Assessing Similarities and Differences Between News Organizations

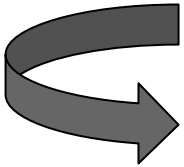
Devanshi Deswal, Samar Dikshit, Connor Higgins, Kartheek Karnati, Oliver Spohngellert

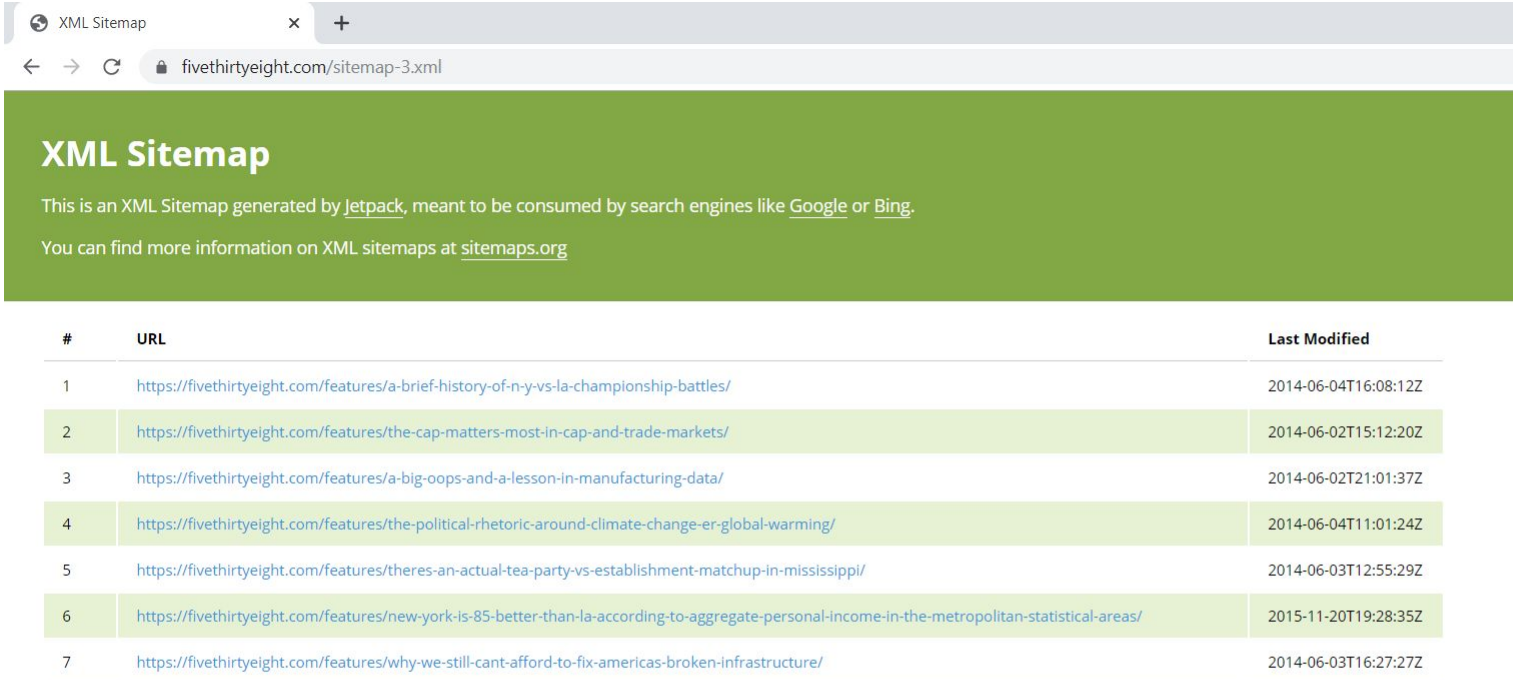
Introduction - Goals

- Explore the differences and similarities between left, right, and center news organizations.
- Hypothesized that these differences could be seen in headlines, and text of articles on political news sites.
- See differences between news sites on the same side of the political spectrum.

- All of our data was collected directly from political news sites
- Procedure: Assemble list of article links, then collect data
  - **All available articles since August 2019** were collected
  - Links were assembled using the website sitemap (see below)
  - Dataset attributes include **article link, article headline, author, article text, news site,** and **political lean**

Data Collection





#	URL	Last Modified
1	https://fivethirtyeight.com/features/a-brief-history-of-n-y-vs-la-championship-battles/	2014-06-04T16:08:12Z
2	https://fivethirtyeight.com/features/the-cap-matters-most-in-cap-and-trade-markets/	2014-06-02T15:12:20Z
3	https://fivethirtyeight.com/features/a-big-oops-and-a-lesson-in-manufacturing-data/	2014-06-02T21:01:37Z
4	https://fivethirtyeight.com/features/the-political-rhetoric-around-climate-change-er-global-warming/	2014-06-04T11:01:24Z
5	https://fivethirtyeight.com/features/theres-an-actual-tea-party-vs-establishment-matchup-in-mississippi/	2014-06-03T12:55:29Z
6	https://fivethirtyeight.com/features/new-york-is-85-better-than-la-according-to-aggregate-personal-income-in-the-metropolitan-statistical-areas/	2015-11-20T19:28:35Z
7	https://fivethirtyeight.com/features/why-we-still-cant-afford-to-fix-americas-broken-infrastructure/	2014-06-03T16:27:27Z

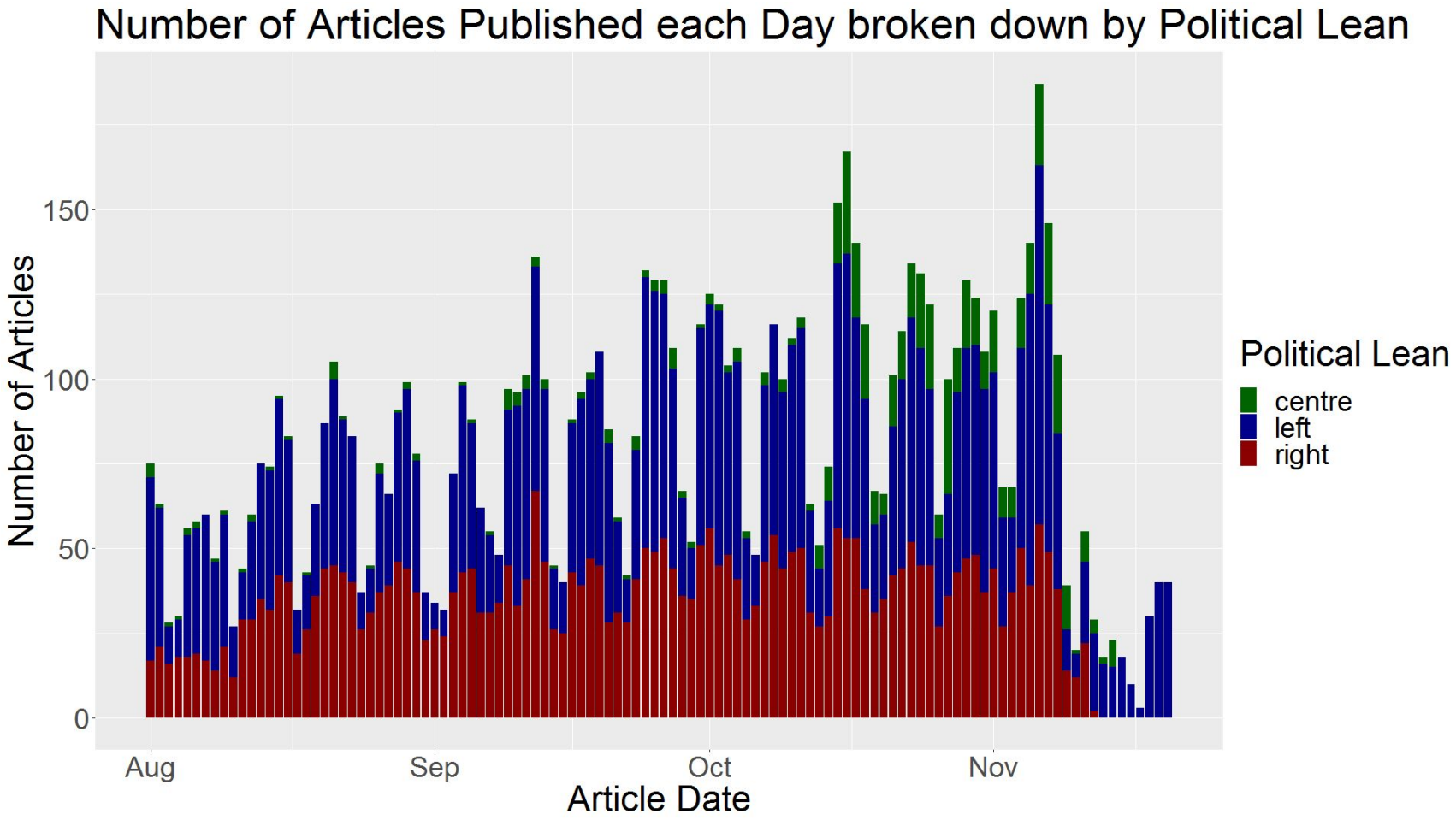
```
# A tibble: 8,992 x 14
   source      link      lastmod  filed_under article_date article_headline     article_author
   <chr>       <chr>     <chr>    <chr>       <chr>        <chr>                <chr>
 1 https://ww~ https://~ 2019-11~ world-us-c~ 2019-09-27   Trump wall - all yo~ ""
 2 https://ww~ https://~ 2019-11~ world-us-c~ 2019-11-06   Roger Stone: Trump ~ ""
 3 https://ww~ https://~ 2019-11~ world-us-c~ 2019-08-01   Democratic debates:~ ""
 4 https://ww~ https://~ 2019-11~ world-us-c~ 2019-08-01   Kelly Craft: Congre~ ""
 5 https://ww~ https://~ 2019-11~ world-us-c~ 2019-08-01   Democratic debate w~ ""
 6 https://ww~ https://~ 2019-11~ world-us-c~ 2019-08-01   Steve Bannon and Ga~ ""
 7 https://ww~ https://~ 2019-11~ world-us-c~ 2019-08-02   John Ratcliffe: Tru~ ""
 8 https://ww~ https://~ 2019-11~ world-us-c~ 2019-08-03   INF nuclear treaty:~ ""
 9 https://ww~ https://~ 2019-11~ world-us-c~ 2019-08-04   El Paso and Dayton:~ ""
10 https://ww~ https://~ 2019-11~ world-us-c~ 2019-08-05   US mass shootings: ~ ""
# ... with 8,982 more rows
```

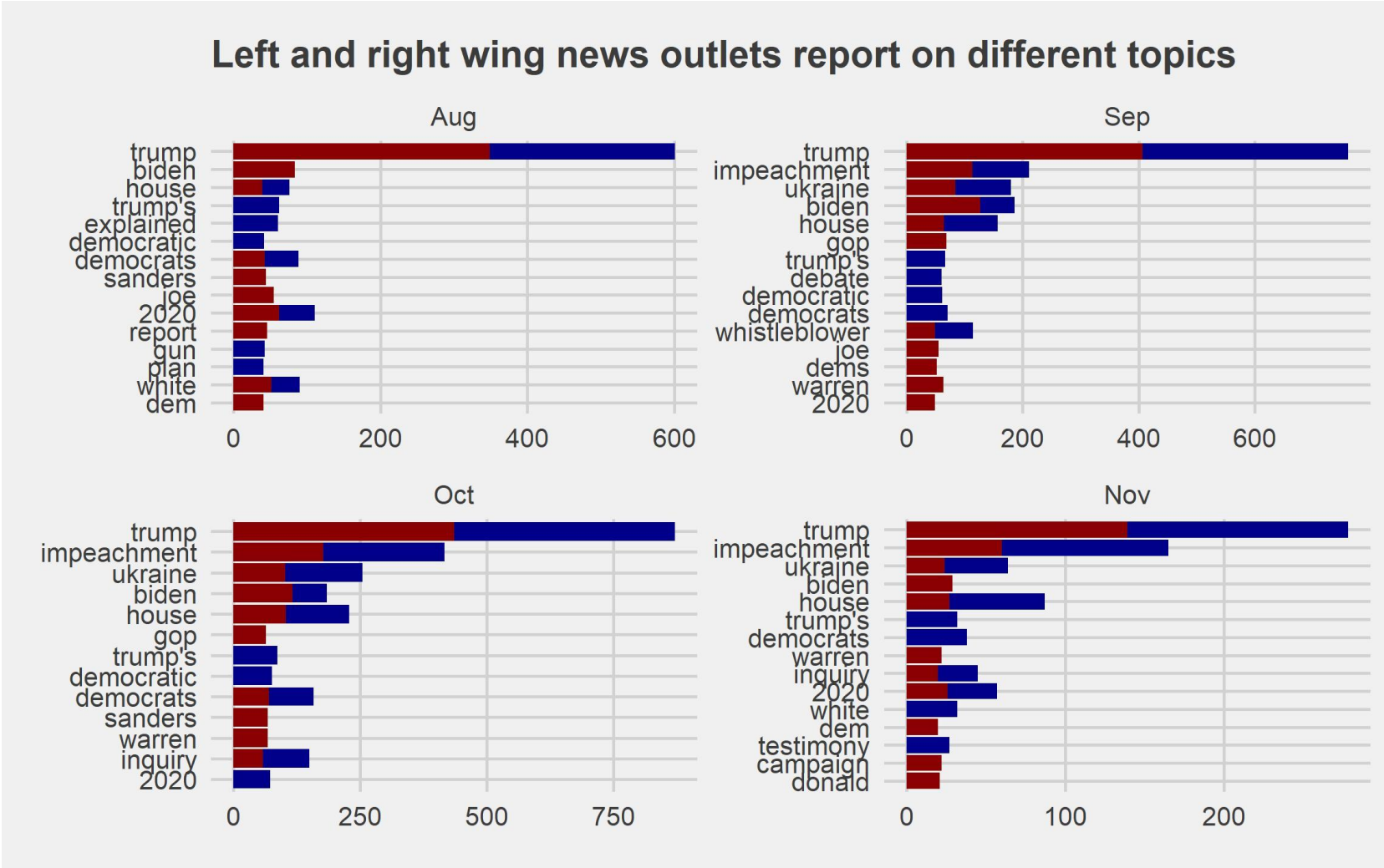
article_tag	article_text	http_status_code	collection_date	political_lean	news_site
<chr>	<chr>	<int>	<chr>	<chr>	<chr>
Mexicoâ€"US bord~	"President Donald Trump d~	200	2019-11-14T23:~	centre	bbc
Mueller Trump-Ru~	"Roger Stone, a long-time~	200	2019-11-14T23:~	centre	bbc
US election 2020	"Former Vice-President Jo~	200	2019-11-14T23:~	centre	bbc
Donald TrumpUnit~	"The US senate has confir~	200	2019-11-14T23:~	centre	bbc
US election 2020	"Another month, another d~	200	2019-11-14T23:~	centre	bbc
Donald TrumpUS p~	"This is the tale of two ~	200	2019-11-14T23:~	centre	bbc
Donald TrumpUnit~	"US President Donald Trum~	200	2019-11-14T23:~	centre	bbc
Nuclear weaponsR~	"US President Donald Trum~	200	2019-11-14T23:~	centre	bbc
US gun violenceN~	"It's become a familiar r~	200	2019-11-14T23:~	centre	bbc
US gun violenceD~	"President Donald Trump s~	200	2019-11-14T23:~	centre	bbc

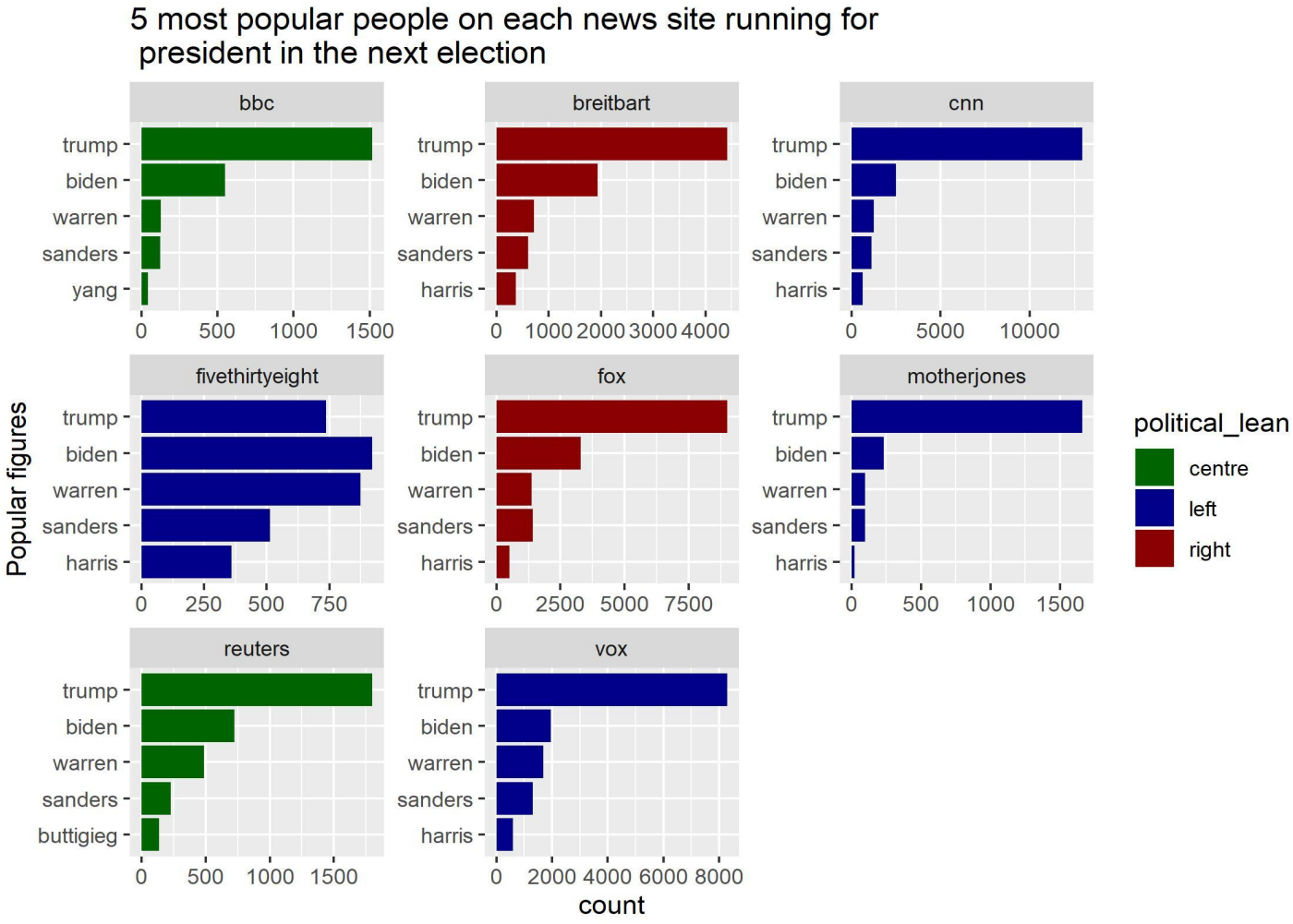
Filtering the Data

- We kept articles published only on or after 1st August 2019
- Web scraping gave us articles related to many different news categories, but using article tags and keywords, we filtered them down to political articles
- Basic filter pipeline:
  1. Use tags/keywords to filter articles
  2. Check if the article text isn't *NA*
  3. Apply date filter
  4. Add political lean and news site columns

Data after Filtering



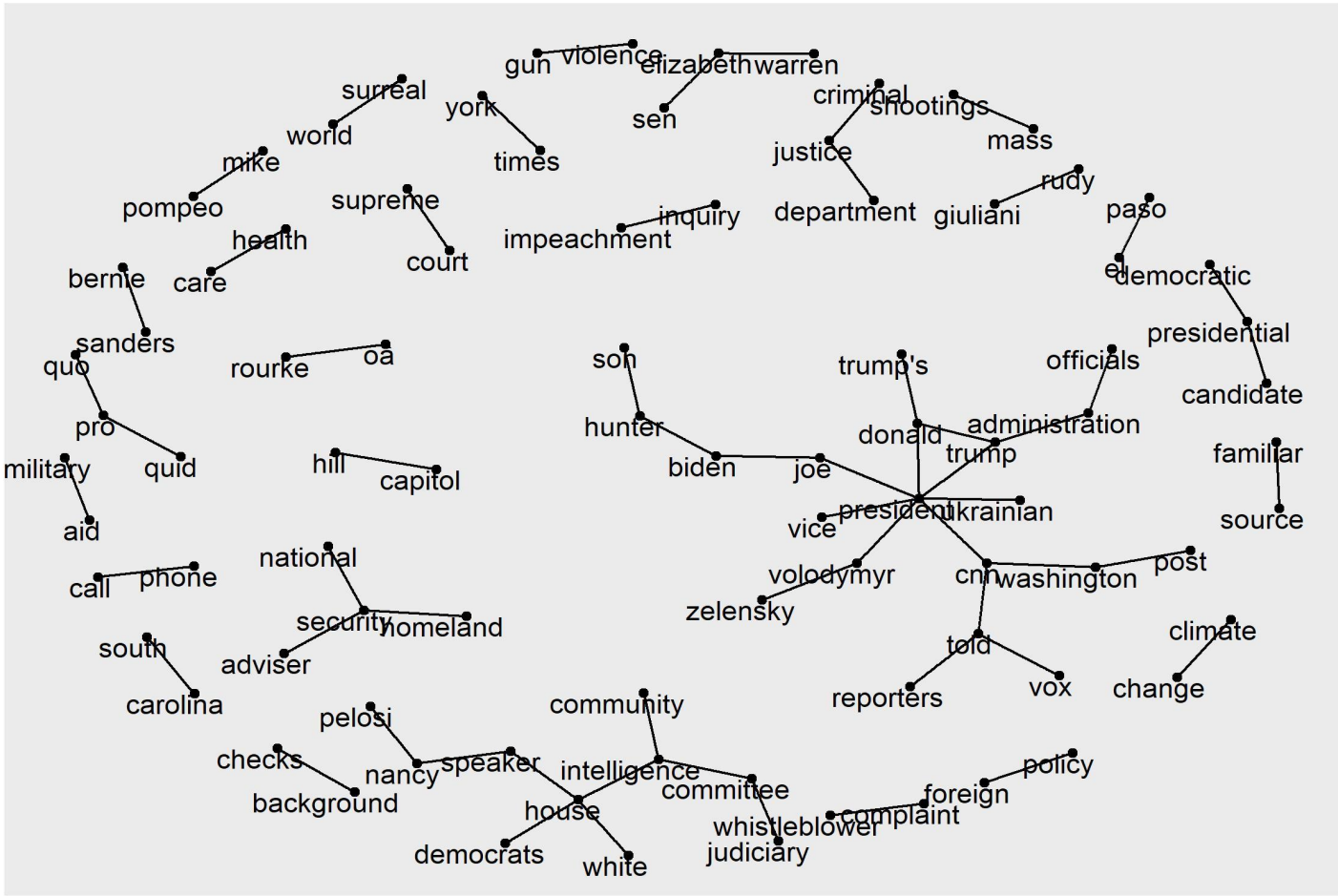

Left and right wing news outlets report on different topics
Aug
trump
biden
house
trump's
explained
democratic
democrats
sanders
joe
2020
report
gun
plan
white
dem
0
200
400
600
Sep
trump
impeachment
ukraine
biden
house
gop
trump's
debate
democratic
democrats
whistleblower
joe
dems
warren
2020
0
200
400
600
Oct
trump
impeachment
ukraine
biden
house
gop
trump's
democratic
democrats
sanders
warren
inquiry
2020
0
250
500
750
Nov
trump
impeachment
ukraine
biden
house
trump's
democrats
warren
inquiry
2020
white
dem
testimony
campaign
donald
0
100
200


5 most popular people on each news site running for president in the next election
bbc
breitbart
cnn
fivethirtyeight
fox
motherjones
reuters
vox
trump
biden
warren
sanders
harris
yang
buttigieg
0
500
1000
1500
2000
3000
4000
5000
10000
2500
7500
250
750
2000
4000
6000
8000
Popular figures
count
political_lean
centre
left
right

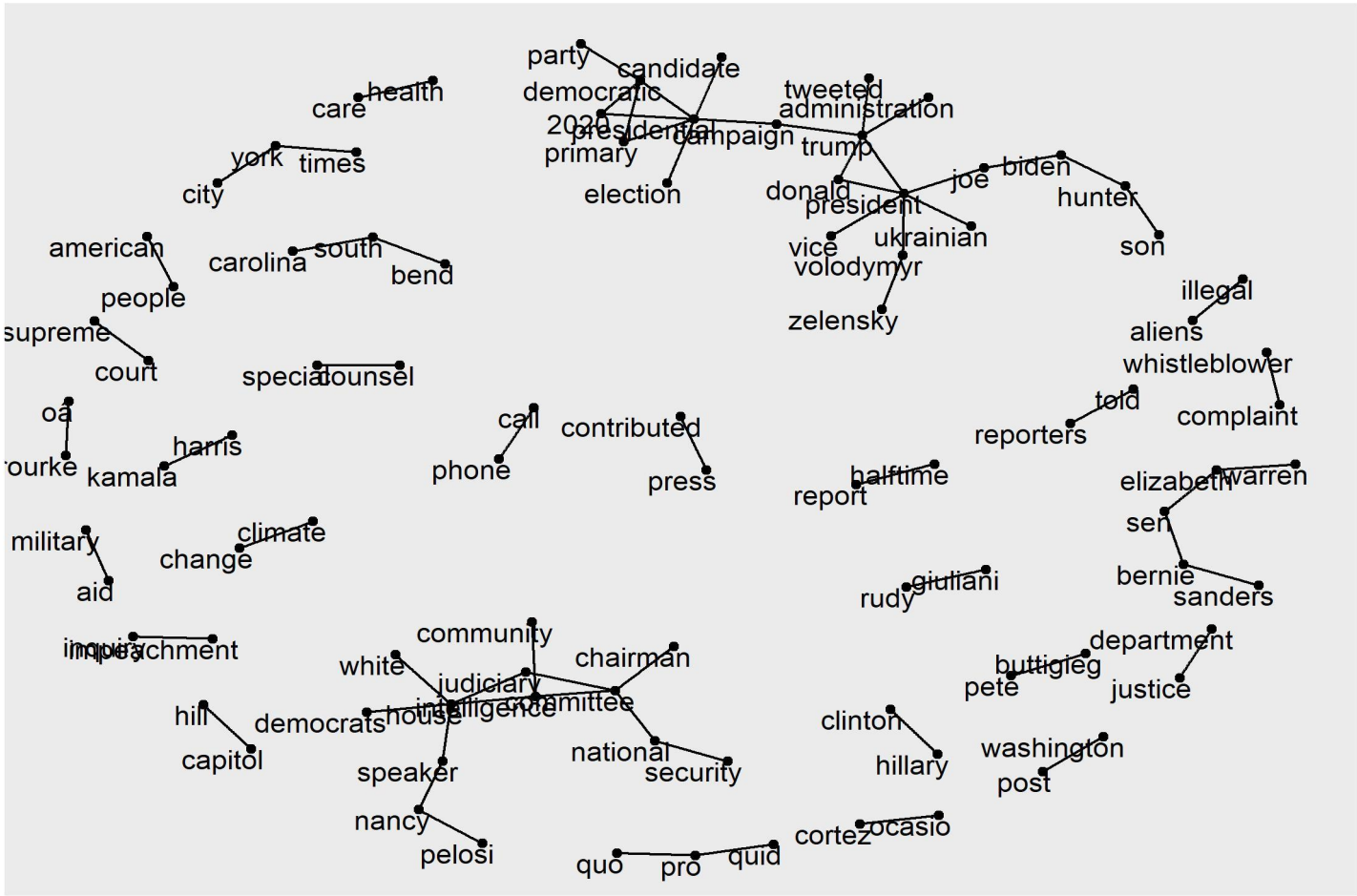
Least popular candidate on each news site, running for office in the next election

```
# A tibble: 8 x 4
# Groups:   news_site, political_lean [8]
  news_site       political_lean word   number_of_mentions
  <chr>           <chr>          <chr>                <int>
1 fox             right          steyer                 128
2 fivethirtyeight left           yang                   114
3 vox             left           steyer                 112
4 cnn             left           steyer                  84
5 breitbart       right          steyer                  45
6 reuters         centre         steyer                  28
7 bbc             centre         steyer                   7
8 motherjones     left           yang                     4
```

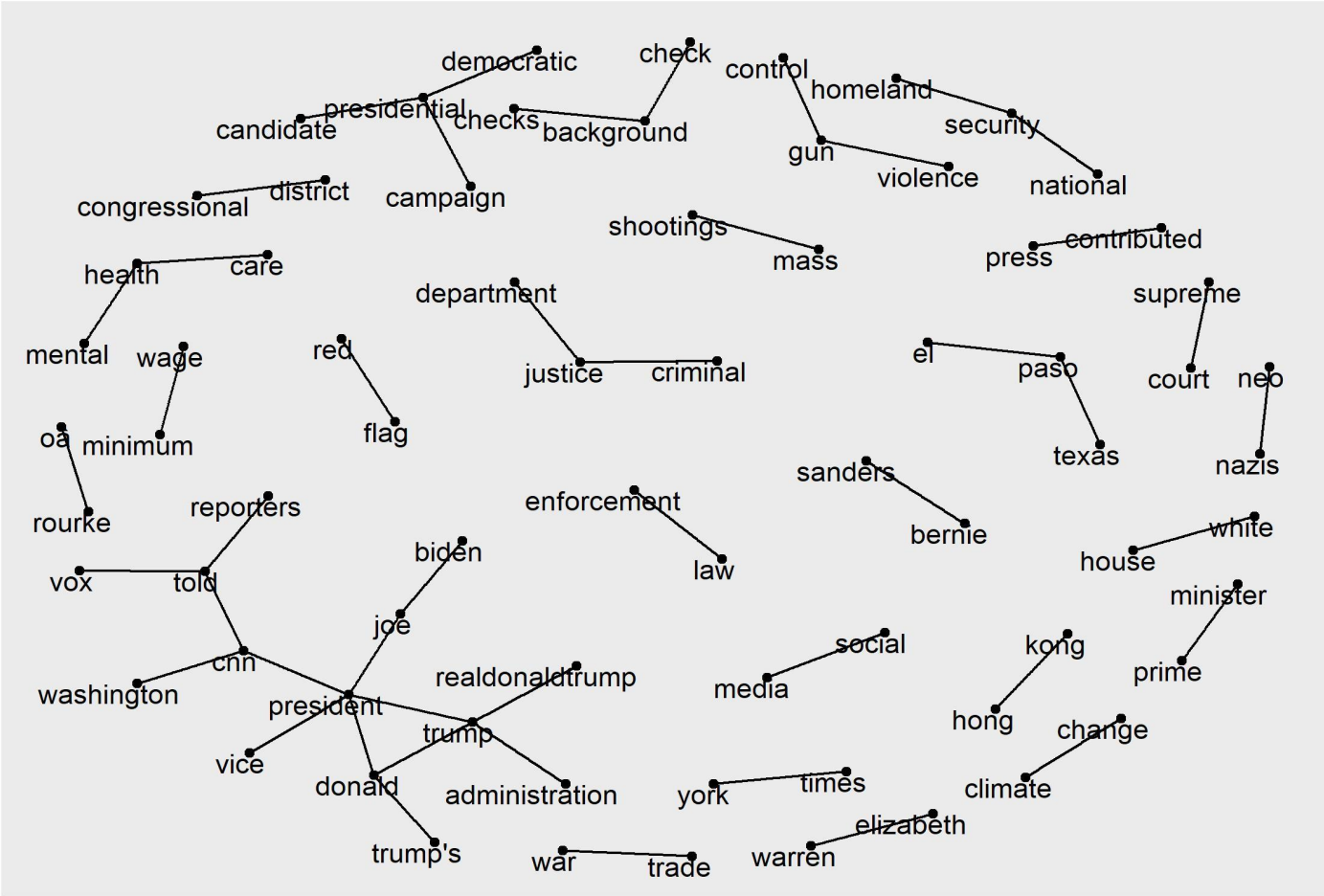
Most common bigrams in left leaning news sites



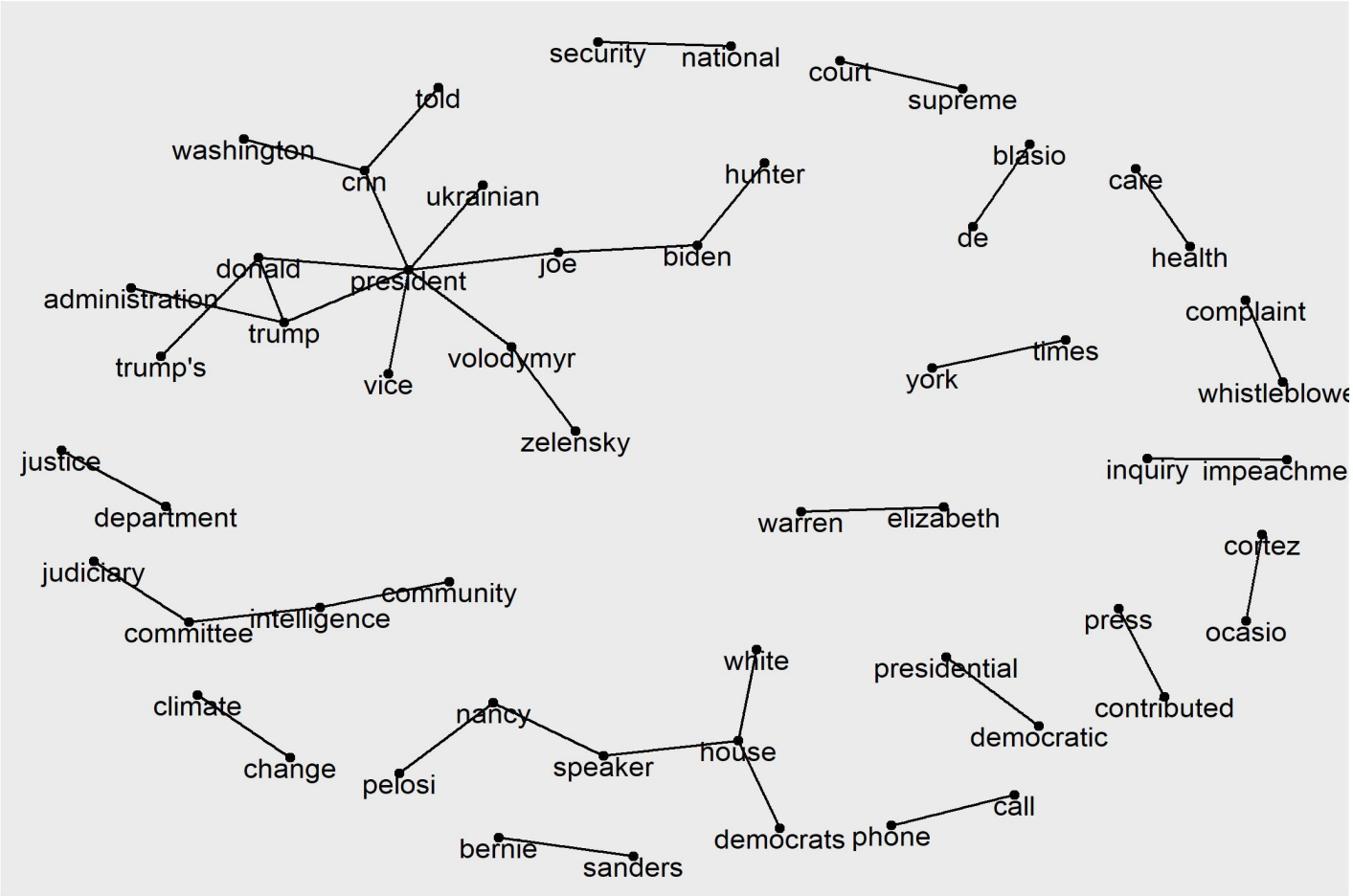
Most common bigrams in right leaning news sites



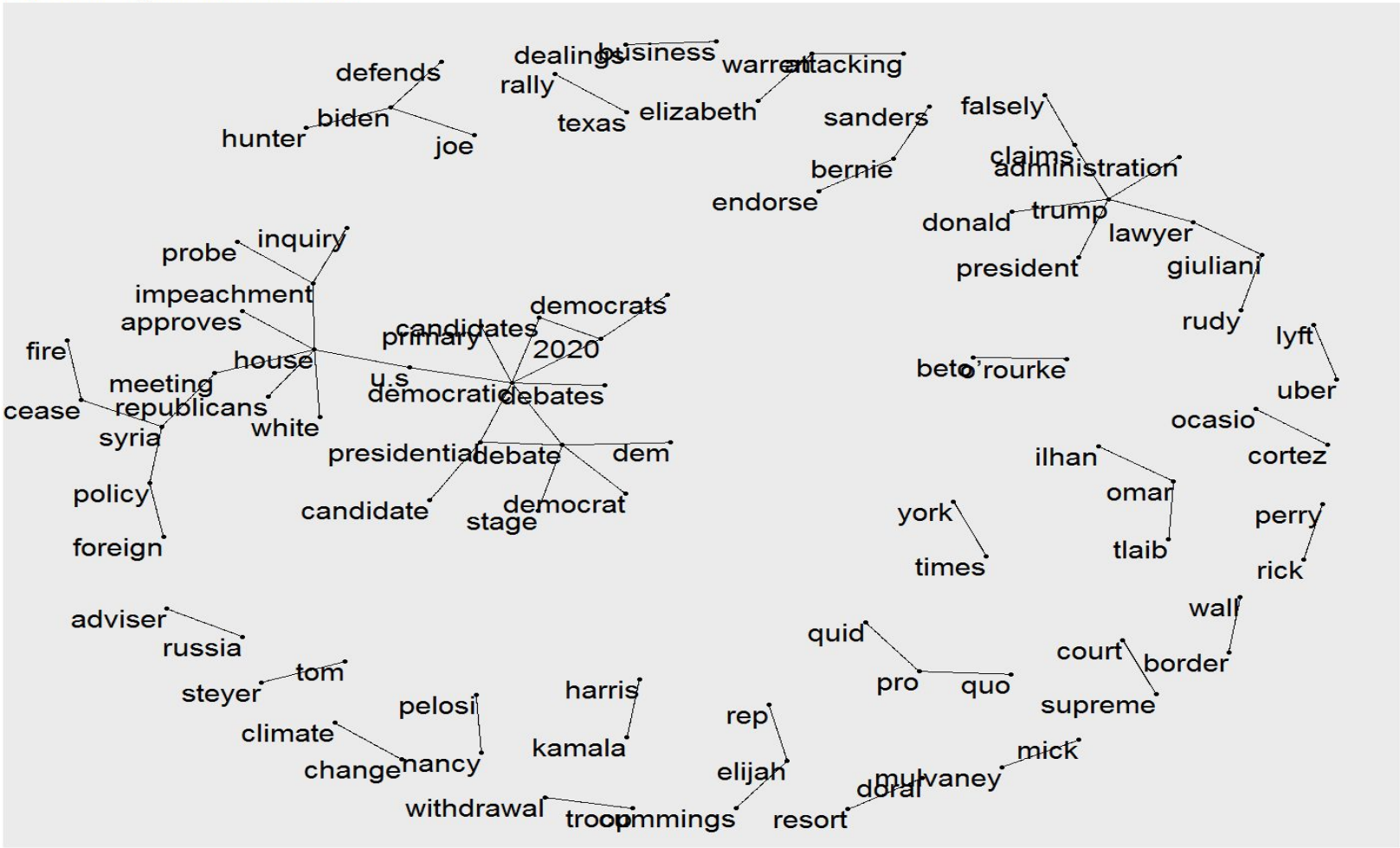
Most common bigrams in news sites in the month of August



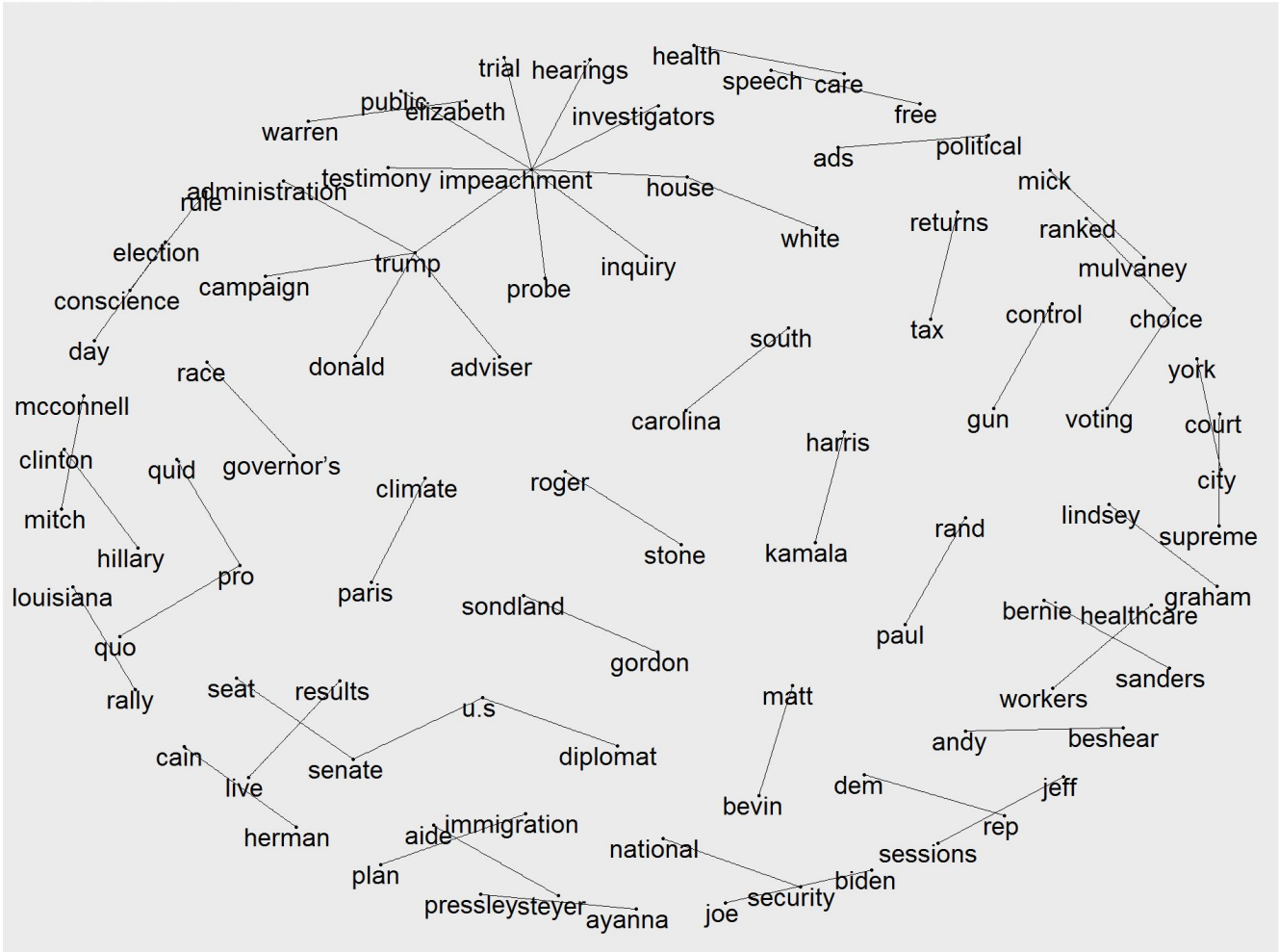
Most common bigrams in news sites in the month of September



Most Common Bigrams - 15th to 17th October

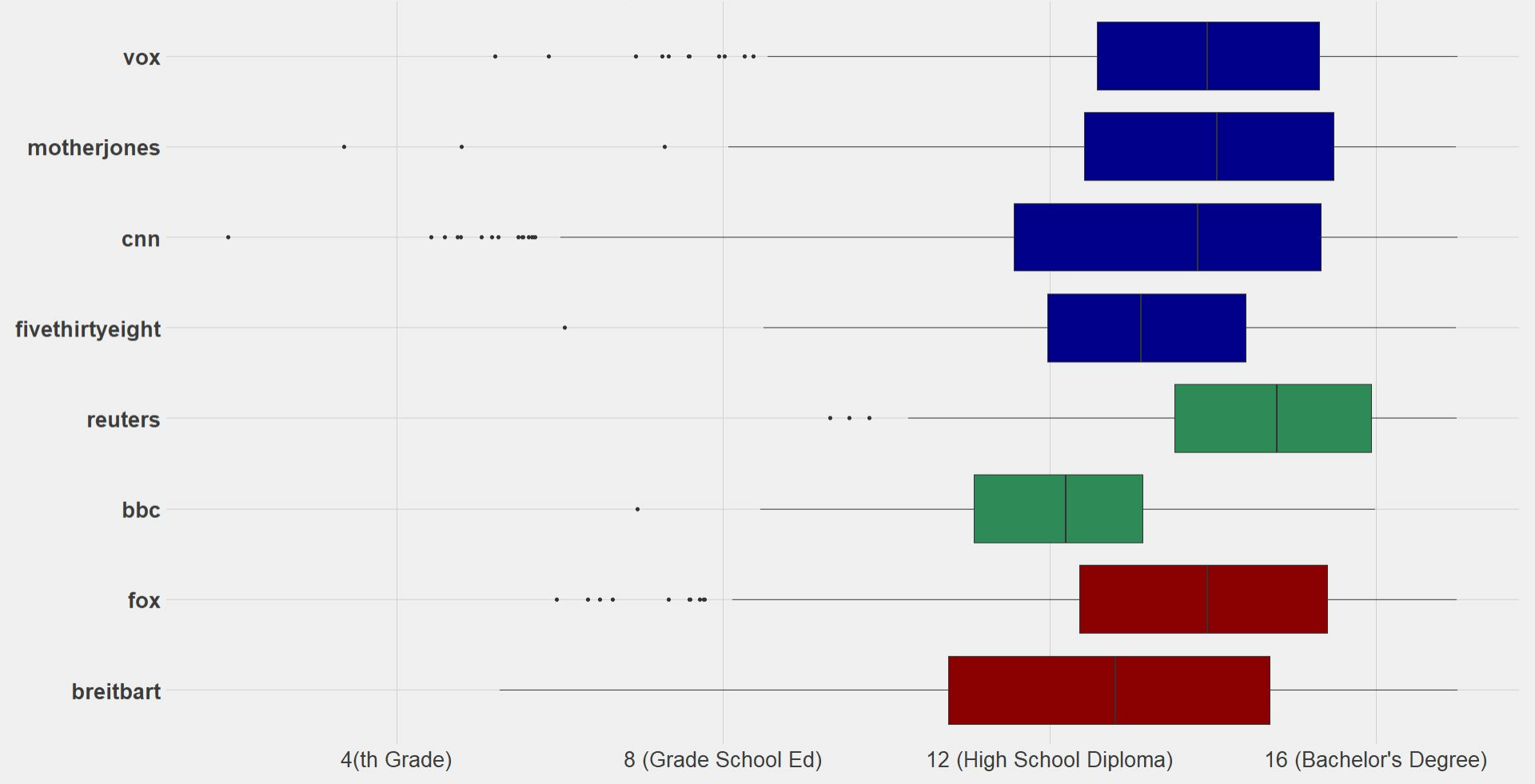


Most Common Bigrams - 5th to 7th November



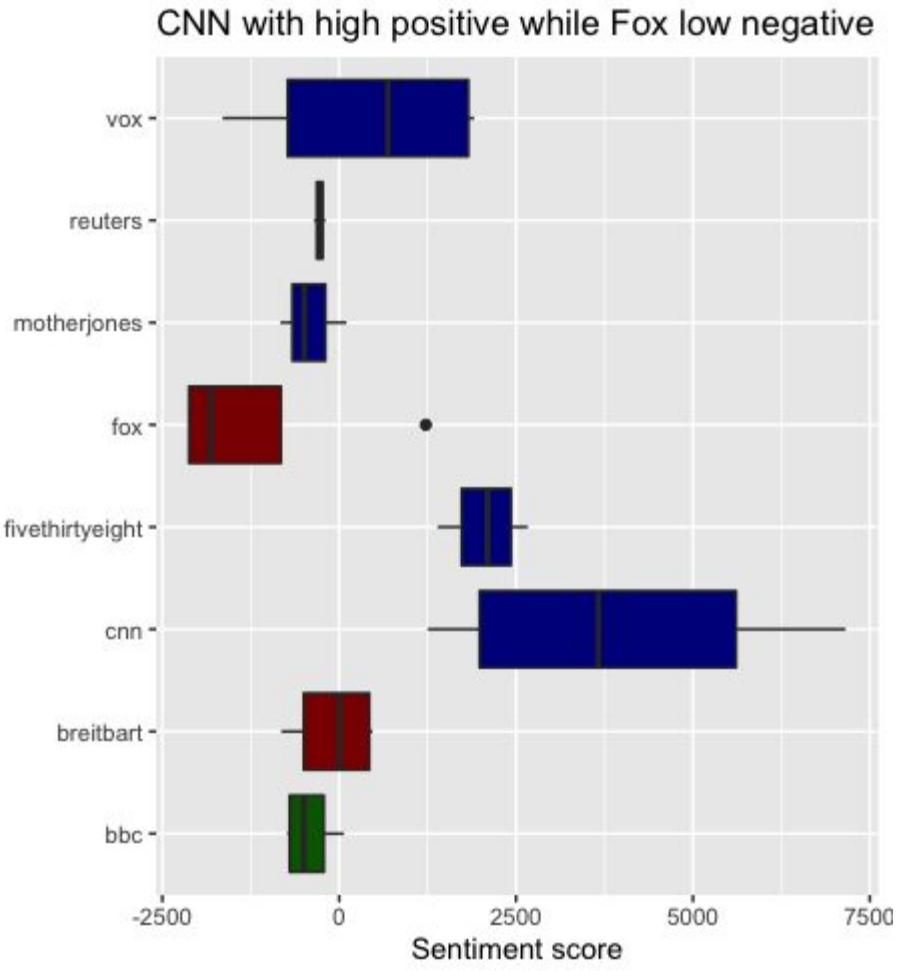
Article Reading Levels by Politics and Sites

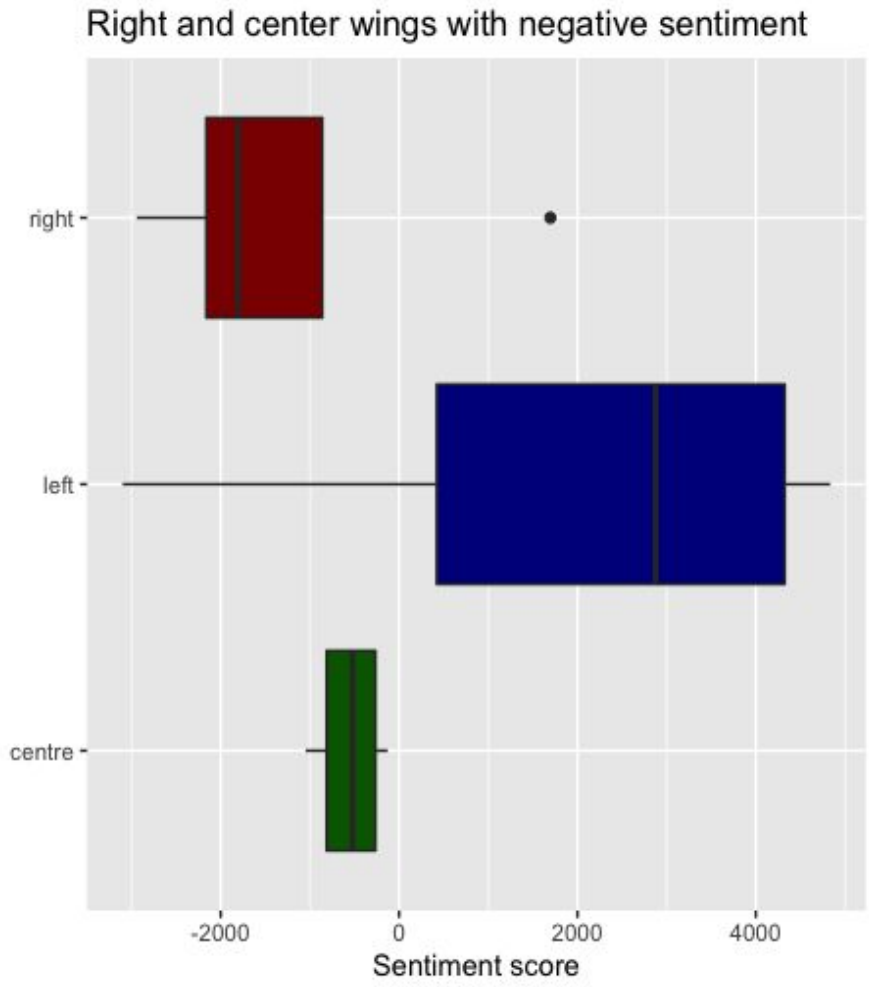
Based on Flesch-Kincaid scale. Excluding scores over 17



like
false
support
well
right
intelligence
complaint
work
top
vice
good
undefined
issue
supreme
clear
issues
better
lead
win
attack
conservative
concerns
allegations
enough
great
strict
problem
corruption
criminal
important
inactive
approval
oversight
protection
progressive
reform
scandal
worked
hard
wrong
threat
strong
death
won
assault
crime
best
attacks
popular
concern
favor
criticized
super
denied
supported
supporting
controversy
effectively
supports
successful
grand
warning
fake
broke
suicide

paradoxically
fuss
syndrome
courageously
congratulate
molestation
obliterate
impeccable
brash
overstated
lambaste
bearish
overweight
impoverished
oblivious
favored
indigent
deviation
authoritative
godsend
delight
willing
delicious
booming
divisiveness
demonize
falsehood
authoritarian
crush
bothers
celebrate
disinterest
undermined
demean
inopportune
perfectly
beautiful
slower
crass
topple
ample
liable
crashes
slowest
crashing
hectic
heedless
deadlock
sketchy
glamorous
fabulously
decency
sacrificed

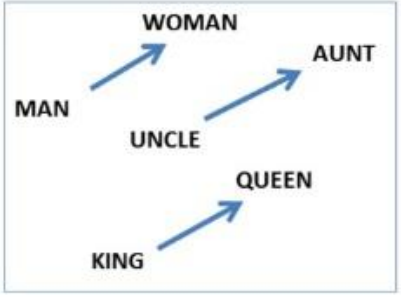

CNN with high positive while Fox low negative
vox
reuters
motherjones
fox
fivethirtyeight
cnn
breitbart
bbc
-2500
0
2500
5000
7500
Sentiment score

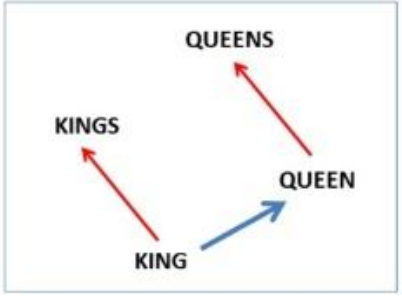

Right and center wings with negative sentiment
right
left
centre
-2000
0
2000
4000
Sentiment score

Fasttext

- Fasttext is a NLP library that learns representations of words in text through “ngrams”.
  - ngrams is the breaking down of words into subsegments and learning those. For example, the word “where” could be broken into <wh, whe, her, ere, re>.
- These embeddings encode the meaning of the word, which are used in classification

vec(“man”) – vec(“king”) + vec(“woman”) = vec(“queen”)

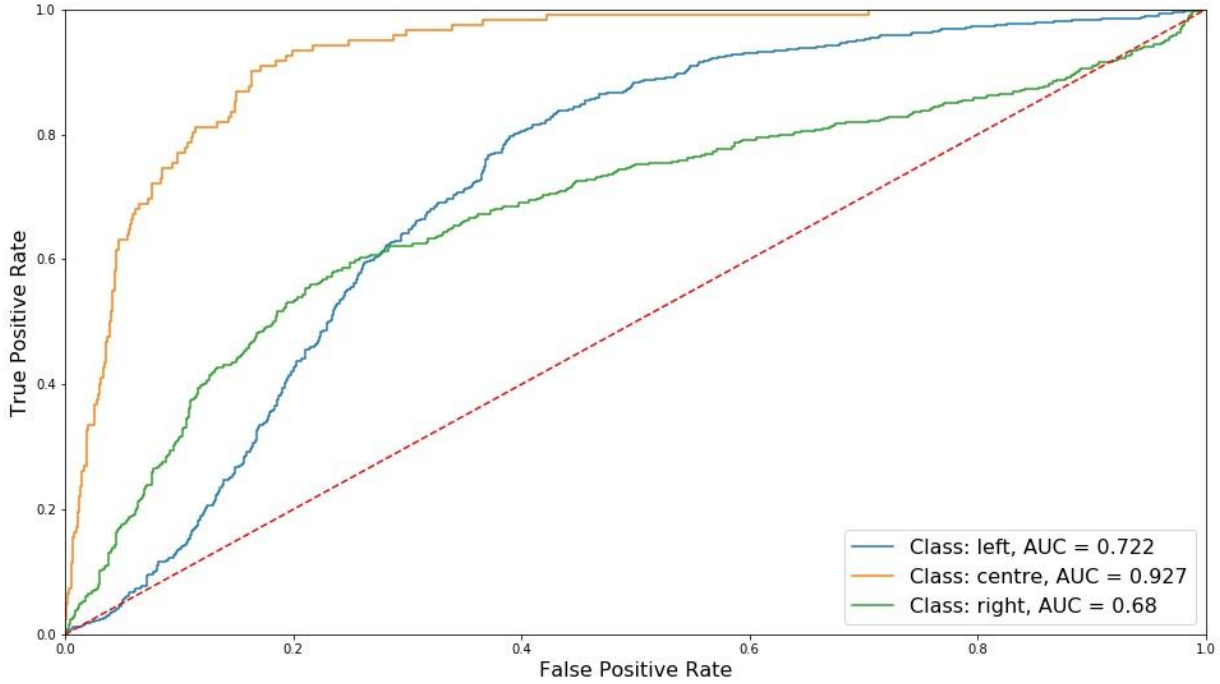




Fasttext Results

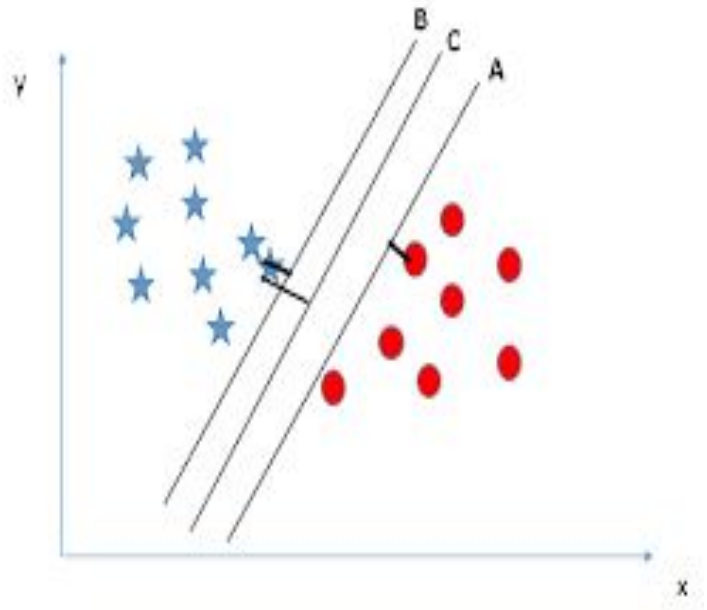
- Found that best model used ngram=2 from cross validation.
- Headline Accuracy: 74%
- By class:
  - Center: 99% precision, 80% recall, 88% F1 Score
  - Left: 67% precision, 97% recall, 79% F1 Score
  - Right: 91% precision, 44% recall, 60% F1 Score

Fasttext ROC



SVM

- Support Vector Machine, abbreviated as SVM is a Machine Learning algorithm that is widely used for classification.
- The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space (N is the no. of features) that distinctly classify the data points.



SVM Model

- The SVM model classifies the words in the text into either left or right political lean.
- The algorithm classifies left as ‘0’ and right as ‘1’.

SVM Results

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 871  55
         1  67 682

               Accuracy : 0.9272
                 95% CI : (0.9137, 0.9392)
    No Information Rate : 0.56
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8525

 Mcnemar's Test P-Value : 0.3193

            Sensitivity : 0.9286
            Specificity : 0.9254
         Pos Pred Value : 0.9406
         Neg Pred Value : 0.9105
             Prevalence : 0.5600
         Detection Rate : 0.5200
   Detection Prevalence : 0.5528
      Balanced Accuracy : 0.9270

       'Positive' Class : 0
```

Thank you! Questions?