Third Obama-Romney Presidential Debate

Claimbuster Data, EDA

```
library(rvest)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
```

```
all_sentences = read_csv("data/all_sentences.csv")
```

```
## Rows: 32072 Columns: 10

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): Text, Speaker, Speaker_title, Speaker_party, Speaker_role, File_id
## dbl (4): Sentence_id, Length, Line_number, Sentiment

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
crowdsourced = read_csv("data/crowdsourced.csv")
```

```
## Rows: 22501 Columns: 10
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Text, Speaker, Speaker_title, Speaker_party, File_id
## dbl (5): Sentence_id, Length, Line_number, Sentiment, Verdict
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
groundtruth = read_csv("data/groundtruth.csv")
```

```
## Rows: 1032 Columns: 10
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Text, Speaker, Speaker_title, Speaker_party, File_id
## dbl (5): Sentence_id, Length, Line_number, Sentiment, Verdict
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
rom_ob_all = all_sentences %>%
  filter(File_id == "2012-10-22.txt")

rom_ob_crowd = crowdsourced %>%
  filter(File_id == "2012-10-22.txt")

rom_ob_truth = crowdsourced %>%
  filter(File_id == "2012-10-22.txt")
```

```
rom_ob_entire <- rom_ob_all %>%
  filter(Speaker == "Mitt Romney" |
         Speaker == "Barack Obama")
```

```
full_join(rom_ob_crowd, rom_ob_truth, Sentence_id = Sentence_id)
```

```
## Joining, by = c("Sentence_id", "Text", "Speaker", "Speaker_title", "Speaker_party", "File_id", "Leng
```

```
## # A tibble: 835 x 10
##    Sentence_id Text          Speaker  Speaker_title Speaker_party File_id Length
##          <dbl> <chr>         <chr>    <chr>         <chr>         <chr>    <dbl>
##  1       26902 And thank yo~ Mitt Ro~ Governor      REPUBLICAN    2012-1~     11
##  2       26903 Thank you to~ Mitt Ro~ Governor      REPUBLICAN    2012-1~      9
##  3       26904 And Mr. Pres~ Mitt Ro~ Governor      REPUBLICAN    2012-1~     10
```

```
##  4       26905 We were toge~ Mitt Ro~ Governor      REPUBLICAN    2012-1~     21
##  5       26907 This is obvi~ Mitt Ro~ Governor      REPUBLICAN    2012-1~     40
##  6       26908 With the Ara~ Mitt Ro~ Governor      REPUBLICAN    2012-1~     40
##  7       26909 But instead,~ Mitt Ro~ Governor      REPUBLICAN    2012-1~     13
##  8       26910 Of course we~ Mitt Ro~ Governor      REPUBLICAN    2012-1~     15
##  9       26911 We see in --~ Mitt Ro~ Governor      REPUBLICAN    2012-1~     29
## 10       26912 Our hearts a~ Mitt Ro~ Governor      REPUBLICAN    2012-1~     10
## # ... with 825 more rows, and 3 more variables: Line_number <dbl>,
## #   Sentiment <dbl>, Verdict <dbl>
```

```
#essentially, these datasets are identical
write.csv(rom_ob_truth, 'rom_ob.csv', row.names=FALSE)
```

README Info from Claimbuster Datasets

The ClaimBuster dataset consists of six files, the three used are the following: groundtruth.csv, (22,501 sentences) crowdsourced.csv, (1032 sentences) all_sentences.csv, (32,072 sentences)

Both groundtruth.csv and crowdsourced.csv files contain the following attributes. - Sentence_id: A unique numerical identifier to identify sentences in the dataset. - Text: A sentence spoken by a debate participant. - Speaker: Name of the person who verbalized the Text. - Speaker_title: Speaker's job at the time of the debate. - Speaker_party: Political affiliation of the Speaker. - File_id: Debate transcript name. - Length: Number of words in the Text.
- Line_number: A numerical identifier to to indicate the order of the Text in the debate transcript. - Sentiment: Sentiment score of the Text. The score ranges from -1 (most negative sentiment) to 1 (most positive sentiment). - Verdict: Assigned class label (1 when the sentence is CFS,0 when the sentence is UFS, and -1 when sentence is NFS).

all_sentences.csv file contains all presidential debate sentences. It has all the features shown above except for "Verdict' '. It also includes the following attribute: - Speaker_role: It depicts the role of the Speaker in the debate as a participant.