

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.

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:

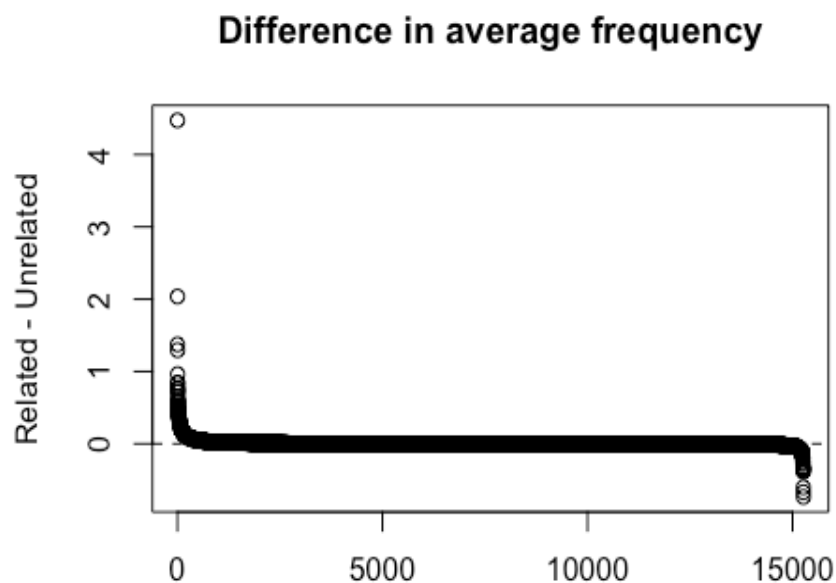
```
## coronavirus      barr      say      draconian      rule      may
## 3.545226131 0.040201005 1.494974874 0.010050251 0.278894472 0.723618090
##      need      revisit      soon      attorney
## 0.891959799 0.002512563 0.165829146 0.155778894
```

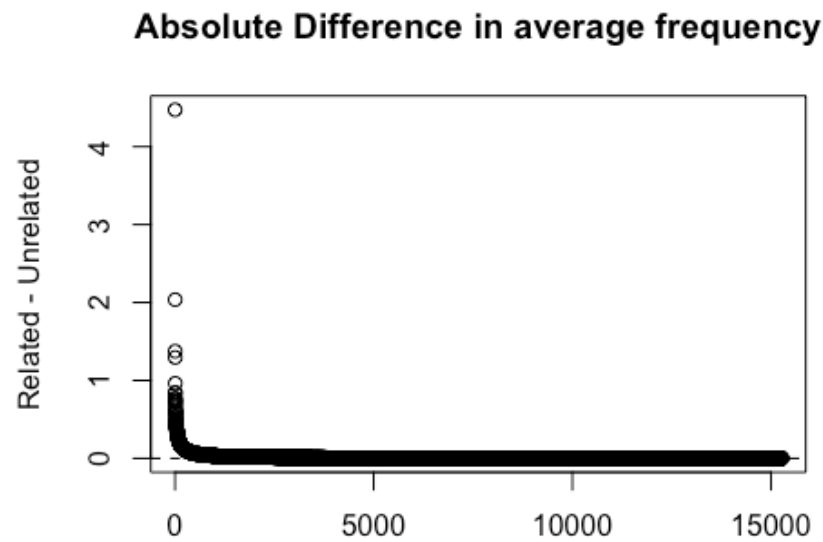
Plotting Data

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
```

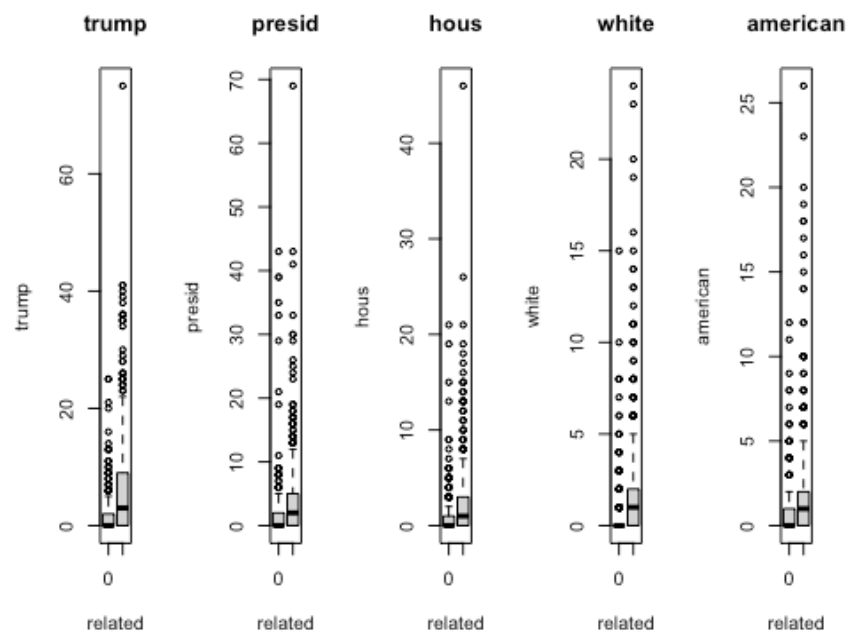
These are the words that appear to be most strongly associated with related articles. We can visualize the distribution of raw and absolute differences in frequencies over all words:



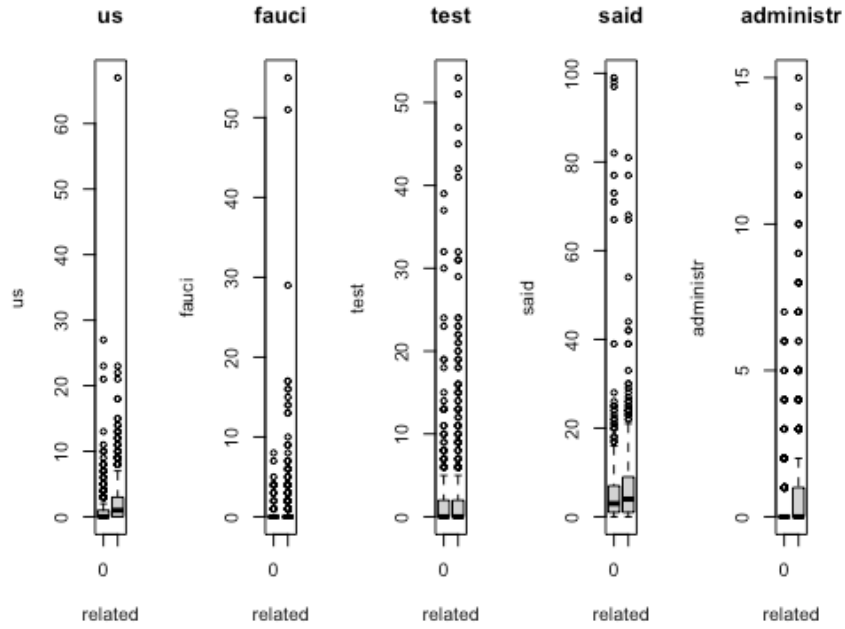


We will consider the top 50 predictors in absolute difference value for the following analysis. We will visualize the differences in these predictors with box plots.

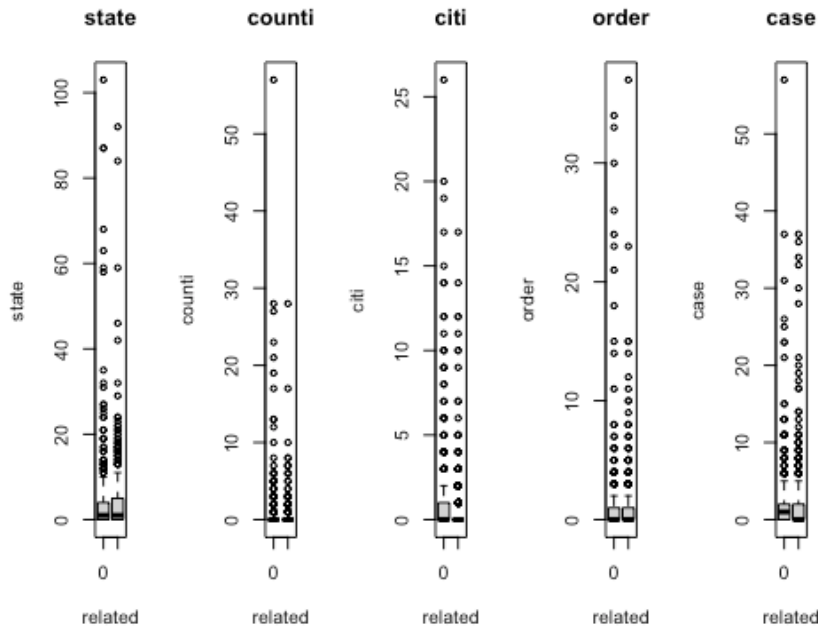
The 5 variables with the strongest positive difference (i.e., more frequent in “related” than “non-related”):



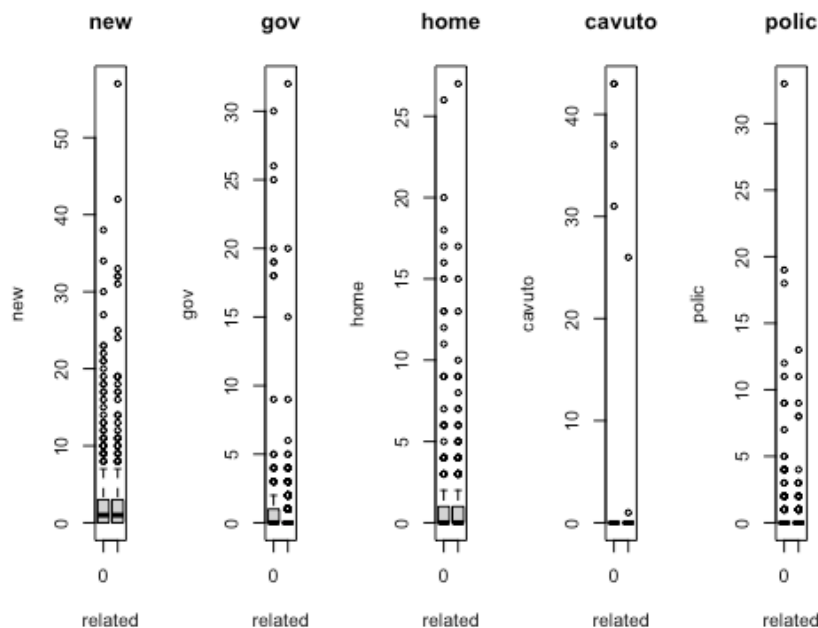
The variables with the next 5 strongest positive differences:



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 for classification.

```
## [1] 4.1912063 2.2425153 1.8227165 1.3681726 1.2755564 1.1660214 1.0800051
## [8] 0.9726369 0.9619879 0.9344458 0.9281251 0.9053505 0.8988751 0.8665701
## [15] 0.8273307 0.8096184 0.7873033 0.7686862 0.7552953 0.7350643 0.7207339
## [22] 0.7123993 0.6955466 0.6795812 0.6657618 0.6529894 0.6485141 0.6323966
## [29] 0.6095759 0.5949783 0.5927999 0.5724817 0.5607616 0.5450591 0.5236203
## [36] 0.5061590 0.4993489 0.4734990 0.4699699 0.4533407 0.4392762 0.4253091
## [43] 0.4206691 0.3718023 0.3622366 0.3460633 0.3374982 0.3066742 0.2843116
## [50] 0.2710060
```

```

## [1] 0.351324206 0.100577497 0.066445909 0.037437925 0.032540883
0.027192120
## [7] 0.023328221 0.018920450 0.018508414 0.017463779 0.017228324
0.016393191
## [13] 0.016159531 0.015018874 0.013689522 0.013109639 0.012396928
0.011817570
## [19] 0.011409419 0.010806389 0.010389148 0.010150255 0.009675702
0.009236613
## [25] 0.008864775 0.008527902 0.008411411 0.007998510 0.007431654
0.007079985
## [31] 0.007028235 0.006554705 0.006289072 0.005941789 0.005483565
0.005123939
## [37] 0.004986986 0.004484026 0.004417433 0.004110355 0.003859271
0.003617757
## [43] 0.003539249 0.002764739 0.002624307 0.002395197 0.002278101
0.001880981
## [49] 0.001616662 0.001468885

```

##	PC1	PC2	PC3	PC4	PC5
## trump 0.167424104	-0.13140881	0.16213292	0.177274645	-0.1489277137	
## presid 0.084055799	-0.16537348	0.18868126	-0.037412401	-0.1806726620	
## hous 0.052084540	-0.14706814	0.15224410	0.153992580	-0.2472049106	
## white 0.032678149	-0.13729504	0.17551737	0.195054211	-0.2375821317	
## american 0.163260147	-0.13279631	0.15936469	0.026664609	0.0700477536	
## us 0.130616049	-0.14233038	0.10097547	0.009607129	0.2583902579	-
## fauci 0.418039297	-0.07424760	0.12165414	0.177291289	0.1954569555	-
## test 0.050700242	-0.13209850	-0.12806438	0.041342505	-0.0249797315	
## said 0.010360378	-0.17354651	-0.20817838	0.082252943	-0.0911458190	
## state 0.003976998	-0.16555225	-0.26240362	-0.040637832	-0.0420849725	-
## administr 0.254552729	-0.12520832	0.10943817	0.201336392	-0.1009248748	
## brief 0.181821408	-0.09178593	0.09901662	0.123468298	-0.1744077159	-
## counti 0.087776316	-0.10541803	-0.30704897	-0.011953113	-0.0967187208	-
## go 0.098443377	-0.15989686	0.14646683	-0.296926037	-0.0251216161	-
## citi 0.046416780	-0.12377439	-0.16617442	-0.144820013	-0.0774009928	-

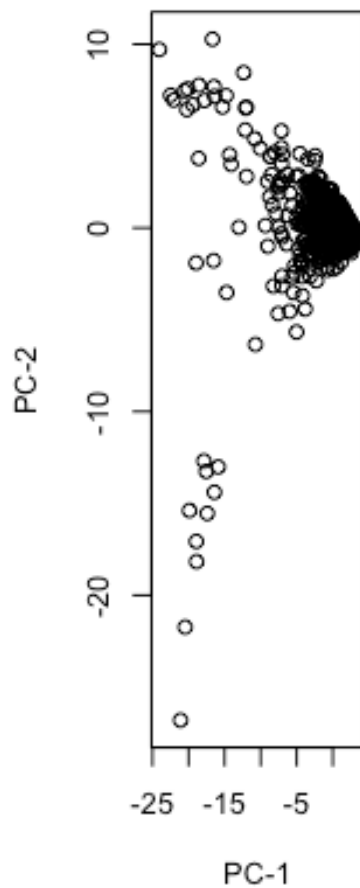
## coronavirus	-0.17103376	-0.20829006	0.121985949	-0.0218859034
0.023663583				
## nation	-0.16520906	0.02990027	0.091282103	0.0449557415
0.150280191				
## virus	-0.17076799	-0.10013189	0.077482070	0.1610546364
0.090858954				
## respons	-0.13041728	0.05067697	0.200624869	-0.1033417550
0.126786499				
## public	-0.15177287	-0.10935926	0.170603918	0.0043634236
0.021196287				
## need	-0.17293380	0.05139384	-0.142951259	0.0263754713
0.012289401				
## penc	-0.06185538	0.07727450	0.114787669	-0.2192751000
0.002733546				
## countri	-0.14540135	0.08758194	-0.021396996	0.2820530126
0.080929589				
## task	-0.10251024	0.11858270	0.256384967	0.0437658868 -
0.393681776				
## forc	-0.12251318	0.10111286	0.218677526	0.0227513627 -
0.349148995				
## diseas	-0.13791876	-0.09761069	0.141657549	0.2372003694
0.038524743				
## dr	-0.14598677	0.07234098	0.035777503	0.2506225060 -
0.206680606				
## china	-0.07766890	0.09322219	-0.001152061	0.2795344723
0.232561699				
## use	-0.12841558	0.01142601	0.037373121	-0.0009696921
0.174482208				
## can	-0.18247140	0.04621153	-0.213828354	-0.0645198184 -
0.033650165				
## s	-0.14180804	-0.24719495	0.063008860	-0.0361145199
0.029888515				
## say	-0.17807028	0.08566853	-0.126382495	0.0441226372 -
0.120546749				
## pandem	-0.15821049	-0.11023534	0.076573310	0.1144800418
0.126731160				
## get	-0.17661866	0.11581399	-0.257322729	-0.0256518145 -
0.063450603				
## time	-0.18456205	0.08093056	-0.119830182	-0.0197467806
0.014639009				
## work	-0.18044416	0.02014644	-0.122113560	-0.0232022329
0.060069441				
## order	-0.12645092	-0.26779129	-0.024116661	-0.1074304197 -
0.058405296				
## expert	-0.11275008	0.06455507	0.184874540	0.1647965953
0.041304499				
## make	-0.17786510	0.09312514	-0.186484651	-0.0101436564 -
0.038781837				
## ask	-0.16259515	0.04573716	-0.051677755	-0.0820675790 -
0.081089523				


```

## new      -0.15496589 -0.16061466 -0.089593683  0.0232687636 -
0.022663748
## news     -0.14881244  0.03945187 -0.067316715 -0.1057623911
0.055189902
## gov      -0.12587004 -0.32774554 -0.002069597 -0.0961764758 -
0.065229789
## case     -0.14366234 -0.22196130 -0.006150618  0.1239485915 -
0.136626179
## press    -0.09464894  0.08674826  0.131898783 -0.2412647128 -
0.036080441
## even     -0.17724720  0.07580088 -0.086407687  0.0272796221
0.040228602
## unit     -0.10239023  0.03238822 -0.041856851  0.2555243820
0.236275138
## cavuto   -0.06550154  0.08851518 -0.306220289 -0.0694460136 -
0.066019222
## like     -0.12708535  0.09399307 -0.225018500 -0.0573034088 -
0.050652758
## vaccin   -0.05235750  0.04752924  0.078795943  0.2168198701
0.128393349

```

Plot the first vs. second principal component scores:



From the PCA, 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. Also, it is, in general, hard to draw inferences about the first few components from the words as features. The first component, however, represents the sum of the features associated with all the words, while the second component represents the differences 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”.

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.