News Article Classification: EDA Report

Lauren Contard, Archit Datar, Bobby Lumpkin, Yue Li, Haihang Wu

3/1/2021

Introduction

Our project focuses on classification of news articles covering the White House's delivery of news related to covid-19. We begin with a sample of about 8000+ articles from ten mainstream media outlets which were selected using a keyword search. All of these articles are related to covid-19 and the White House in some way; however, not all are focused on White House covid briefings, which is the desired focus of our research. 1022 of the articles in our sample have been classified by hand into:
0 = "not related to White House briefings about covid-19" or
1 = "related to White House briefings about covid-19"
(Note: these articles were randomly selected from the larger sample, so they should be representative of the full sample of articles.)

Our goal will be to use this sample to build a classifier for the remaining articles. We will do this using the counts of various words that appear in the articles' text; in this exploratory data analysis, we will examine which words may be the most useful as predictors.

After all the news articles are classified, we will next classify the relavant articles into multiple categories based on the content, such as threat of covid-19, organizational response, self-congratualation, criticizing the government, etc. We will also conduct sentiment analysis to analyze the tone of the news articles. These two steps will not be covered in the exploratory data analysis as our dataset is not clean yet.

Text Preprocessing

We began by processing the text of the 1022 classified articles using the "quanteda" package. In this step we:
-created tokens for all words that appear
-removed stop words such as "a", "the", etc.
-stemmed the tokens, e.g. converting "learning" and "learned" to "learn"
-filtered out words that appear in less than 2.5% and more than 97.5% of articles, as these words may be less useful for prediction

The head of the document-feature matrix is below:

```
## Document-feature matrix of: 10 documents, 10 features (1.0% sparse) and 12
docvars.
##                                      features
## docs                                  said trump state coronavirus presid
s
```

```
##   dataframe_with_article_data.csv.39    98     1   103          51      2
67
##   dataframe_with_article_data.csv.183   68    40    10          42     33
21
##   dataframe_with_article_data.csv.555   67     6    87          50      8
62
##   dataframe_with_article_data.csv.915   67     4    84          56      2
54
##   dataframe_with_article_data.csv.347   77     7    92          44      7
48
##   dataframe_with_article_data.csv.993   82     1    87          49      3
59
##                                      features
## docs                                  peopl test new health
##   dataframe_with_article_data.csv.39     29   32  34     57
##   dataframe_with_article_data.csv.183    23   16  10     11
##   dataframe_with_article_data.csv.555    37   19  21     38
##   dataframe_with_article_data.csv.915    31   31  33     37
##   dataframe_with_article_data.csv.347    27   51  32     44
##   dataframe_with_article_data.csv.993    32   24  22     37
## [ reached max_ndoc ... 4 more documents ]
```

Exploring the Data

We now have a data frame with the counts of each tokenized word. 15,267 words were included; the first 50 words are shown here as examples:

```
##  [1] "doc_id"      "coronavirus" "barr"        "say"         "draconian"
##  [6] "rule"        "may"         "need"        "revisit"     "soon"
## [11] "attorney"    "general"     "william"     "wednesday"   "call"
## [16] "restrict"    "effect"      "mani"        "state"       "mitig"
## [21] "spread"      "said"        "next"        "month"       "ask"
## [26] "fox"         "news"        "host"        "laura"       "ingraham"
## [31] "balanc"      "protect"     "peopl"       "like"        "stayathom"
## [36] "order"       "feder"       "govern"      "keep"        "care"
## [41] "eye"         "use"         "broad"       "power"       "regul"
## [46] "live"        "citizen"     "offici"      "make"        "surethat"
```

We can now compare the distribution of words in the relevant and irrelevant articles. The distribution of the response is:

```
##
##   0   1
## 398 624

##
##         0         1
## 0.3894325 0.6105675
```

i.e., about 61.1% of the articles focused on White House briefings, and 38.9% did not (based on human classification).

Below, we look to see how frequently a given word appears in related articles, vs. how frequently in non-related articles. We can then examine the words with the largest difference between those two groups. These words might be most useful as features.

View the mean frequency of a word's appearance in related articles, for a sample of 10 words:

```
## coronavirus        barr         say   draconian        rule         may
## 4.139423077 0.024038462 1.897435897 0.011217949 0.189102564 0.786858974
##        need     revisit        soon    attorney
## 1.403846154 0.006410256 0.275641026 0.091346154
```

And the same for non-related articles:

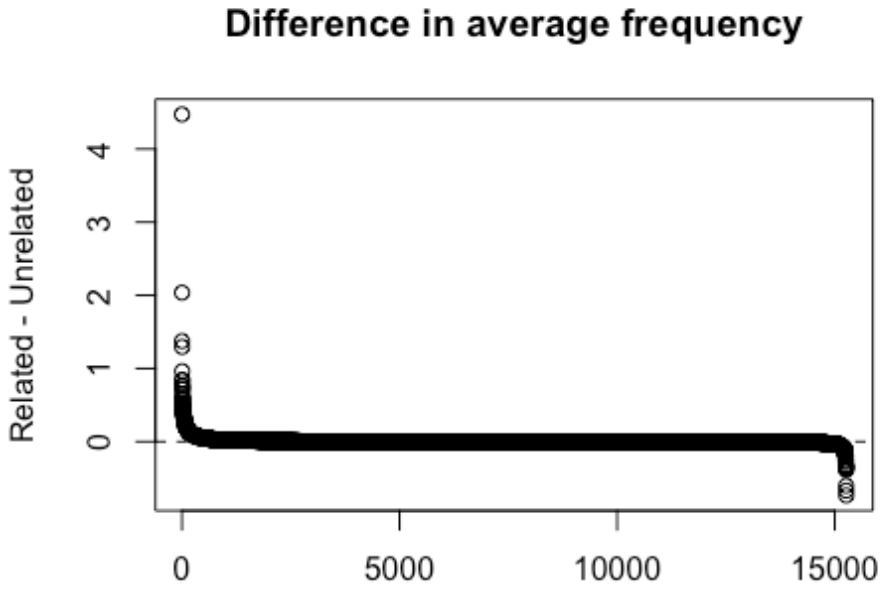
```
## coronavirus        barr         say   draconian        rule         may
## 3.545226131 0.040201005 1.494974874 0.010050251 0.278894472 0.723618090
##        need     revisit        soon    attorney
## 0.891959799 0.002512563 0.165829146 0.155778894
```

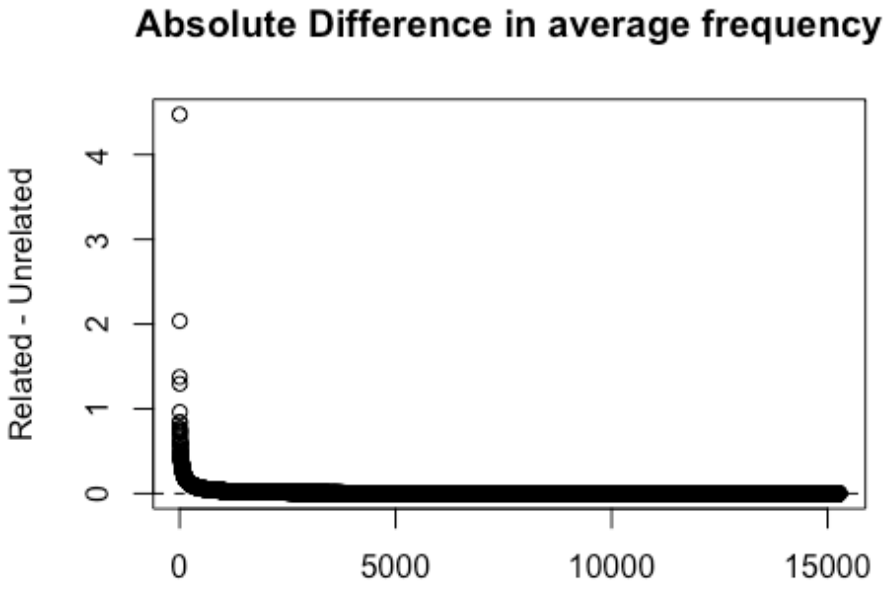
Plotting Data

The 10 largest differences (related cases - non-related cases) between these frequencies are below:

```
##     trump    presid      hous     white  american        us     fauci
test
##    4.4728    2.0365    1.3746    1.2947    0.9629    0.8456    0.8362
0.7740
##      said administr
##    0.7588    0.7268
```

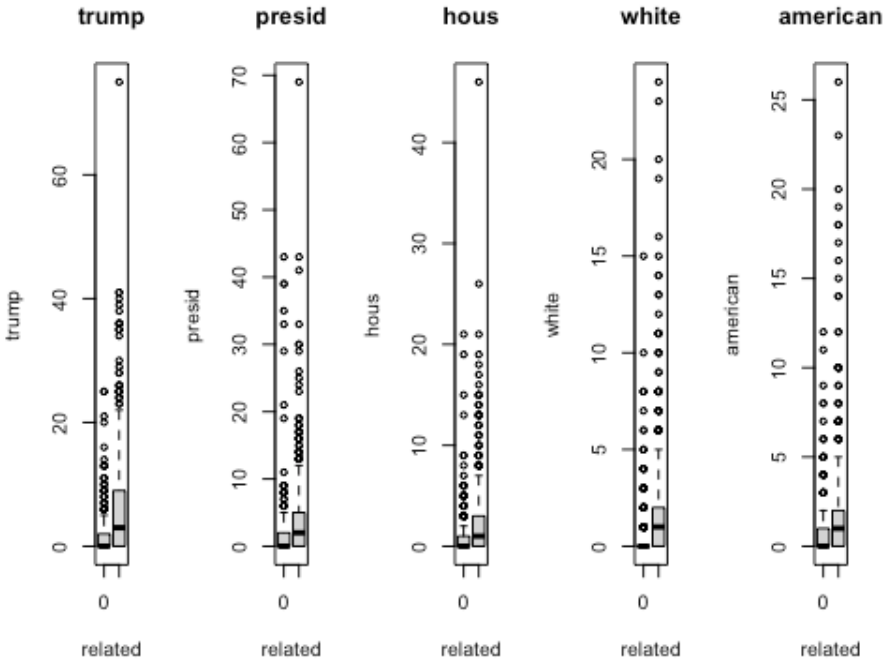
These are the words that appear to be most strongly associated with related articles. We can visualize the distribution of raw and absolute differences in frequencies over all words:



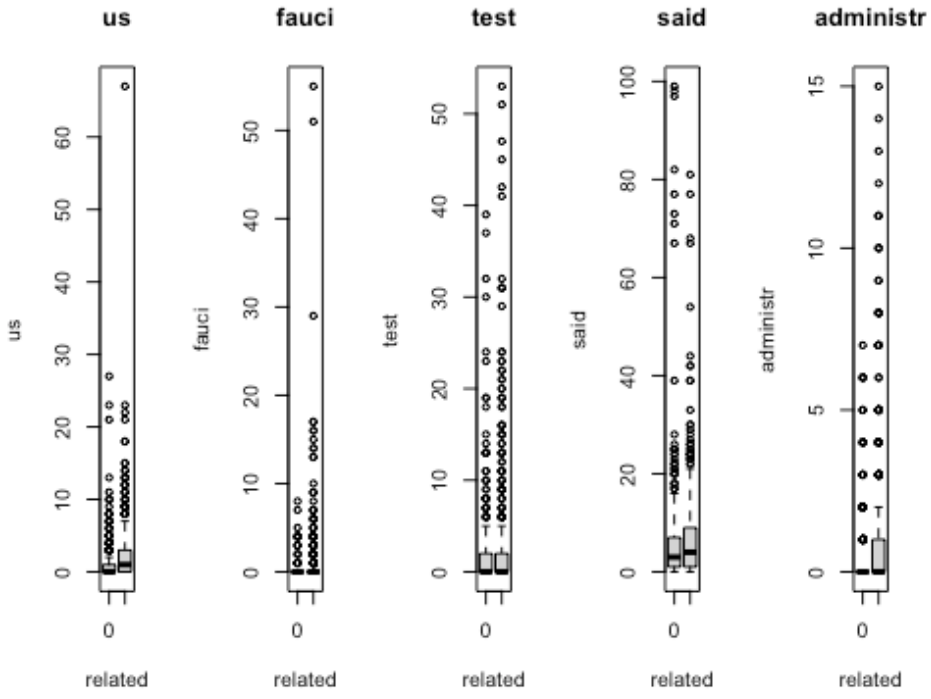


We will consider the top 50 predictors in absolute difference value for the following analysis. We will visualize the differences in these predictors with box plots.

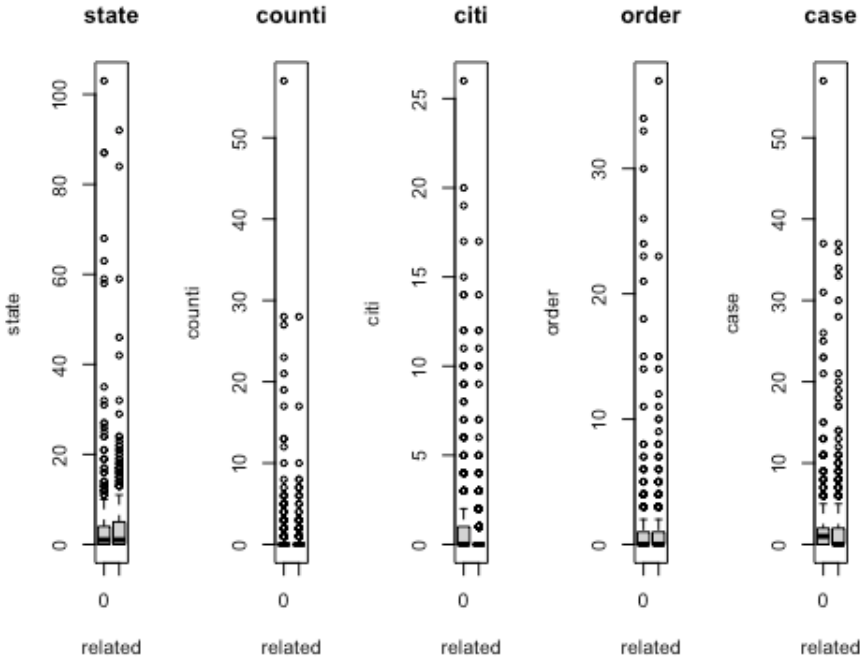
The 5 variables with the strongest positive difference (i.e., more frequent in "related" than "non-related"):



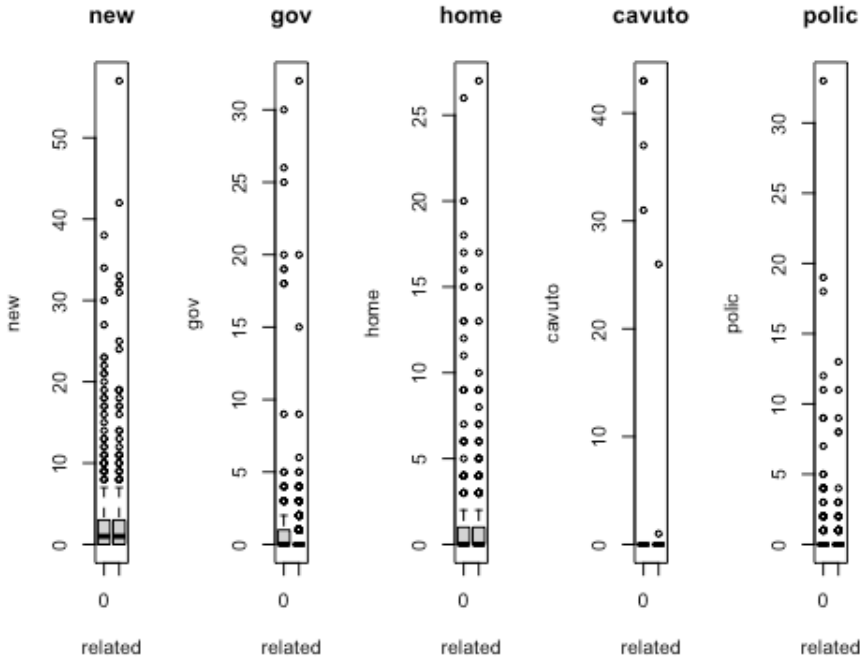
The variables with the next 5 strongest positive differences:



The 5 variables with the strongest negative differences (i.e., more frequent in “non-related” than “related”:



And the variables with the next 5 strongest negative differences:



In general, it seems that all of these words have heavily right-skewed distributions. In other words, any given word does not appear in most articles, but appears many times in a few articles. This is to be expected, but may need to be taken into account if we attempt to use classification methods that require assumptions on the distribution of the predictors.

Further, we can see that the differences in distributions appear to be larger for the positive effects. There are more words that clearly appear more frequently in related articles, than there are words that clearly appear more frequently in non-related articles. This makes sense, as the related articles are, by definition, all about the same topic, while the non-related articles may be about many different topics that happen to mention certain covid-related keywords. Thus, we expect less variation among the related articles. This suggests that the predictors with positive differences may end up being the most useful in classification.

Principal Component Analysis

We also perform a principal component analysis to examine which are the most useful variables to visualize the data in lower dimensions.

PCA was performed using the 50 variables with the largest differences. Below are the ratios of overall variance explained by the first 5 components:

```
## [1] 0.35132421 0.10057750 0.06644591 0.03743792 0.03254088
```

We can see that the variances for the first 5 components represent > 50% of the data, but those from the other components cannot be neglected.

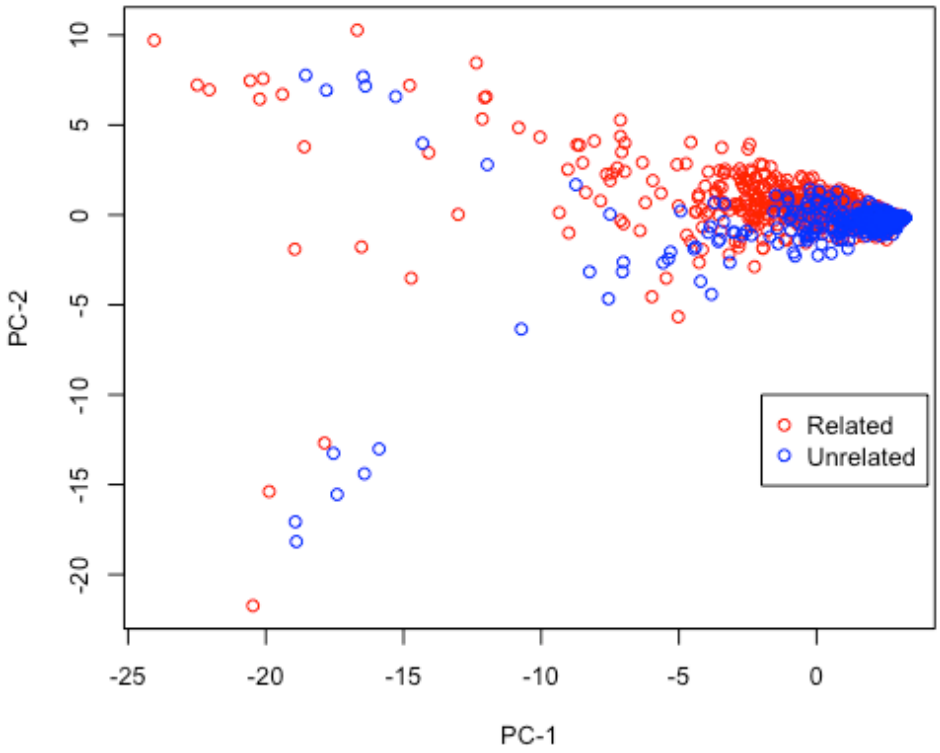
Below are the principal component scores for the first two components:

```
##                     PC1         PC2
## trump       -0.13140881  0.16213292
## presid      -0.16537348  0.18868126
## hous        -0.14706814  0.15224410
## white       -0.13729504  0.17551737
## american    -0.13279631  0.15936469
## us          -0.14233038  0.10097547
## fauci       -0.07424760  0.12165414
## test        -0.13209850 -0.12806438
## said        -0.17354651 -0.20817838
## state       -0.16555225 -0.26240362
## administr   -0.12520832  0.10943817
## brief       -0.09178593  0.09901662
## counti      -0.10541803 -0.30704897
## go          -0.15989686  0.14646683
## citi        -0.12377439 -0.16617442
## coronavirus -0.17103376 -0.20829006
## nation      -0.16520906  0.02990027
## virus       -0.17076799 -0.10013189
## respons     -0.13041728  0.05067697
## public      -0.15177287 -0.10935926
## need        -0.17293380  0.05139384
## penc        -0.06185538  0.07727450
## countri     -0.14540135  0.08758194
## task        -0.10251024  0.11858270
## forc        -0.12251318  0.10111286
## diseas      -0.13791876 -0.09761069
## dr          -0.14598677  0.07234098
## china       -0.07766890  0.09322219
## use         -0.12841558  0.01142601
## can         -0.18247140  0.04621153
## s           -0.14180804 -0.24719495
## say         -0.17807028  0.08566853
## pandem      -0.15821049 -0.11023534
## get         -0.17661866  0.11581399
## time        -0.18456205  0.08093056
## work        -0.18044416  0.02014644
## order       -0.12645092 -0.26779129
## expert      -0.11275008  0.06455507
## make        -0.17786510  0.09312514
## ask         -0.16259515  0.04573716
## new         -0.15496589 -0.16061466
## news        -0.14881244  0.03945187
## gov         -0.12587004 -0.32774554
## case        -0.14366234 -0.22196130
## press       -0.09464894  0.08674826
## even        -0.17724720  0.07580088
## unit        -0.10239023  0.03238822
```

```
## cavuto       -0.06550154  0.08851518
## like         -0.12708535  0.09399307
## vaccin       -0.05235750  0.04752924
```

It is, in general, hard to draw inferences about the first few components from the words as features. The first component, however, reprensents the sum of the features associated with all the words, while the second component represents the differenes in the features associated with America-specific words such as “trump”, “presid”, “hous”, “white”, “fauci” and more general words like “state”, “citi”, “coronavirus”, “virus”, and “public”.

Plot the first vs. second principal component scores:



By visualizing the data along the 1st and 2nd components, we can see that the related points seem to be at a higher value of the 2nd component. This makes sense since the the second component seems to represent the differences between the values for related and unrelated words. There seems to be a great deal of skew in the data, especially in the first component. Regardless, it does seem that the related and unrelated articles are at least somewhat different based on the features we have used, indicating that they have at least some predictive power.

Conclusions

Overall, our exploratory data analysis has identified which words are likely to be the most useful in predicting whether an article in our sample is related to White House covid-19 briefings. We can see that there are a number of words that appear more frequently in related articles, and a smaller number of words that appear more frequently in non-related articles. These give us a good starting point for what to focus on in building our classifier.