9]{2}-[0-9]{2}")) %>%  
 count(date, sort = TRUE) %>%  
 head(10)  
print("Busiest debate days (top 10):")

## [1] "Busiest debate days (top 10):"

print(daily\_counts)

## date n  
## 1 1977-06-21 1045  
## 2 1976-07-08 928  
## 3 1976-07-27 859  
## 4 1977-07-28 796  
## 5 1975-11-11 757  
## 6 1976-07-12 748  
## 7 1977-02-01 739  
## 8 1977-07-13 734  
## 9 1974-11-28 731  
## 10 1975-04-23 722

# Check data quality  
print("Missing/NA values per column:")

## [1] "Missing/NA values per column:"

sapply(hansard, function(x) sum(is.na(x) | x == "N/A" | x == ""))

## speech\_id name person\_id text year text\_length   
## 0 0 286057 5 0 0

# Sample of unique speech patterns  
print("Sample of different debate patterns (unique speech\_id patterns):")

## [1] "Sample of different debate patterns (unique speech\_id patterns):"

unique\_patterns <- head(unique(str\_extract(hansard$speech\_id, "\\d+a\\.\\d+\\.\\d+")), 10)  
print(unique\_patterns)

## [1] "22a.1.6" "22a.1.7" "22a.1.8" "22a.2.0" "22a.2.1" "22a.2.2" "22a.2.3"  
## [8] "22a.3.0" "22a.3.2" "22a.3.3"

#### Observations:

1. There appears to be a clear hierarchy in participation. The Speaker and Deputy Speaker, who manage parliamentary proceedings, appear most frequently. Among regular MPs, prominent figures like Mr. Foot, Mr. Short, and Mr. Dalyell were particularly active. The Prime Minister’s high frequency (6,220 speeches) indicates regular parliamentary engagement during this period.
2. There are some particularly active sessions, with some days recording over 1,000 speeches (e.g., June 21, 1977 with 1,045 speeches). **Many of the busiest days occur in July**, suggesting increased parliamentary activity before summer recesses.
3. While most fields are complete, all person\_id entries are marked as “N/A”, and there are 5 missing text entries. person\_id are unique identifier of the speaker, which might not be used in this analysis and can be dropped.

# Drop person\_id and clean the data  
hansard\_clean <- hansard %>%  
 select(-person\_id) %>% # Remove person\_id  
 filter(!is.na(text)) # Remove rows with missing text  
  
print(paste("Rows remaining after cleaning:", nrow(hansard\_clean)))

## [1] "Rows remaining after cleaning: 286057"