Social Science Testimony

Julia Tache

2/12/2021

While looking through the Data is Plural archive, I was interested in the dataset on social scientists testifying in congress. I always wondered about the efficacy of "expert witnesses" in court cases as well as in congressional hearings. People who have very extensive experience in a certain topic could provide an excellent resource in the consultation of legislation, especially if the bill or other initiative they are sharing their testimony for is directly related to their field of study. They may be able to persuade the members of congress to vote in a certain way if their testimony is strong enough. Because the individuals in the dataset are social scientists, one can assume that they would prioritize social benefits of whatever legislative process the congressional committee is going through. For example, an economist might make suggestions on the budget based on the financial needs of the country, or a sociologist might suggest ways to make a bill about housing more equitable and fairer to those affected by structural inequities. The presence of social scientists in congress could have a very positive impact overall.

Of course, the opposite could very well be true: what if their testimonies are not listened to and respected, even if their advice is beneficial? What if these scholars, academics, thinktank researchers, etc. have their own special interests they are trying to promote? What if some members of certain disciplines are given more credence than others? I was curious about the tangible effect that these testimonies could have on actual legislative change.

My research question is whether or not social science testimonies make an impact in congressional hearings? By impact, I decided somewhat arbitrarily on a few factors: the total presence of social scientists at hearings, the ability to get bills passed during related hearings, as well as the ability to incite change in specific committees.

```
library(tidyverse)
sst <- read_csv("social science congressional testimony.csv")
gov_df <- read_csv("gov_df.csv")
bills_df <- read_csv("bills_df.csv")

head(sst)
```

```
## # A tibble: 6 x 29
##      X1 hearing_id  year date_begin date_end witness_name witness_affilia~
##   <dbl> <chr>      <dbl> <chr>      <chr>    <chr>        <chr>
## 1     1 HRG-1946-~  1946 May 8, 19~ Jun. 3,~ howard grie~ chief economist~
## 2     2 HRG-1947-~  1947 May 26, 1~ Jun. 16~ philip haus~ dep dir, bur of~
## 3     3 HRG-1947-~  1947 Jun. 27, ~ Jul. 16~ j. ely       chief, foreign ~
## 4     4 HRG-1948-~  1948 Mar. 24, ~ Apr. 15~ j. ely       chief, foreign ~
## 5     5 HRG-1949-~  1949 Aug. 31, ~ Sep. 15~ philip haus~ act dir, bur of~
## 6     6 HRG-1949-~  1949 Oct. 12, ~ Oct. 17~ philip hous~ act dir, bur of~
## # ... with 22 more variables: discipline1 <chr>, discipline2 <chr>,
## #   discipline3 <lgl>, sociologist <dbl>, economist <dbl>,
## #   anthropologist <dbl>, psychologist <dbl>, polscientist <dbl>,
## #   agency <chr>, census <dbl>, nsfsbe <dbl>, thinktank <dbl>,
## #   thinktank1 <chr>, thinktank2 <lgl>, full_committee1 <chr>,
## #   sub_committee1 <chr>, full_committee2 <chr>, sub_committee2 <chr>,
## #   title_description <chr>, filter <dbl>, HID <dbl>, univ_aff <dbl>
```

Using the Social Science Congressional Testimony dataset, I first looked at the most popular committees that social scientists attend. For all of time, the most popular is the Joint Economic Committee, but from

2015 onwards the most popular committee in the data was the House Committee on Financial Services. I looked at 2015 onwards because this represents the most recent sessions of congress in the data: the 114th to 116th sessions. I figured that these data would be the most relevant to my research question because I would not have to take historical factors into consideration as much.

```
# most popular committee
table_com <- as.data.frame(table(sst$full_committee1))
head(table_com[order(-table_com$Freq), ])
```

```
##                                                    Var1 Freq
## 38                            committee on economic. joint 1154
## 125                        committee on ways and means. house  967
## 118                        committee on the judiciary. senate  686
## 21  committee on banking, housing, and urban affairs. senate  682
## 50                            committee on finance. senate  653
## 52                         committee on foreign affairs. house  565
```

```
sst_recent <- sst %>%
  filter(year >= 2015)
```

```
library(data.table)
table_df <- as.data.frame(table(sst_recent$full_committee1)) %>%
  rename("Frequency" = "Freq") %>%
  rename("Committee Name" = "Var1")

top5_comm <- table_df[order(-table_df$Frequency), ][0:5, ]

table_df$`Committee Name` <- str_to_title(table_df$`Committee Name`)

head(data.table(table_df[order(-table_df$Frequency), ]))
```

```
##                                          Committee Name Frequency
## 1:                  Committee On Financial Services. House        39
## 2:                     Committee On Foreign Affairs. House        24
## 3:                      Committee On Ways And Means. House        24
## 4: Committee On Banking, Housing, And Urban Affairs. Senate        21
## 5:                            Committee On Budget. House        17
## 6:                           Committee On Budget. Senate        17
```

I created a similar frequency table for the witnesses who spoke and the most represented disciplines in the field.

```
table_df <- as.data.frame(table(sst_recent$witness_name)) %>%
  rename("Frequency" = "Freq") %>%
  rename("Witness" = "Var1")

table_df$Witness <- str_to_title(table_df$Witness)

data.table(table_df[order(-table_df$Frequency), ])
```

```
##                  Witness Frequency
```

```
##   1: Douglas Holtz-Eakin        19
##   2:        Janet Yellen        13
##   3:         John Taylor         9
##   4:          Keith Hall         9
##   5:       Simon Johnson         8
##  ---
## 186:           W. Wilcox         1
## 187:        Walden Bello         1
## 188:       Wallace Tyner         1
## 189:     Walter Kemmsies         1
## 190:    William Lawrence         1
```

```
table_df <- as.data.frame(table(sst_recent$discipline1)) %>%
  rename("Frequency" = "Freq") %>%
  rename("Discipline" = "Var1")

table_df$Discipline <- str_to_title(table_df$Discipline)

data.table(table_df[order(-table_df$Frequency), ])
```

```
##             Discipline Frequency
## 1:           Economist       228
## 2: Political Scientist        89
## 3:         Sociologist        10
## 4:        Psychologist         4
## 5:      Anthropologist         1
```

Next, I did some NLP and sentiment analysis to understand the nature of the hearings. I used the title_description variable to create a corpus of text, a document term matrix, and another variable calculating the sentiment value of the description using the syuzhet package. This process of sentiment analysis is concerned with "plot" development as well as sentiment by utilizing several different existing dictionaries. I felt that this was suitable and comprehensive enough to calculate the sentiments of the descriptions of the hearings because it would give an idea of the positive/negative attributes of these hearings while also giving us a sense of how "well-developed" the intention of the hearing was.

Code template from https://www.red-gate.com/simple-talk/sql/bi/text-mining-and-sentiment-analysis-with-r/

```
library(tidytext)
library(tm)
library(SnowballC)
library(RColorBrewer)
library(syuzhet)

TextDoc <- Corpus(VectorSource(sst_recent$title_description))

toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
TextDoc <- tm_map(TextDoc, toSpace, "/")
TextDoc <- tm_map(TextDoc, toSpace, "@")
TextDoc <- tm_map(TextDoc, toSpace, "\\|")

# Convert the text to lower case
TextDoc <- tm_map(TextDoc, content_transformer(tolower))
```

```
# Remove numbers
TextDoc <- tm_map(TextDoc, removeNumbers)

# Remove english common stopwords
TextDoc <- tm_map(TextDoc, removeWords, stopwords("english"))

# Remove punctuations
TextDoc <- tm_map(TextDoc, removePunctuation)

# Eliminate extra white spaces
TextDoc <- tm_map(TextDoc, stripWhitespace)
```

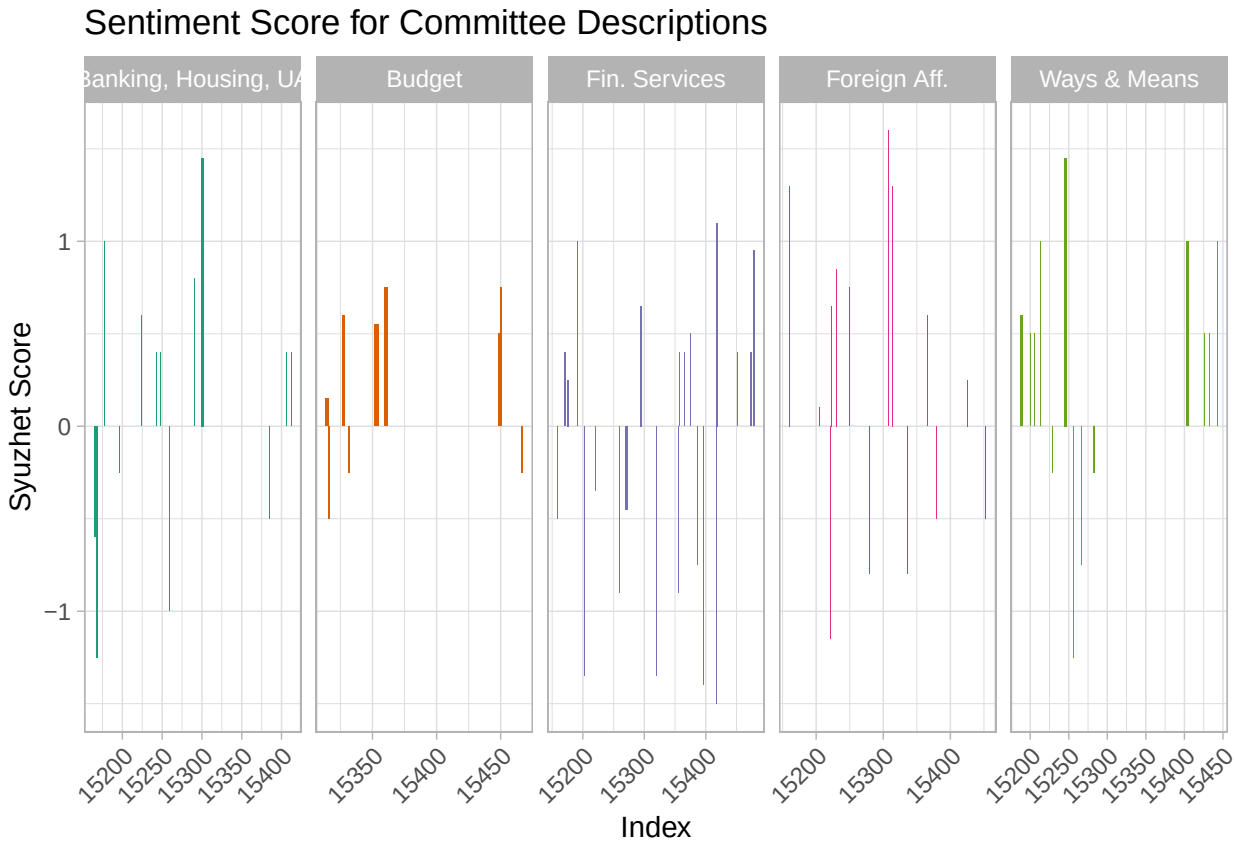
```
# Build a term-document matrix
TextDoc_dtm <- TermDocumentMatrix(TextDoc)
dtm_m <- as.matrix(TextDoc_dtm)

# Sort by descearing value of frequency
dtm_v <- sort(rowSums(dtm_m),decreasing=TRUE)
dtm_d <- data.frame(word = names(dtm_v),freq=dtm_v)
dtm_d <- dtm_d %>%
  rename("Word" = "word") %>%
  rename("Frequency" = "freq")

# Display the top 5 most frequent words
head(dtm_d, 5)
```

```
##              Word Frequency
## federal   federal        27
## budget     budget        27
## policy     policy        26
## economic economic        25
## reform     reform        24
```

Looking at the five most common committees in the data, the sentiments ranged widely from very positive to very negative across hearings. This could suggest that these hearings were of a very pressing nature and required experts to speak on the subject immediately.

```
sst_recent$syuzhet_vector <- get_sentiment(sst_recent$title_description, method="syuzhet")

top5_comm_graph <- sst_recent %>%
  filter(full_committee1 %in% top5_comm$`Committee Name`) %>%
  mutate(full_committee1 = replace(full_committee1, full_committee1 == "committee on financial services
  mutate(full_committee1 = replace(full_committee1, full_committee1 == "committee on foreign affairs. ho
  mutate(full_committee1 = replace(full_committee1, full_committee1 == "committee on banking, housing, a
  mutate(full_committee1 = replace(full_committee1, full_committee1 == "committee on ways and means. hou
  mutate(full_committee1 = replace(full_committee1, full_committee1 == "committee on budget. house", "Bu
  filter(syuzhet_vector != 0)

ggplot(top5_comm_graph, aes(X1, syuzhet_vector, fill = full_committee1)) +
  geom_col(show.legend = FALSE) +
  facet_grid(~full_committee1, scales = "free") +
  theme_light() +
  ggtitle("Sentiment Score for Committee Descriptions") +
```

```
  xlab("Index") +
  ylab("Syuzhet Score") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = "Dark2")
```



Using binary variables in the data, I compared the syuzhet vector to the outcome variable of the field of the expert present. However, none of the results were significant.

```
library(rstanarm)
fit <- glm(economist ~ syuzhet_vector, sst_recent, family = "binomial")
print(summary(fit), digits = 5)
```

```
##
## Call:
## glm(formula = economist ~ syuzhet_vector, family = "binomial",
##     data = sst_recent)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.69717  -1.46414   0.83502   0.88432   1.03417
##
## Coefficients:
##                Estimate Std. Error z value  Pr(>|z|)
## (Intercept)     0.73716    0.12328  5.9798 2.235e-09 ***
## syuzhet_vector  0.21107    0.16304  1.2946    0.1955
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 412.795  on 331  degrees of freedom
## Residual deviance: 411.104  on 330  degrees of freedom
## AIC: 415.104
##
## Number of Fisher Scoring iterations: 4
```

```
fit_2 <- glm(polscientist ~ syuzhet_vector, sst_recent, family = "binomial")
print(summary(fit_2), digits = 5)
```

```
##
## Call:
## glm(formula = polscientist ~ syuzhet_vector, family = "binomial",
##     data = sst_recent)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.99341  -0.82423  -0.77371   1.48903   1.80466
##
## Coefficients:
##                Estimate Std. Error z value  Pr(>|z|)
## (Intercept)    -0.90511    0.12713 -7.1194 1.084e-12 ***
## syuzhet_vector -0.24626    0.16934 -1.4543    0.1459
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 391.894  on 331  degrees of freedom
## Residual deviance: 389.756  on 330  degrees of freedom
## AIC: 393.756
##
## Number of Fisher Scoring iterations: 4
```

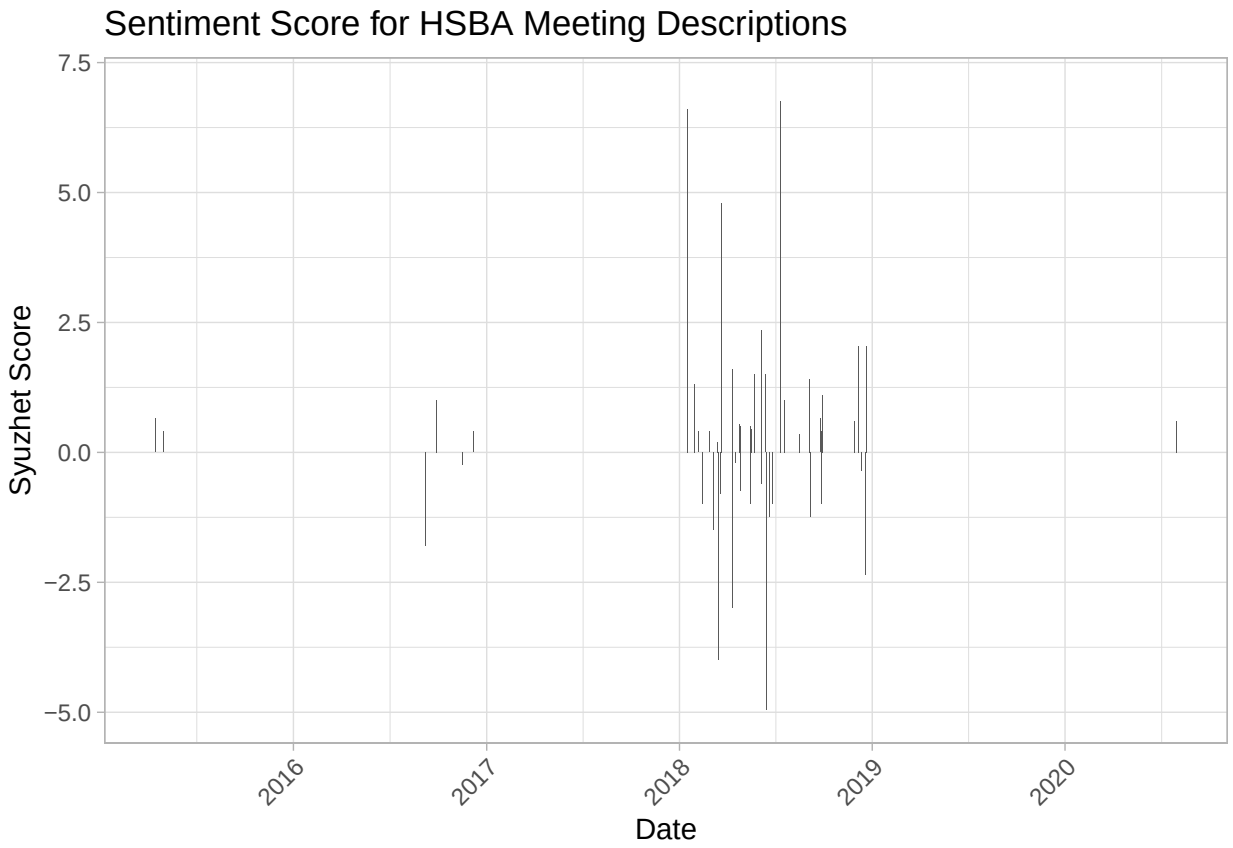
```
fit_3 <- glm(sociologist ~ syuzhet_vector, sst_recent, family = "binomial")
print(summary(fit_3), digits = 5)
```

```
##
## Call:
## glm(formula = sociologist ~ syuzhet_vector, family = "binomial",
##     data = sst_recent)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.31592  -0.25293  -0.24541  -0.23101   2.81023
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.42641    0.32584 -10.516   <2e-16 ***
## syuzhet_vector -0.24528    0.43875  -0.559   0.5761
```

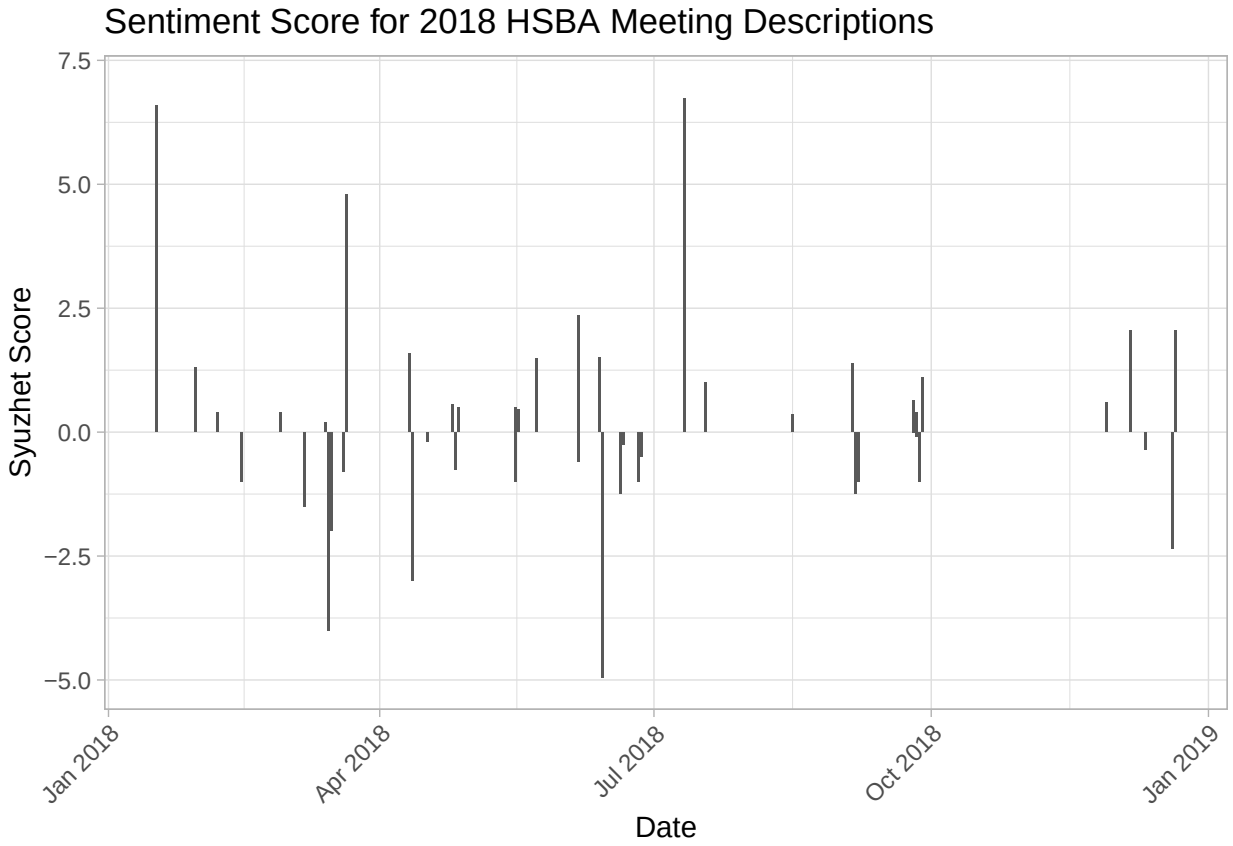
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 89.7467  on 331  degrees of freedom
## Residual deviance: 89.4342  on 330  degrees of freedom
## AIC: 93.4342
##
## Number of Fisher Scoring iterations: 6
```

Because the data were a bit limited, I retrieved some additional data using the ProPublica API for data on congress. For the most popular committee in the original dataset where most social scientists were present, HSBA, the sentiment of meetings spike in intensity in 2018 both positively and negatively.

```
HSBA <- gov_df %>%
  filter(committee_code == "HSBA")

HSBA$syuzhet_vector <- get_sentiment(HSBA$description, method="syuzhet")

HSBA <- HSBA %>%
  filter(syuzhet_vector != 0)

ggplot(HSBA, aes(date, syuzhet_vector)) +
  geom_col(show.legend = FALSE) +
  theme_light() +
  ggtitle("Sentiment Score for HSBA Meeting Descriptions") +
  xlab("Date") +
  ylab("Syuzhet Score") +
  ylim(-5, 7) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
HSBA$year <- str_extract(HSBA$date, "\\d{4}")

HSBA_recent <- HSBA %>%
  filter(year == 2018)

ggplot(HSBA_recent, aes(date, syuzhet_vector)) +
  geom_col(show.legend = FALSE) +
  theme_light() +
  ggtitle("Sentiment Score for 2018 HSBA Meeting Descriptions") +
  xlab("Date") +
  ylab("Syuzhet Score") +
  ylim(-5, 7) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The document term matrix and word frequency table suggest that most of the hearings at the time were related to regulations on investment protection.

```
TextDoc <- Corpus(VectorSource(HSBA_recent$description))

toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
TextDoc <- tm_map(TextDoc, toSpace, "/")
TextDoc <- tm_map(TextDoc, toSpace, "@")
TextDoc <- tm_map(TextDoc, toSpace, "\\|")

# Convert the text to lower case
TextDoc <- tm_map(TextDoc, content_transformer(tolower))

# Remove numbers
TextDoc <- tm_map(TextDoc, removeNumbers)

# Remove english common stopwords
TextDoc <- tm_map(TextDoc, removeWords, stopwords("english"))

# Remove punctuations
TextDoc <- tm_map(TextDoc, removePunctuation)

# Eliminate extra white spaces
TextDoc <- tm_map(TextDoc, stripWhitespace)
```

```
# Build a term-document matrix
TextDoc_dtm <- TermDocumentMatrix(TextDoc)
```

```
dtm_m <- as.matrix(TextDoc_dtm)

# Sort by descearing value of frequency
dtm_v <- sort(rowSums(dtm_m),decreasing=TRUE)
dtm_d <- data.frame(word = names(dtm_v),freq=dtm_v)
dtm_d <- dtm_d %>%
  rename("Word" = "word") %>%
  rename("Frequency" = "freq")

# Display the top 5 most frequent words
head(dtm_d, 5)
```

```
##                   Word Frequency
## act                act       107
## regulatory regulatory        27
## small           small        23
## investment investment        19
## protection protection        18
```

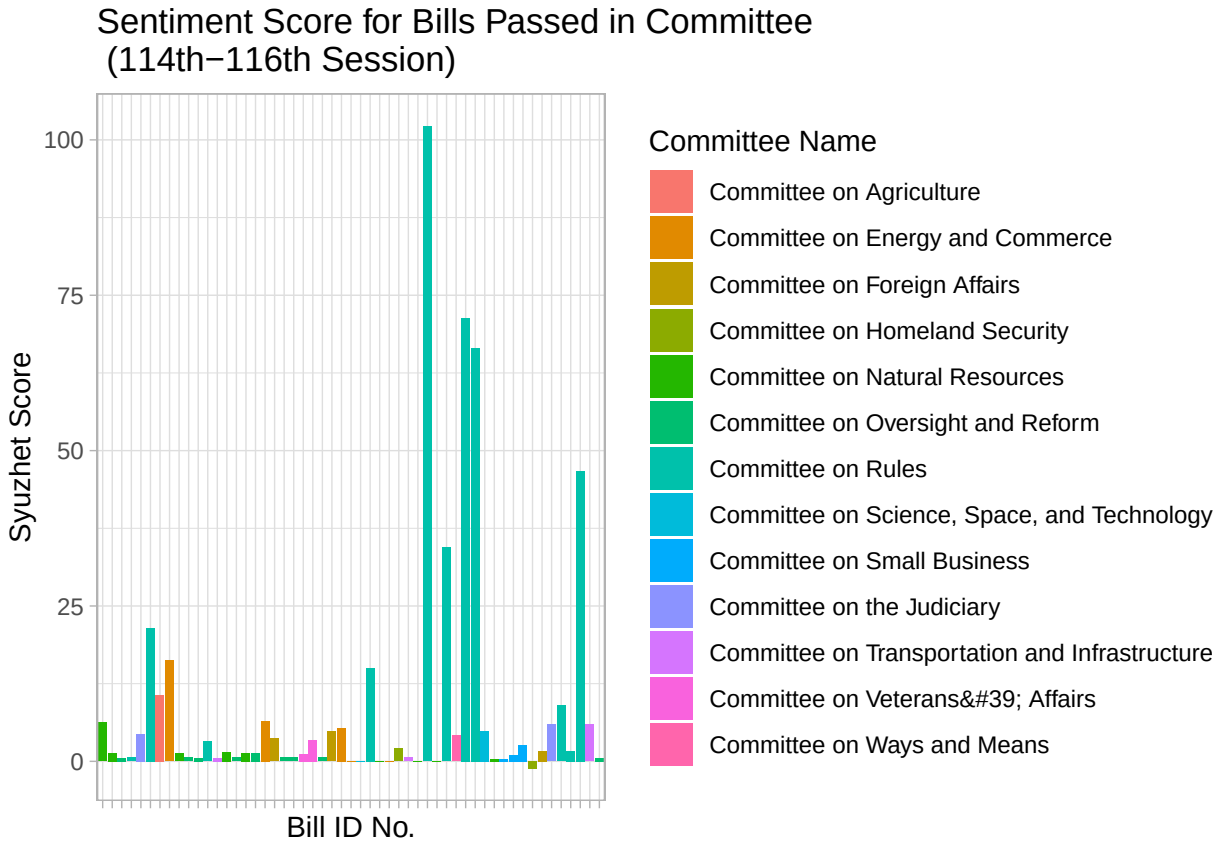
Combining the full API dataset on congressional hearings with one of bills that were passed and a sentiment analysis reveals that the committee most likely to pass bills is the Committee on Rules.

```
gov_df <- gov_df %>%
  rename("bill_id" = "bill_ids")

gov_bills <- gov_df %>% filter(bill_id != "n/a")
bills_df <- as.data.frame(bills_df)
```

```
joined_df <- inner_join(gov_bills, bills_df, by = "bill_id")
```

The chart created also shows that bills passed have overwhelmingly positive sentiments for their summaries.

```
joined_df$syuzhet_vector <- get_sentiment(joined_df$summary, method="syuzhet")

ggplot(joined_df, aes(bill_id, syuzhet_vector, fill = committee)) +
  geom_col(show.legend = TRUE) +
  theme_light() +
  ggtitle("Sentiment Score for Bills Passed in Committee \n (114th-116th Session)") +
  xlab("Bill ID No.") +
  ylab("Syuzhet Score") +
  labs(fill = "Committee Name") +
  theme(axis.text.x = element_blank())
```



This preliminary analysis is very limited with regards to the ability to answer my original research question. It was challenging finding ways to use the data from the government API in concert with the original dataset on testimonies when there were no related variables to join the datasets. However, through exploratory analysis, I did find some factors that do lead to "impactful" congressional hearings, most notably the sentiment and "intensity" of the hearing and/or bill description. The next steps for this project would be perhaps to find actual transcripts of the social scientists present to analyze the information shared and their persuasion tactics used. A hypothesis test can be completed on whether or not hearings with an expert present make a difference or not in the ultimate conclusion of the hearing. If we understand what makes a strong testimony, social scientists hoping to make a beneficial social impact can follow these guidelines to make sure their voices are heard in the elected body.