Sentiment in Reporting: A data-driven analysis of bias in 15 major publications

Summary

By using completely data-driven analysis, mostly free from analyst input, this exploration draws bias-free inference on political slant in 15 different news publications. A few broad and important insights from the analysis are:

- Articles about Republicans are more neutral than those about Democrats
- Articles about Democrats are more negative than those about Republicans
- The New York Times writes the most positive coverage about Republicans
- The New York Times is the most neutral and most consistent of all tested publications
- Fox News falls within the Mainstream Media
- Fox News is the most negative of all tested publications

There are far more conclusions and insights to be found, but the 5 above are perhaps most contrary to the general belief.

Problem Statement

Accusations of bias in new reporting have been commonplace in American political discussion for years. Rhetoric against the media during the administration of President Donald Trump has left the American public especially distrustful of published news. This problem is difficult to address; if a person believes that the news media is conspiratorially biased, then how can that person trust analysis that says otherwise? Any claims that the news is unbiased are immediately written off as proof of the analyst's bias.

This study attempts to bypass accusations of bias by removing as much analyst input as possible; the analysis is purely data-driven. Sentiment in each article is measured by the Valence Aware Dictionary for sEntiment Reasoning (VADER) algorithm, thereby removing the possibility of analyst bias in measuring the bias of news articles.

VADER

VADER was originally developed by C.J. Hutto and Eric Gilbert of the Georgia Institute of Technology. It is a rule-based model for general sentiment analysis, complete with a dictionary of slang, acronyms, and emoticons. Study of the algorithm has shown it to be as accurate as human interpreters, and in F1-score VADER actually outperforms human sentiment analysts.

The algorithm catches punctuation, capitalization, degree modifiers, polarity shift (due to conjunctions), and polarity negation. It is a very powerful method for sentiment analysis.

VADER returns judgements on both polarity and intensity; it determines whether the sentiment is positive or negative and also *how* positive or *how* negative. In Python, VADER can be implemented from the vaderSentiment or NLTK packages.

Data Gathering and Preparation

Data was downloaded from https://www.kaggle.com/snapcrack/all-the-news; there are three .csv files, with 142,570 articles news articles, published between 2011 and 2017, with most of the articles falling between January 2016 and July 2017. Publications represented in the dataset are:

- Breitbart
- Atlantic
- Business Insider
- Buzzfeed News
- CNN
- Fox News
- Guardian
- National Review
- New York Post
- New York Times
- NPR
- Reuters
- Talking Points Memo
- Vox
- Washington Post

The data contains the text of the articles, titles of the articles, publication, author, and date, along with a unique ID.

The text of the articles appeared as written; VADER uses information such as capitalization and punctuation, so there was minimal preprocessing of the text.

The text, as printed, was fed through a SentimentIntensityAnalyzer from vaderSentiment, which returns four measurements:

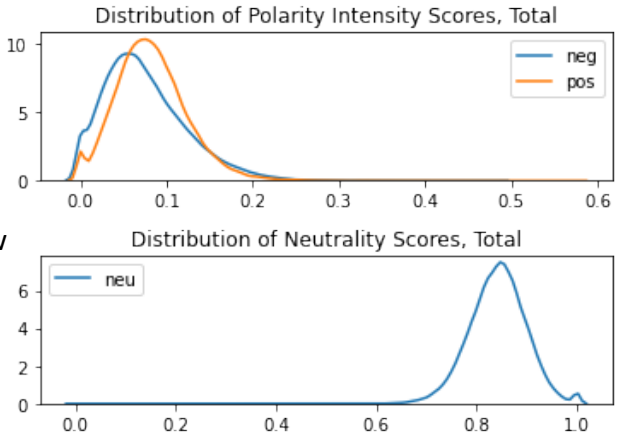
- 'neg' – Negative sentiment intensity, scaled between 0 and 1
- 'pos' – Positive sentiment intensity, scaled between 0 and 1
- 'neu' – A measure of neutrality
- 'compound' – A measure of combined overall sentiment, between -1 and +1

In the compound score, negative numbers indicate a preponderance of negative sentiment, while positive numbers mean the positive sentiment is stronger.

After analyzing the sentiment of the dataset as a whole, the articles were separated by topic. Two lists of words were formed, ["Republican", "Republicans", "Trump"], and ["Democratic", "Democrats", "Democrat", "Obama", "Clinton"]. If words from the first list appeared in an article more than words in the second list, the article was judged to be about Republicans, and vice versa. If no words from either list appeared in an article, then the article was deemed "non-political" and discarded.

Overall Analysis

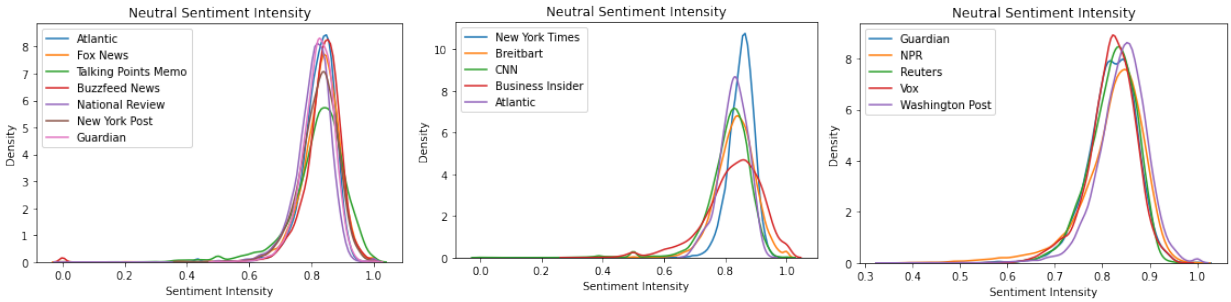
Overall, positive sentiment was slightly more prevalent than negative sentiment, and neutrality scores were high.



The neutrality and compound sentiment scores are the most important scores to the goals of this research and will be the focus of this report from now on.

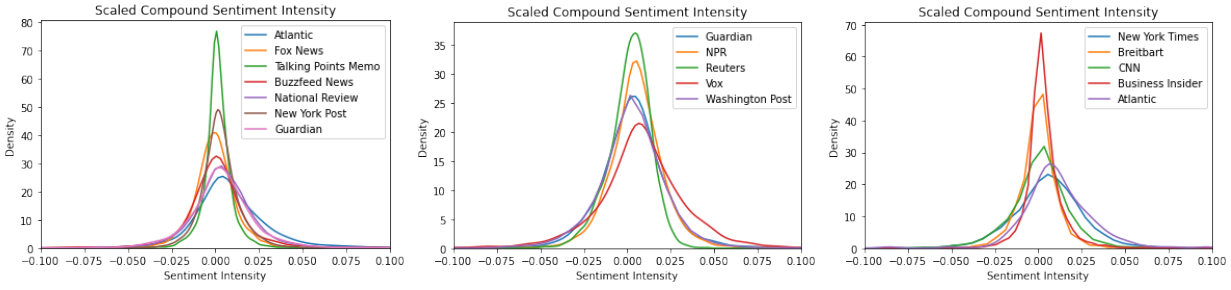
The data came in three large .csv files, and was first examined in groups according to those files.

The overall neutrality scores are quite similar with three notable exceptions. Talking Points Memo and Business Insider have unusually low neutrality peaks, indicating more variance in neutrality, while the New York Times has an unusually high peak.



The scaled compound scores have more differentiation than the neutrality scores. In these scores, Talking Points Memo has the highest peak. There are many publications with low peaks and wide tails, including Atlantic, VOX, and the New York Times.

The *meaning* of the compound score is much more abstract than the meaning of the neutrality score. The location of the peak is easier to interpret than the height of the peak; peaks farther to the right indicate generally more positive reporting.



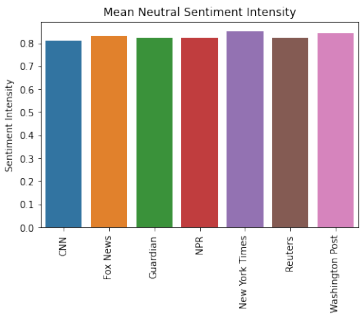
Grouping Publications

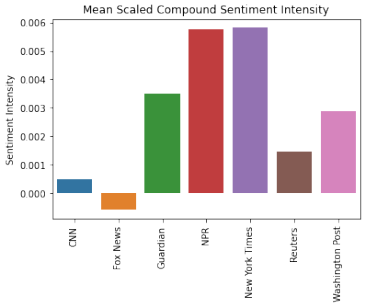
The drawback to the analysis above is that which publications are compared is based on which file the data was in, and not on any reason to compare them. Thus, the next step in overall analysis was separating the publications into groups for more appropriate comparison. Deciding on these groups was the only specifically analyst-defined area in this research, but it does not affect the outcome; the analyst only decided which lines to plot on which graph. The three groups were:

- Traditional News – Older, established media publications
    - New York Times
    - CNN
    - Fox News
    - Guardian
    - NPR
    - Reuters
    - Washington Post
- Fun News – Publications that sometimes run "tabloid"-type stories
    - Buzzfeed News
    - Talking Points Memo
    - New York Post
- Blog-style News – Publications that include significant amounts of analysis, explanation, and editorial content
    - Atlantic
    - Breitbart
    - National Review
    - Vox

Traditional News

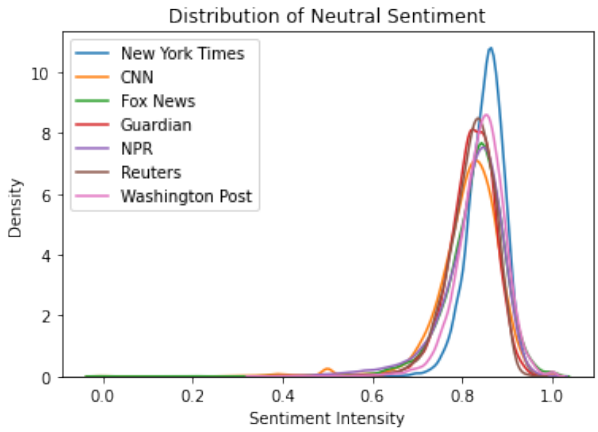
The traditional news sources all had mean neutral intensities quite near 0.8; the New York Times has a mean neutrality slightly higher than the others.

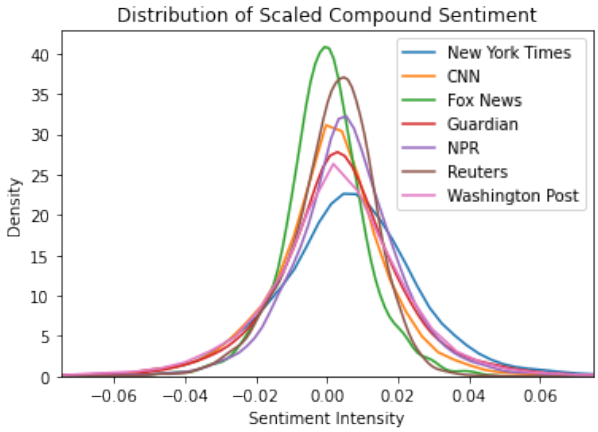




Mean compound sentiment in the traditional news sources was quite diverse. Fox News actually has a negative mean compound sentiment! It is the only publication with a negative mean compound sentiment.

The distribution of neutral sentiment among the traditional news sources shows that they are mostly the same, except the New York Times has a very high peak that is to the right of other peaks; this indicates more neutrality as well as more consistency.

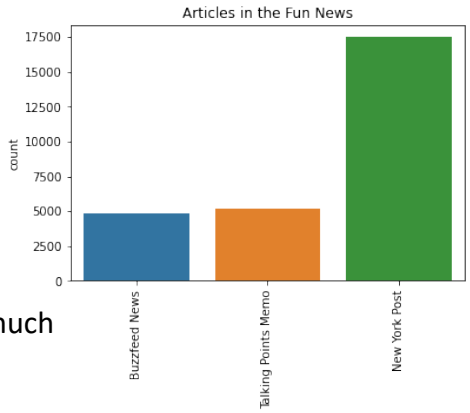




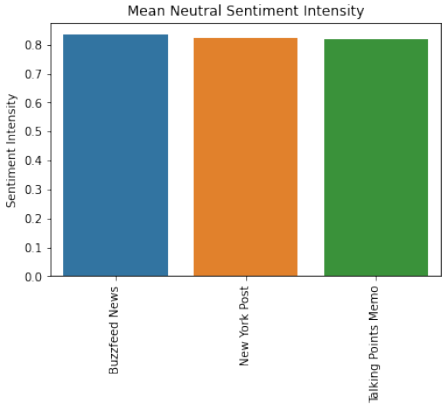
The distribution of compound sentiment scores shows Fox News with a relatively high peak that is far to the left of the others. Again, Fox News is the only publication to have a negative mean compound sentiment.

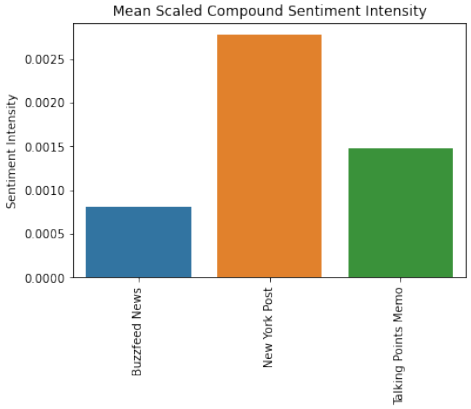
Fun News

The Fun News category is not balance; there are more than three times as many articles from the New York Post as there are from Buzzfeed News or Talking Points Memo.

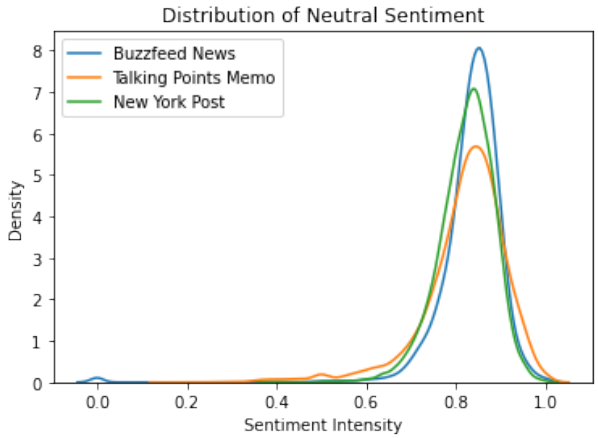


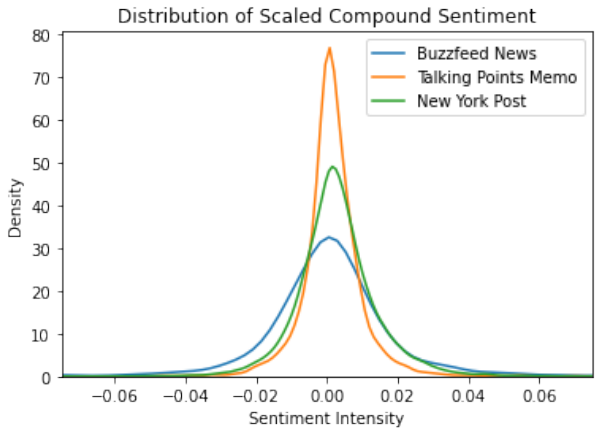
There was little difference in mean neutrality score between the three publications. They were all quite close to 0.8; but there was a wide difference in the compound sentiment score. The New York Post has a much more positive compound score than Buzzfeed News or Talking Points Memo.



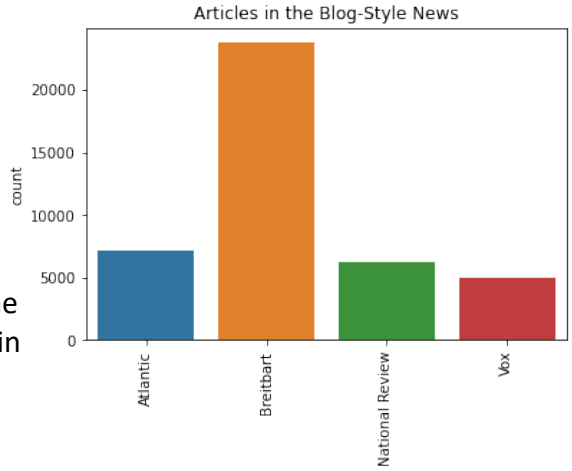


The distributions of neutral and compound sentiment show similar patterns.



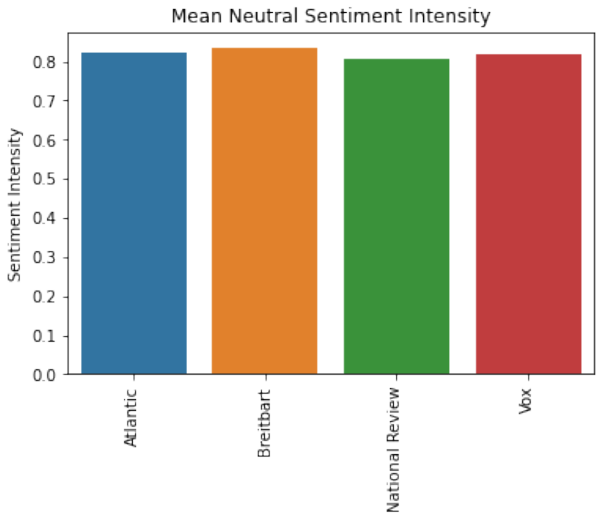


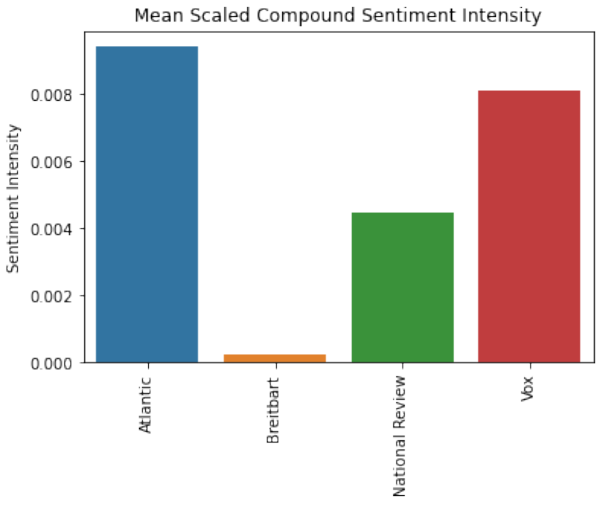
Blog-Style News



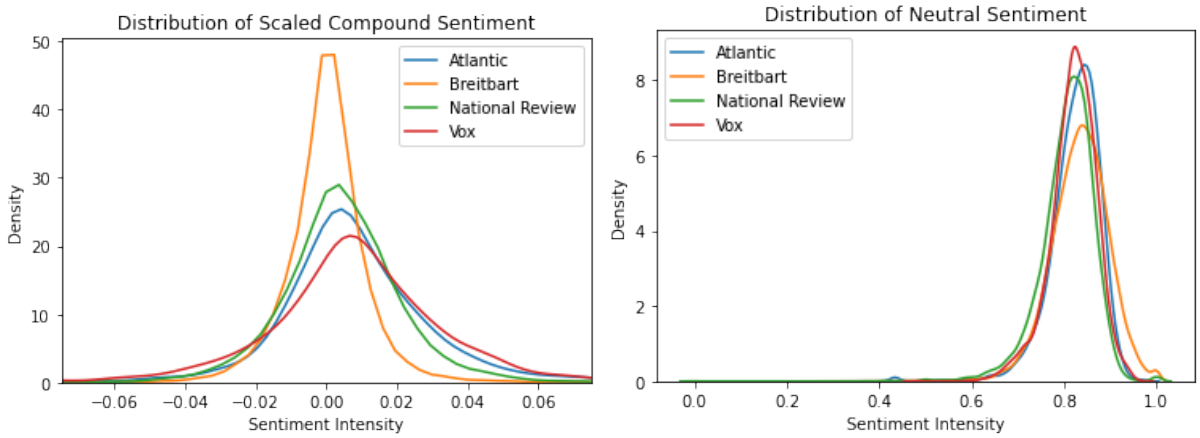
This is the least balanced of all three groups. Breitbart has the most articles of any publication; more than 4 times as many as Vox.

There is a visible difference between the four publications; Breitbart has a slightly higher mean and National Review slightly lower, with Vox and Atlantic in the middle. However, there is quite a pronounced difference in mean compound sentiment score; Breitbart's mean compound
Score is much lower than the other three. This makes it one of the most negative publications.



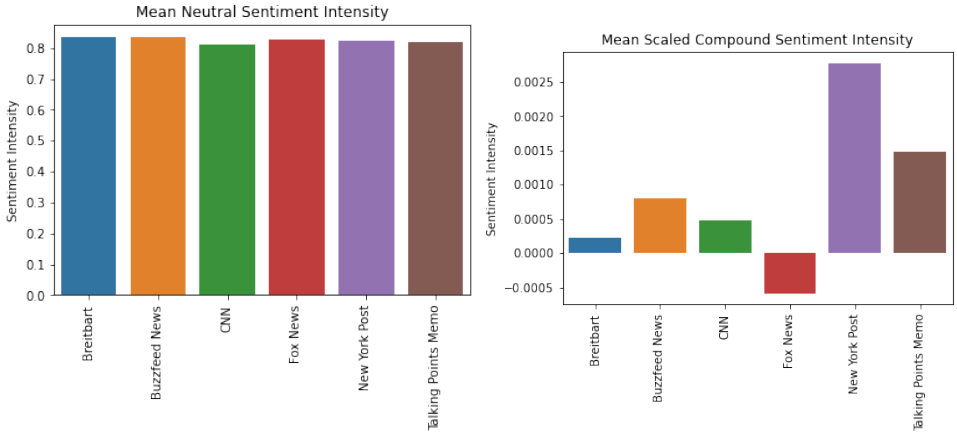


The relationship between the four publications is more apparent in distribution plots.

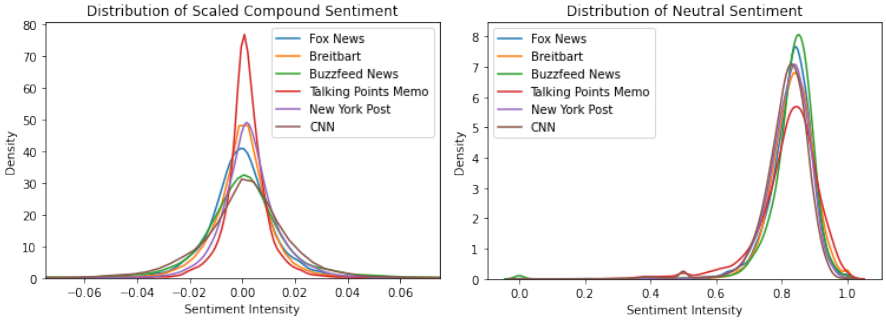


Final Group

In the group analysis, there were several publications that had especially low compound sentiment scores. Those were grouped for a final comparison.

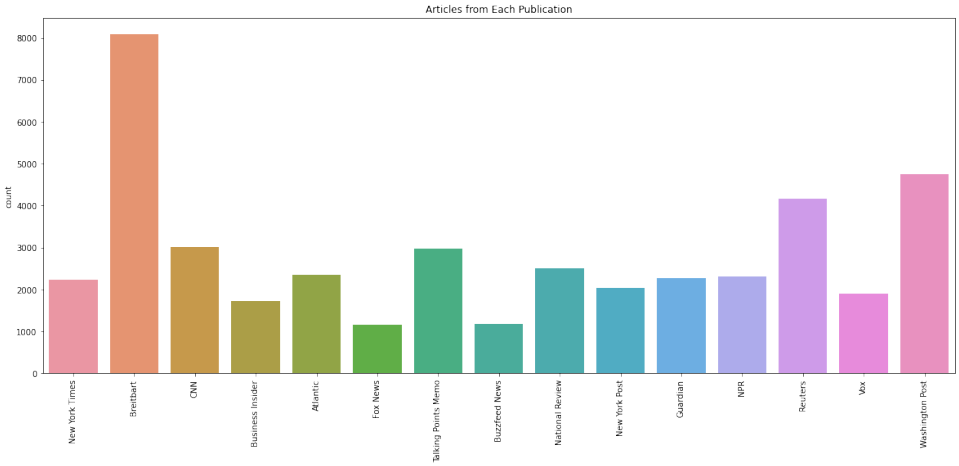


The mean neutrality scores in this group were once again quite similar, but the mean compound scores still show significant differences. Fox News is the only publication where the negative reporting outweighed the positive. However, Breitbart is the second-lowest, with a compound score quite near zero.

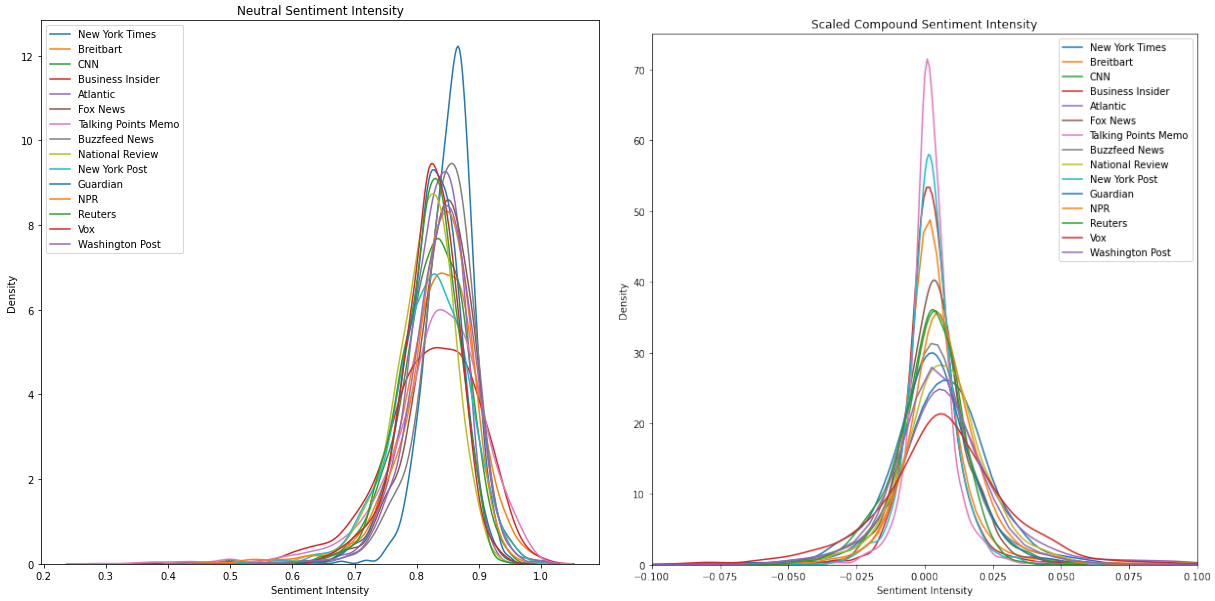


Articles about Republicans

There were 42,689 articles predominantly about Republicans.



When the distribution of neutrality scores is plotted for every publication, most of the publications are quite similar. There are four publications with relatively low peaks and fat tails, indicating less consistent neutrality; sometimes they are more neutral, and sometimes less. These inconsistent publishers are, from lowest to highest: Business Insider, Talking Points Memo, New York Post, and Breitbart. The distribution plot also shows the New York Times as having by far the highest and farthest-right peak, indicating the most neutrality and most consistency.



The distrbution of compound scores is interesting; the lower the peak, the futher right it is. Low peaks indicate inconsistency, and area to the right indicates positive sentiment; although Vox has the least consistent compound score, it also has the most positive compound sentiment.

Since the distributions at times appeared skewed, the median rather than the mean was used to compare the "average" sentiment in the table to the right.

publication	neg	pos	neu	scaled_compound
Talking Points Memo	0.043791	0.058905	0.837215	0.001747
Business Insider	0.044385	0.056351	0.829800	0.001984
New York Post	0.054500	0.067840	0.824889	0.002048
Guardian	0.068061	0.075172	0.826952	0.002048
Breitbart	0.053059	0.070294	0.839412	0.002087
Reuters	0.058712	0.066636	0.822550	0.002321
Fox News	0.048094	0.063364	0.846683	0.003412
Buzzfeed News	0.053629	0.067396	0.850700	0.003484
CNN	0.055463	0.070591	0.829976	0.003729
Washington Post	0.059066	0.072366	0.842964	0.004345
NPR	0.048256	0.068800	0.838362	0.005846
Vox	0.066083	0.081571	0.823695	0.006533
National Review	0.068867	0.089765	0.816939	0.006782
Atlantic	0.060452	0.078858	0.833343	0.006955
New York Times	0.054853	0.072959	0.856793	0.007725

The analysis continued with the construction of a Linear Model and then an Analysis of Variance (ANOVA) to see if the publications do in fact have significantly different sentiment in their reporting on Republicans. In addition to showing that the data fulfilled the required assumptions for an ANOVA, they were also examined for stationarity; the articles were published over time, and so if there was any trend or change in sentiment from one time period to the next, then the linear model would need to include Date as a variable. The data was shown to be stationary, and so linear model included publication as the only independent variable.

The statistical test found a highly significant difference between neutrality in articles about Republicans, based on the publisher ($p = 0.00$). The only publication that was found to have a non-significant coefficient in the linear model was NPR; this means that NPR has the average neutrality (approx. 0.83). Breitbart, Buzzfeed News, Fox News, Washington Post, and New York Times had positive coefficients, indicating higher-than-average neutrality; Business Insider, CNN, Guardian, National Review, New York Post, Reuters, Talking Points Memo, and Vox had negative coefficients, indicating lower-than-average neutrality. The linear model summary table is on the next page.

The same analysis was performed for the compound sentiment score; this time, all publications had significantly non-zero coefficients. All coefficients were also positive, meaning that all publications typically had slightly positive sentiment in articles about Republicans. Interestingly, the New York Times had the highest coefficient, indicating most positive coverage, followed by National Review.

Lowest Compound Sentiment		**Highest Compound Sentiment**	
Publication	Coefficient	Publication	Coefficient
Reuters	0.001	New York Times	0.008
Guardian	0.0016	National Review	0.0074
Talking Points Memo	0.002	Vox (tie)	0.0069
New York Post	0.0022	NPR (tie)	0.0069
Business Insider	0.0027	Washington Post	0.0054

OLS Regression Results

Dep. Variable:	neu	**R-squared:**	0.035
Model:	OLS	**Adj. R-squared:**	0.035
Method:	Least Squares	**F-statistic:**	106.7
Date:	Fri, 08 Jan 2021	**Prob (F-statistic):**	5.61e-305
Time:	11:35:58	**Log-Likelihood:**	59962.
No. Observations:	40615	**AIC:**	-1.199e+05
Df Residuals:	40600	**BIC:**	-1.198e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.8278	0.001	726.657	0.000	0.826	0.830
C(publication)[T.Breitbart]	0.0083	0.001	6.417	0.000	0.006	0.011
C(publication)[T.Business Insider]	-0.0047	0.002	-2.690	0.007	-0.008	-0.001
C(publication)[T.Buzzfeed News]	0.0172	0.002	8.690	0.000	0.013	0.021
C(publication)[T.CNN]	-0.0040	0.002	-2.590	0.010	-0.007	-0.001
C(publication)[T.Fox News]	0.0130	0.002	6.530	0.000	0.009	0.017
C(publication)[T.Guardian]	-0.0042	0.002	-2.570	0.010	-0.007	-0.001
C(publication)[T.NPR]	0.0009	0.002	0.545	0.586	-0.002	0.004
C(publication)[T.National Review]	-0.0174	0.002	-10.966	0.000	-0.021	-0.014
C(publication)[T.New York Post]	-0.0070	0.002	-4.165	0.000	-0.010	-0.004
C(publication)[T.New York Times]	0.0253	0.002	15.521	0.000	0.022	0.029
C(publication)[T.Reuters]	-0.0105	0.001	-7.363	0.000	-0.013	-0.008
C(publication)[T.Talking Points Memo]	-0.0047	0.002	-2.357	0.018	-0.009	-0.001
C(publication)[T.Vox]	-0.0073	0.002	-4.263	0.000	-0.011	-0.004
C(publication)[T.Washington Post]	0.0109	0.001	7.843	0.000	0.008	0.014

Omnibus:	7576.869	**Durbin-Watson:**	1.940
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	25801.904
Skew:	-0.934	**Prob(JB):**	0.00
Kurtosis:	6.429	**Cond. No.**	17.4

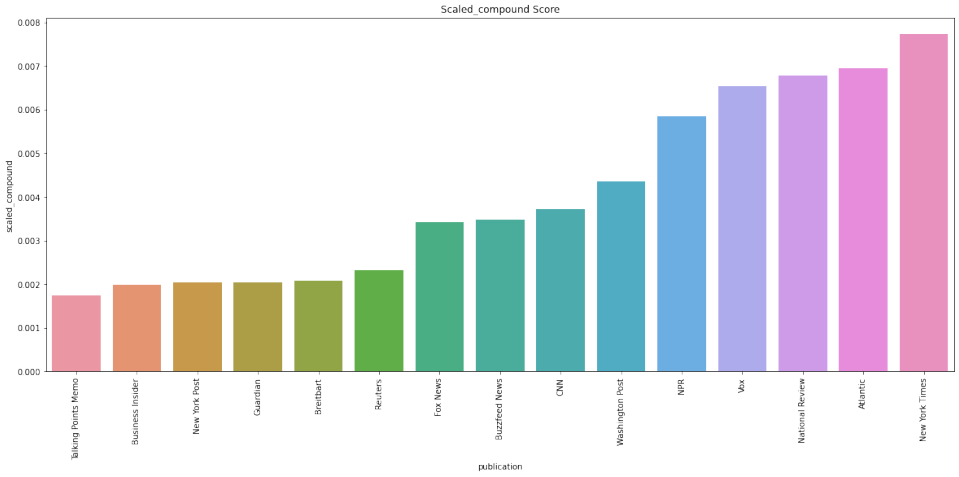
The final step in analysis of articles about Republicans was a Tukey test for Honest Significant Differences; this test compares two publications directly to one another. If the result is significant, then the second publication has significantly higher neutrality than the first.

The high number of comparisons being performed, combined with software restrictions to p-values of 0.001 or greater, meant that the lower significance level that could be achieved here was $\alpha = 0.21$, which is quite high. The full output of the Tukey Test on neutrality is in Appendix A. The full output of the Tukey Test on compound sentiment is in Appendix B.

A few interesting notes:

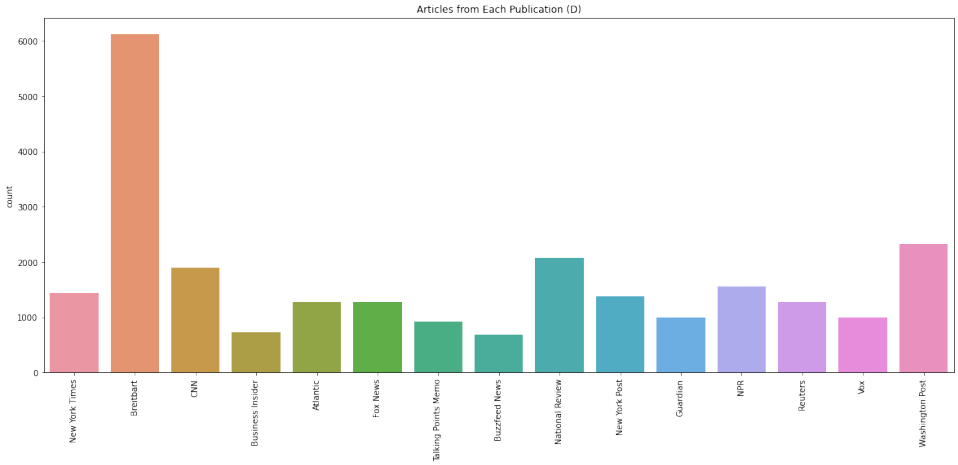
- CNN is not significantly different from Fox News, Breitbart, or the Washington Post
- Breitbart has a lower compound sentiment score than the New York Times, Washington Post, and NPR
- Fox News also has a lower compound score than the New York Times or NPR
- The New York Times has a higher compound score than Reuters, Washington Post, and Talking Points Memo

Finally, here is a barplot of median compound sentiment, so publications can be easily compared directly.

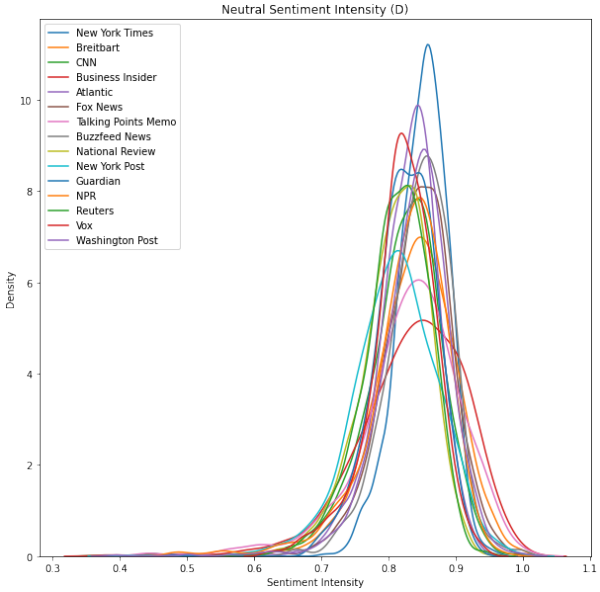


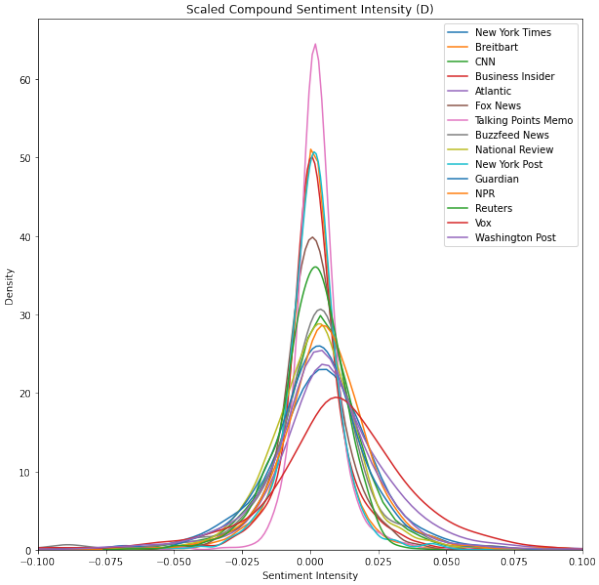
Articles about Democrats

There were 24,984 articles about Democrats.



When the distributions of neutrality scores of all publications are plotted together, they again form a major cluster with a few outlying publishers. Once again, the New York Times has the highest and farthest-right peak, indicating most neutral and consistent reporting, and Business Insider has the lowest peak and fattest tails, indicating inconsistent neutrality.





In compound score, the lower peaks again trended to the right, with the lowest and most positive being Vox.

The table on the right shows the median sentiment intensity for all publications. National Review had the highest negative sentiment, Vox had the highest positive and compound sentiments, and New York Times had the highest neutrality score.

publication	neg	pos	neu	scaled_compound
Fox News	0.052991	0.059065	0.846038	0.000976
Breitbart	0.054789	0.064182	0.842442	0.001021
Business Insider	0.046911	0.055267	0.843566	0.001101
New York Post	0.062033	0.069553	0.812900	0.001109
Reuters	0.061882	0.067750	0.815727	0.001273
National Review	0.074111	0.081290	0.815439	0.002064
Talking Points Memo	0.041375	0.060625	0.837812	0.002180
Guardian	0.067922	0.074818	0.827318	0.002767
Buzzfeed News	0.053317	0.066959	0.850164	0.002850
CNN	0.056224	0.070250	0.828718	0.003403
Washington Post	0.059214	0.070612	0.844415	0.003630
New York Times	0.058855	0.070381	0.853903	0.004402
NPR	0.050598	0.067815	0.837319	0.005458
Atlantic	0.058387	0.080612	0.833536	0.006964
Vox	0.060315	0.086095	0.823400	0.010673

As with the articles about Republicans, the analysis continued with a Linear Model and an ANOVA. The data was shown to fulfill the necessary assumptions of an ANOVA and was also checked for stationarity over time. The New York Times was remarkable consistent in their reporting between 2016 and 2017.

The linear model was constructed with publication as the only independent variable; the ANOVA found publication to have a highly significant effect on sentiment (p=0.00). This time, Guardian, Talking Points Memo, and NPR were found to have insignificant coefficients, meaning that all three publications had neutrality scores indistinguishable from the industry average (approx. 0.83). The publications with negative coefficients, indicating below-average neutrality, were CNN, National Review, New York Post, Reuters, and Vox. The publications with positive coefficients, higher-than-average neutrality, were Breitbart, Business Insider, Buzzfeed News, Fox News, New York Times, and Washington Post. The linear model summary table is on the next page.

Publication was also found to have a highly significant effect on compound sentiment scores, and once again all estimated coefficients were positive; however, there were three relatively high p-values and one extremely high p-value, meaning that the coefficients were indistinguishable from zero. Buzzfeed News, Fox News, and the New York Post had compound sentiment coefficients near zero (p= 0.068, 0.08, 0.077), and Reuters had a zero-coefficient (p=0.689). This means that those publishers were definitely less positive in articles about Democrats than they were in articles about Republicans.

Lowest Compound Sentiment

Publication	Coefficient
Reuters	0.0002
Breitbart	0.0005
Fox News (tie)	0.0009
New York Post (tie)	0.0009
Buzzfeed News	0.0013

Highest Compound Sentiment

Publication	Coefficient
Vox	0.0107
Atlantic	0.0089
NPR	0.0059
New York Times	0.0037
Washington Post (tie)	0.0032
Talking Points Memo (tie)	0.0023

Not only did articles on Democrats have lower lowest-compound-scores, but their highest compound scores were also generally lower than those for Republicans. In general, Democrats get less favorable news coverage than Republicans.

OLS Regression Results

Dep. Variable:	scaled_compound	**R-squared:**	0.022
Model:	OLS	**Adj. R-squared:**	0.021
Method:	Least Squares	**F-statistic:**	37.69
Date:	Mon, 11 Jan 2021	**Prob (F-statistic):**	1.93e-102
Time:	12:16:29	**Log-Likelihood:**	61559.
No. Observations:	23954	**AIC:**	-1.231e+05
Df Residuals:	23939	**BIC:**	-1.230e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
C(publication)[Atlantic]	0.0089	0.001	17.148	0.000	0.008	0.010
C(publication)[Breitbart]	0.0005	0.000	2.056	0.040	2.29e-05	0.001
C(publication)[Business Insider]	0.0022	0.001	3.148	0.002	0.001	0.003
C(publication)[Buzzfeed News]	0.0013	0.001	1.823	0.068	-9.77e-05	0.003
C(publication)[CNN]	0.0027	0.001	5.242	0.000	0.002	0.004
C(publication)[Fox News]	0.0009	0.001	1.748	0.080	-0.000	0.002
C(publication)[Guardian]	0.0022	0.001	3.674	0.000	0.001	0.003
C(publication)[NPR]	0.0059	0.000	12.495	0.000	0.005	0.007
C(publication)[National Review]	0.0014	0.000	3.407	0.001	0.001	0.002
C(publication)[New York Post]	0.0009	0.000	1.769	0.077	-9.52e-05	0.002
C(publication)[New York Times]	0.0037	0.000	7.514	0.000	0.003	0.005
C(publication)[Reuters]	0.0002	0.001	0.400	0.689	-0.001	0.001
C(publication)[Talking Points Memo]	0.0032	0.001	4.258	0.000	0.002	0.005
C(publication)[Vox]	0.0107	0.001	18.110	0.000	0.010	0.012
C(publication)[Washington Post]	0.0032	0.000	8.394	0.000	0.002	0.004

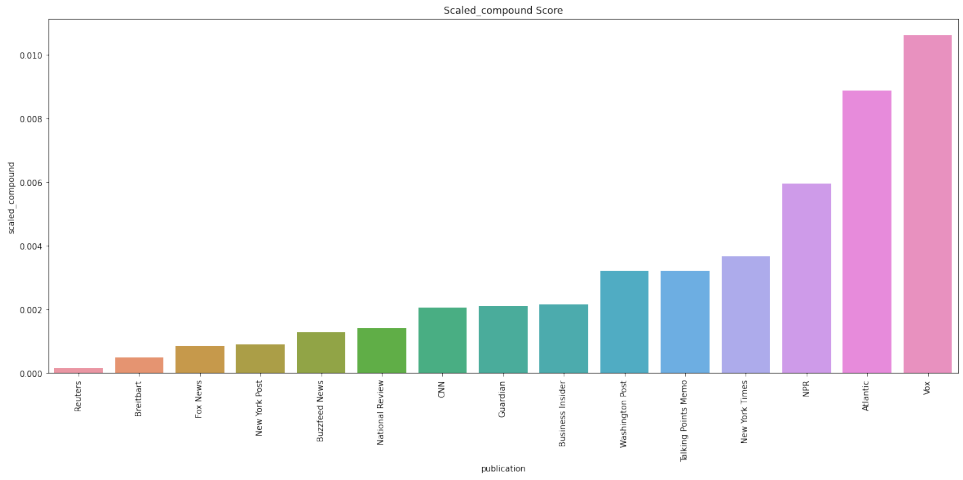
Omnibus:	10027.845	**Durbin-Watson:**	1.928
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1404988.856
Skew:	-0.975	**Prob(JB):**	0.00
Kurtosis:	40.468	**Cond. No.**	3.13

A second Tukey test was performed, and once again the lowest significance level that could be achieved was $\alpha = 0.21$, which is quite high. The full output of the Tukey test for neutrality scores is in Appendix C, and the full output of the Tukey test for compound scores is in Appendix D.

There were a total of 210 comparisons made in the two Tukey tests, too many to discuss. But a few key insights are:

- The New York Times has significantly higher neutrality in articles about Democrats than any other publication except for Fox News and Buzzfeed News
- Vox has a significantly higher compound sentiment score in articles about Democrats than nearly any other publication
- National Review is only significantly different from Vox
- Most publications are only significantly different from the three highest: Vox, Atlantic, and NPR; the other publications seem to form a group

The following bar chart makes non-rigorous comparison of compound sentiment across publications quite easy.



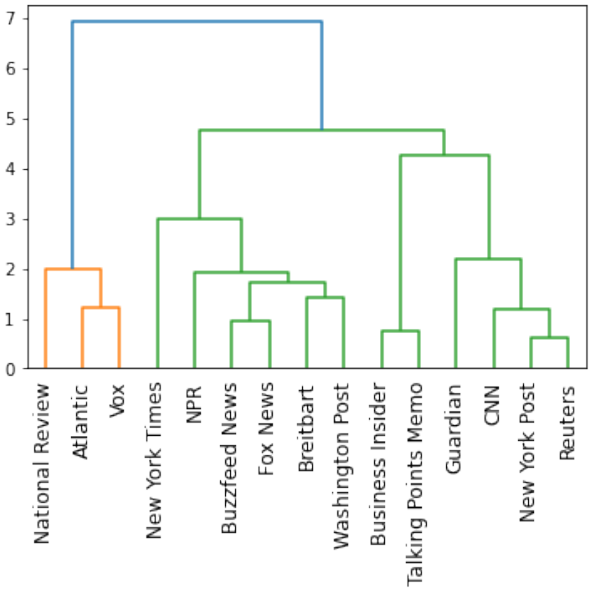
Clustering

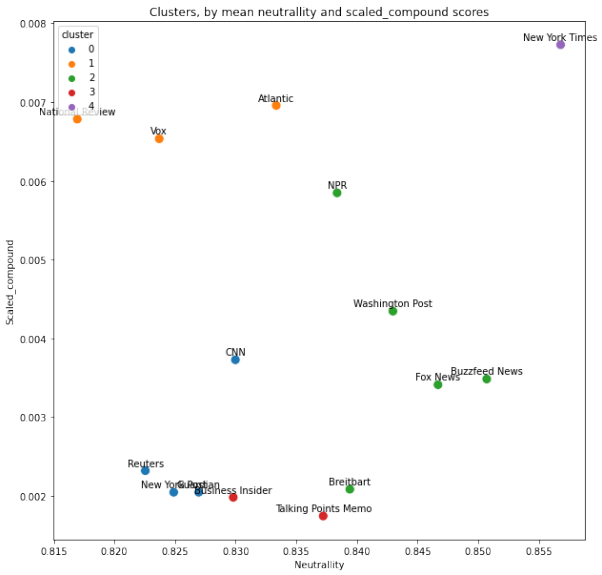
The distribution graphs and Tukey tests showed that most publications are fairly similar, and that natural groups seem apparent. To explore this, clustering was performed in three different ways. First, agglomerative hierarchical clustering was used to find which publications are "closest" together in terms of sentiment. Then, K-Means clustering was used to split articles into groups. Finally, the DBSCAN algorithm was used to find hidden, naturally occurring groups within the data.

Agglomerative Hierarchical Clustering

In this type of clustering, the average distance between all pairs of publications is calculated; the two nearest publications are grouped; the pairwise distance between the group and all remaining publications is calculated, and the two nearest are grouped; and so on until all publications are linked together. A graphical visualization of this clustering shows which publications are most similar to one another.

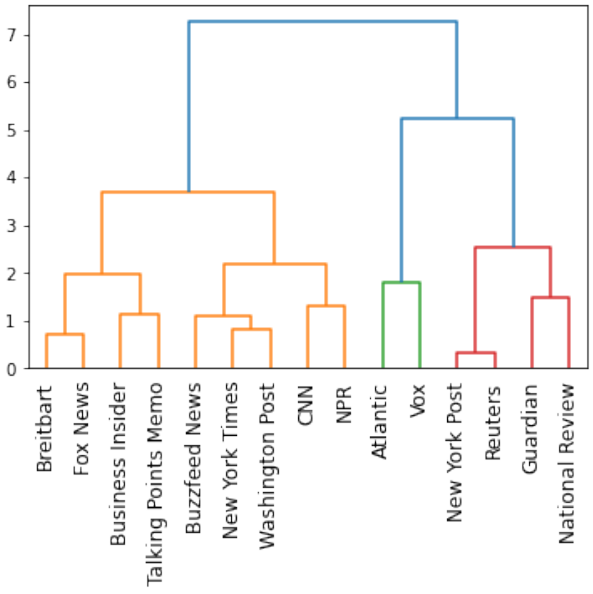
In articles about Republicans, the two most similar publications are the New York Post and Reuters. Breitbart's nearest neighbor is the Washington Post, and Fox News is closest to Buzzfeed News. National Review, Atlantic, and Vox, three categorized earlier as "Blog-Style News", are in a separate cluster from all other publications.

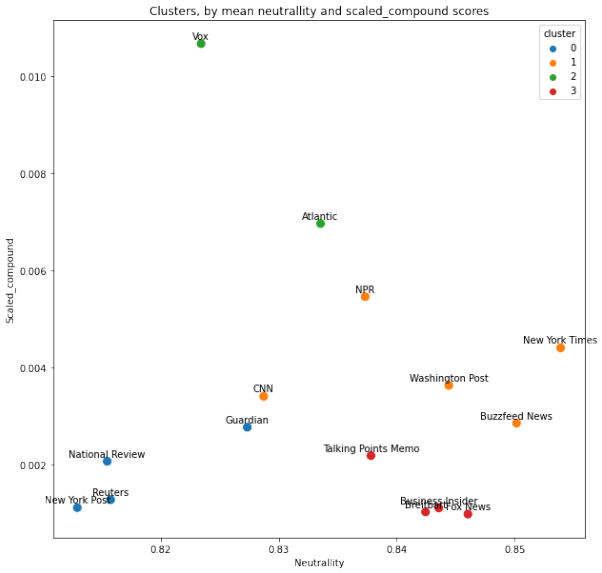




When plotted on axes of neutrality and compound scores, colored by 5-cluster agglomeration, the New York Times appears in its own group, far off to the upper right (more neutrality, more positive compound sentiment) than all the others.

Comparing articles about Democrats, the two most similar publications are again the New York Post and Reuters. This time, the New York Times and Washington Post are nearest-neighbors, CNN and NPR are nearest neighbors, and Fox News is closest to Breitbart. Atlantic and Vox appear in their own cluster, as do the New York Post, Reuters, Guardian, and National Review.

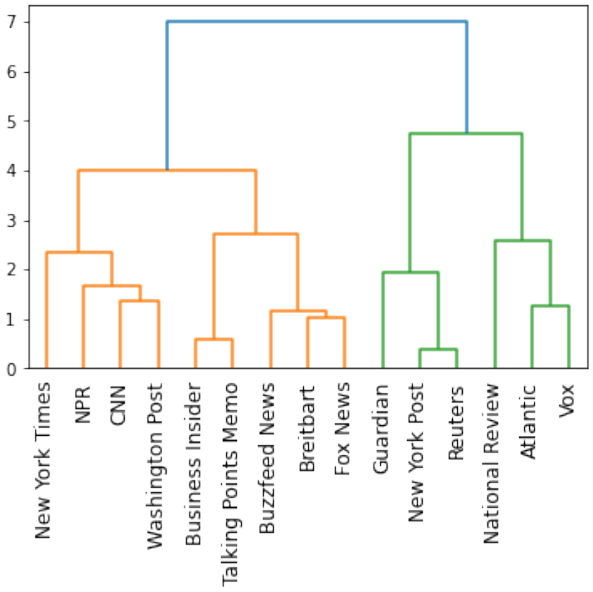


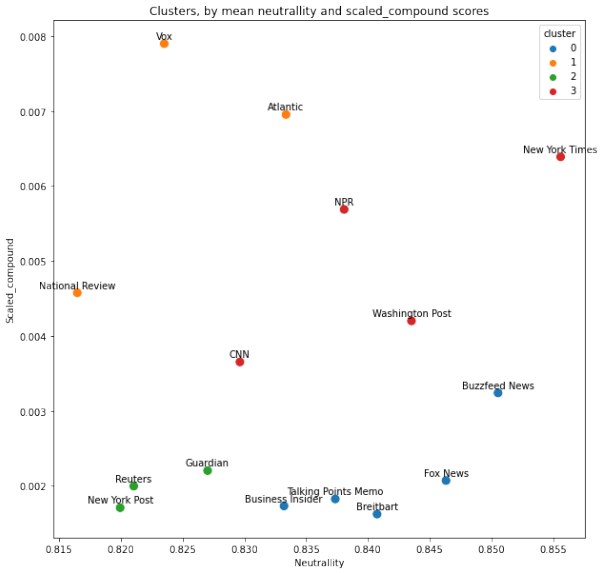


In this plot, with axes of neutrality and compound sentiment, the high outlier is Vox, far more postivie than the others but in the lower range of neutrality. The New York Times appears in the far right, most neutral, area, but is in the lower half on the compound sentiment axis. In articles about Republicans, the New York Times is both highly neutral and comparatively positive; in articles about Democrats, it is neutral but more middle-of-the-road in compound sentiment.

When all political articles are considered together, the results are quite interesting. Once again, Fox News and Breitbart are closest together, as well as Atlantic and Vox, and the New York Post and Reuters. This time, CNN and the Washington Post are closest, then NPR is added and finally the New York Times. It almost appears as though there should be four groups:

Group 1	**Group 2**	**Group 3**	**Group 4**
New York Times	Business Insider	Guardian	National Review
NPR	Talking Points Memo	New York Post	Atlantic
CNN	Buzzfeed News	Reuters	Vox
Washington Post	Breitbart		
	Fox News		





In the plot with axes of neutrality and compound sentiment, the publications are more spread out than before. The New York Times is still far off to the right and relatively high. Vox is quite high but close to the left end. The cutoff lines for the four groups are easy to see.

K-Means Clustering

In K-Means clustering, the algorithm is told the number of groups to form, and it separates the articles into those groups but minimizing the within-cluster sum of squares. There are several ways to estimate how many clusters should be formed; this analysis used both elbow plots and silhouette plots, and in every case found three to be the optimal number of clusters. All features were standardized before clustering.

In all groups of articles, the three clusters followed the same pattern:

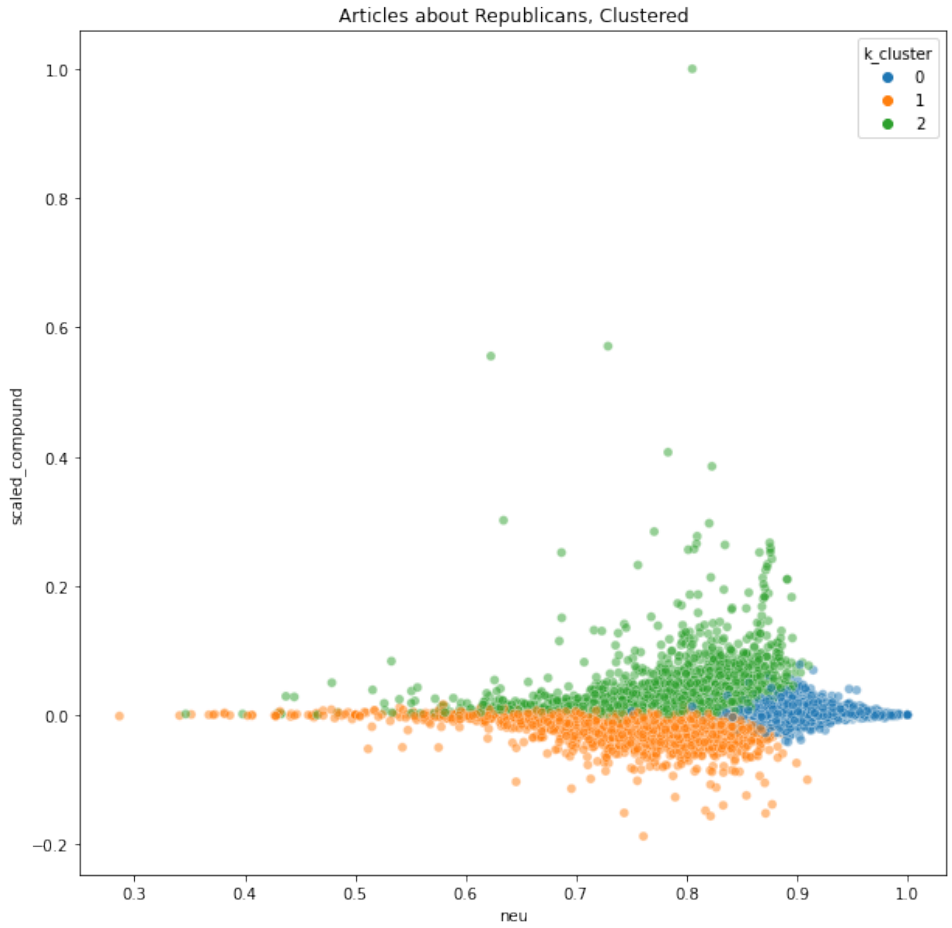
- A cluster of articles with negative compound scores and relatively low neutrality
- A cluster of articles with positive compound scores and relatively low neutrality
- A cluster of articles with high neutrality and compound scores near zero

After clustering, the representation of each publication in each cluster was compared with its proportion of articles in the entire group, and a difference of more than 0.01 was taken to indicate that the specific publisher was either under-represented or over-represented in that cluster.

Articles about Republicans

UNDER-Represented		
Low neutrality, negative	**Low neutrality, positive**	**High neutrality**
NPR	Business Insider	Guardian
New York Times	Reuters	National Review
Talking Points Memo	Talking Points Memo	Vox
Breitbart		

OVER-Represented		
Low neutrality, negative	**Low neutrality, positive**	**High neutrality**
Guardian	Atlantic	Business Insider
National Review	Vox	New York Times
Reuters	National Review	Talking Points Memo
Vox		



Articles about Democrats

UNDER-Represented

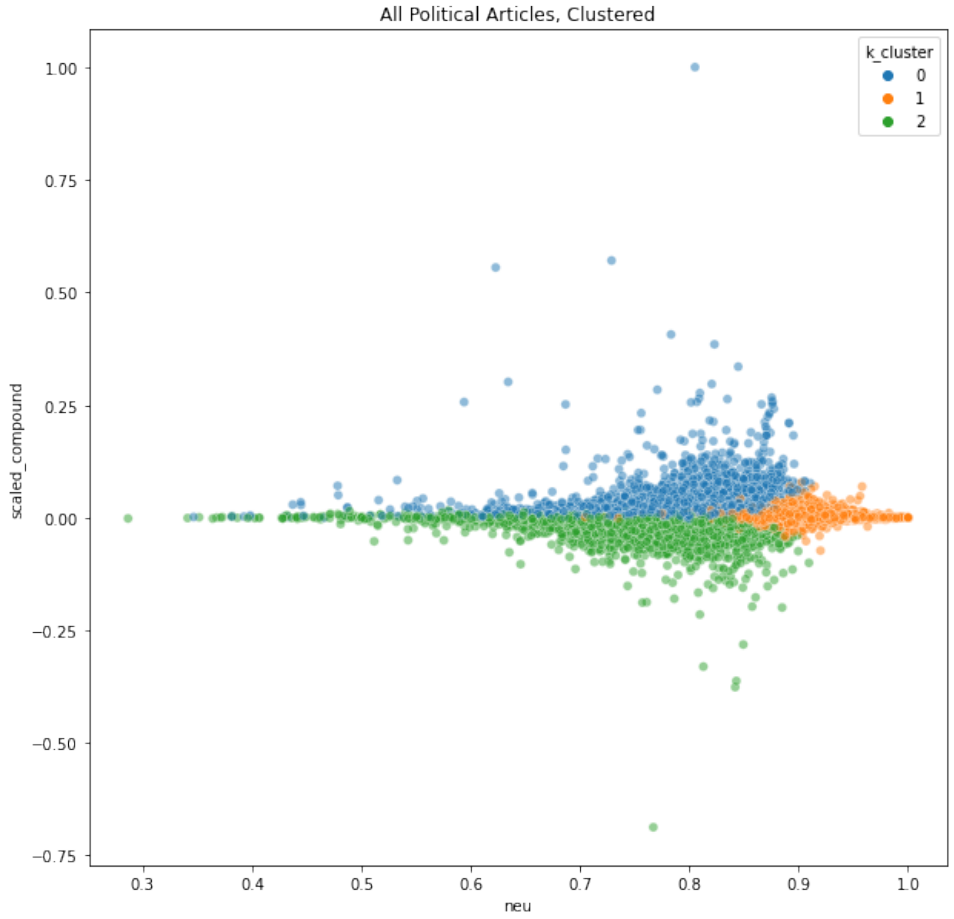
Low neutrality, negative	Low neutrality, positive	High neutrality
Talking Points Memo	Breitbart Fox News	Atlantic National Review Guardian Vox

OVER-Represented

Low neutrality, negative	Low neutrality, positive	High neutrality
National Review New York Post Reuters	Atlantic Vox National Review	Breitbart Fox News Talking Points Memo

All Articles

UNDER-Represented		
Low neutrality, negative	**Low neutrality, positive**	**High neutrality**
NPR New York Times Talking Points Memo	Breitbart Business Insider Fox News Talking Points Memo Reuters	Atlantic Guardian National Review Vox
OVER-Represented		
Low neutrality, negative	**Low neutrality, positive**	**High neutrality**
Guardian National Review Reuters	Atlantic National Review Vox	Breitbart Business Insider Fox News New York Times Talking Points Memo



DBSCAN

DBSCAN is a well-known and often-used clustering algorithm. It groups data-points together based on density; if finds the core points, those with many neighbors with a certain distance, and then groups points together by distance to those core points. The algorithm does not need to know how many clusters to make, it only needs to know the maximum distance between two points that can be considered neighbors. Since there was a large number of articles (more than 60,000), this analysis required at least 100 articles to be in a group.

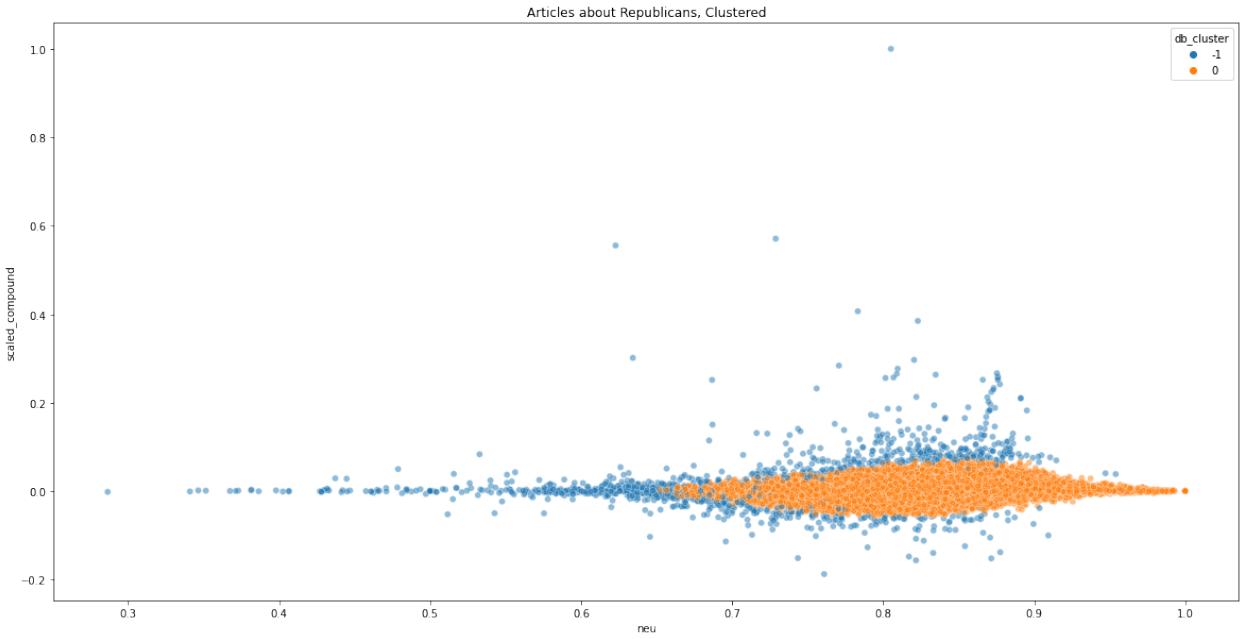
Interestingly, DBSCAN concluded that there is only one real cluster, and all other articles are outliers. In each group of articles analyzed, the one real cluster contains articles with neutrality above approx. 0.65 and compound scores near zero. The one cluster was taken to indicate articles in the mainstream, while outliers are considered just that; outside the mainstream.

After clustering, the proportion of articles in the mainstream or in the outlier cluster was compared to the proportion of all articles from that publication; proportional differences greater than 0.01 were taken to represent over-representation or under-representation.

Articles about Republicans

In articles about Republicans, there were no publications that were disproportionately represented in the Mainstream cluster. Fox News is under-represented in the outliers, meaning they are more like the mainstream media, whereas Breitbart is over-represented and therefore outside the mainstream.

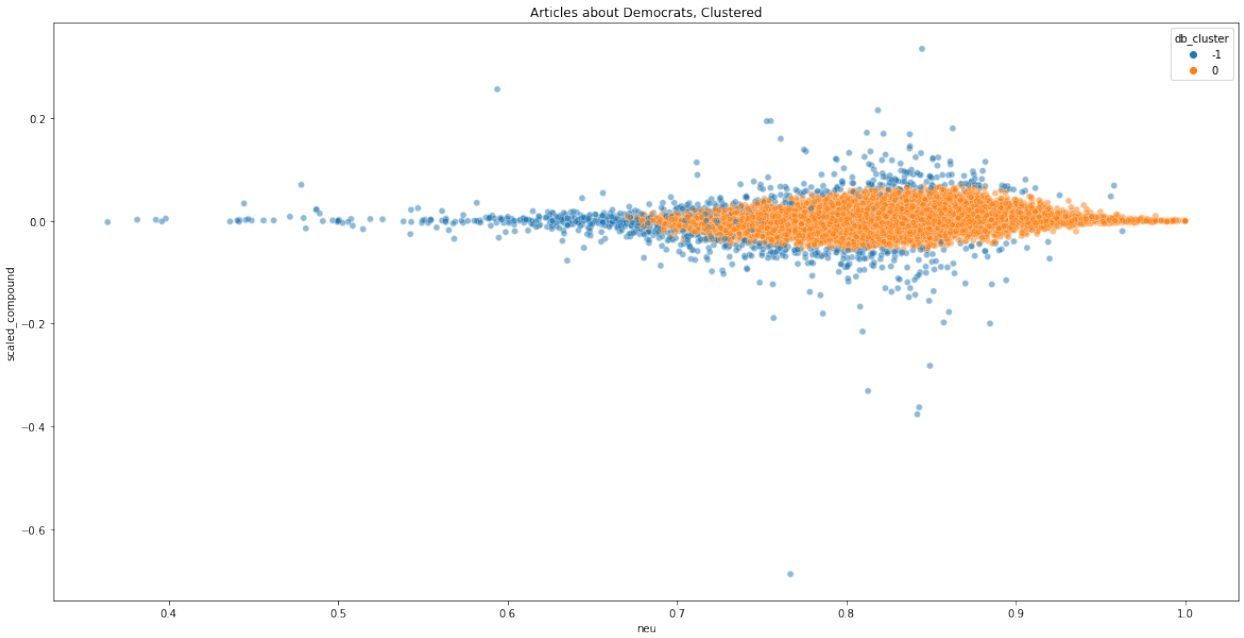
UNDER-Represented	
Mainstream	**Outliers**
	CNN
	Fox News
	Guardian
	New York Times
	Reuters
	Washington Post
OVER-Represented	
Mainstream	**Outliers**
	Atlantic
	Breitbart
	Business Insider
	Talking Points Memo
	Vox



Articles about Democrats

Again, there were no publications that were disproportionately represented in the Mainstream cluster. Both Breitbart and Fox News are under-represented in the outliers, meaning they are more like the mainstream media, while NPR is over-represented and therefore outside of the mainstream.

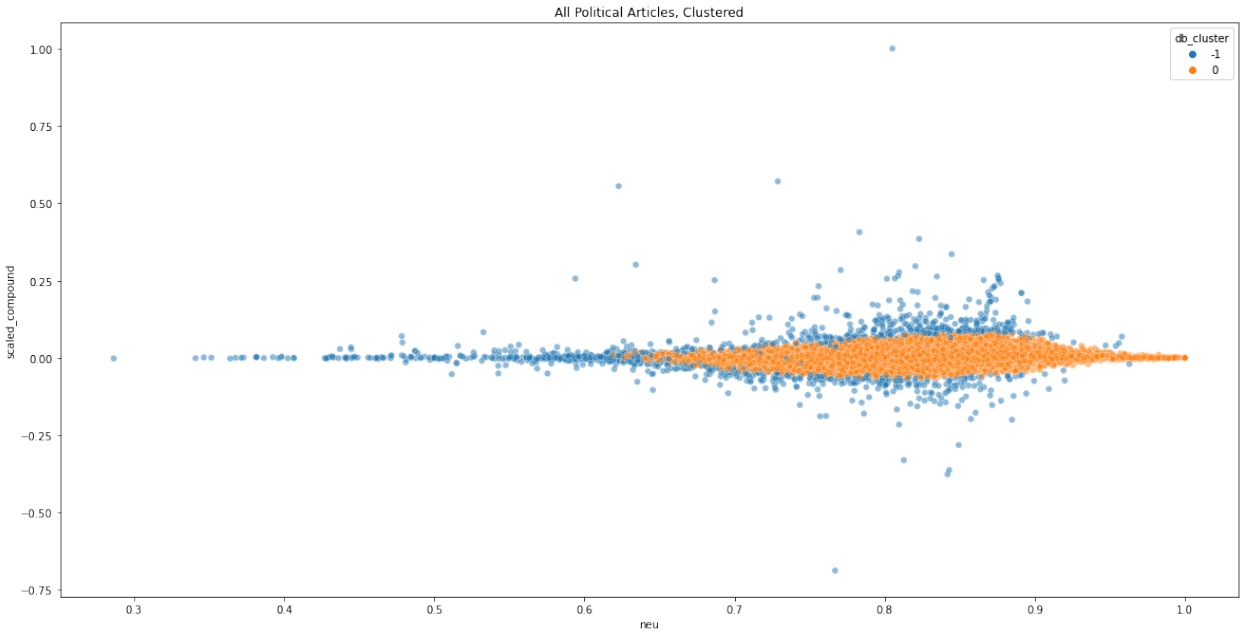
UNDER-Represented	
Mainstream	**Outliers**
	Breitbart
	Fox News
	New York Times
	Reuters
OVER-Represented	
Mainstream	**Outliers**
	Atlantic
	Business Insider
	NPR
	Talking Points Memo
	Vox



All Articles

There are still no publications that are disproportionately represented in the Mainstream cluster. Fox News is under-represented in the outliers, meaning it is more like the mainstream media, and both Breitbart and NPR and over-represented.

UNDER-Represented	
Mainstream	**Outliers**
	CNN
	Fox News
	Guardian
	New York Times
	Reuters
	Washington Post
OVER-Represented	
Mainstream	**Outliers**
	Atlantic
	Breitbart
	Business Insider
	NPR
	Talking Points Memo
	Vox



Conclusion

There are many conclusions that can be drawn from this over-arching exploration of sentiment in political articles. Among the most meaningful are:

- Fox News seems to be fully within the mainstream, despite being the only publication with a negative mean compound score
- NPR tends to stray from the mainstream
- Articles about Democrats tend to have stronger negative sentiment than articles about Republicans
- The New York Times is incredibly neutral and consistent in its reporting
- The New York Times is the most positive publication in articles about Republicans
- National Review is more similar to Atlantic and Vox than to other Breitbart or Fox News

In further research, it would be interesting to perform a time-series analysis of sentiment in news reports during the Trump administration. It would also be interesting to compare more web-based media outlets to the traditional television/paper media and look for conclusions as to the quality of reporting on different platforms.

There is an ever-growing wealth of articles available to be analyzed, and with the current political unrest in the United States, the quality of news being ingested by the population is more important than ever.

Recent improvements in data-based text analysis allow for drawing bias-free inference about political slant; this analysis should be trust-worthy even to the most entrenched consumer.