Final Project

IST 707  
Thomas Bahng, Ted Tinker, Tim Zalk, Michael Znidarsic

[Introduction 3](#_Toc19384822)

[Analysis 4](#_Toc19384823)

[About the Data 4](#_Toc19384824)

[Incorrect and Unusable Data 8](#_Toc19384825)

[Data Transformation 9](#_Toc19384826)

[Subset Data 9](#_Toc19384827)

[Text Mining and Document Term Matrices 10](#_Toc19384828)

[Sentiment Analysis 12](#_Toc19384829)

[Item Matrix for Association Rule Mining 19](#_Toc19384834)

[Models 20](#_Toc19384835)

[Association Rule Mining 20](#_Toc19384836)

[K-Means Clustering 37](#_Toc19384840)

[Hierarchical Clustering 37](#_Toc19384841)

[Decision Trees 37](#_Toc19384842)

[Naïve Bayes 41](#_Toc19384843)

[Support Vector Machines 44](#_Toc19384844)

[Results 52](#_Toc19384853)

[Association Rules 52](#_Toc19384854)

[K-Means Clustering 55](#_Toc19384855)

[Hierarchical Clustering 58](#_Toc19384856)

[Decision Trees 61](#_Toc19384857)

[Naive Bayes 61](#_Toc19384858)

[Support Vector Machines 61](#_Toc19384859)

[Text Mining: Sentiment 64](#_Toc19384862)

[Conclusions 65](#_Toc19384863)

# Introduction

Current events consume many American’s daily attention. Throughout a day, society is bombarded with news through social media, mobile device notifications, and traditional media like radio, TV, and print. Although there is a myriad of news sources people can choose from, there are only several dozen large, national publishers – most of which are owned by six primary corporations[[1]](#footnote-1).

The evolution of the internet and the pervasiveness of social media have changed the way many people access and consume news content. Longform news articles have shifted to short, media-driven content. As the use of social media has increased, newspaper readership has continued to decline[[2]](#footnote-2). With the increase of digital consumption, there have also been concerns over the validity and accuracy of published news content. Particularly since the 2016 US presidential election and the Facebook/Cambridge Analytica scandal, there have been growing concerns of fake news[[3]](#footnote-3) and algorithmically generated content, known as deepfake content[[4]](#footnote-4).

Analysis of published news content can aid in confirming the validity of articles and their claims. This can be done by detecting how similar the content of the article is to that of known fake news or other works by the same journalist that is known to be factual. These same analysis techniques can also be a vital tool in identifying articles for archival and research practices, such as identifying an unknown author from an old article. The analyses can also aid in answering questions about publications and their authors. For example, authors and publications can be identified based on the words that they typically use. Changes in sentiment can be predicted based on articles’ typical responses to world and political events. The change of quality over time can also be measured based on the complexity of words used to determine if articles are getting more or less complicated.

# Analysis

## About the Data

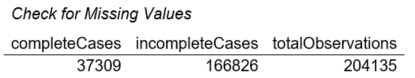
The All The News dataset was retrieved from the Components Publishing website components.one. The original dataset was made available on Kaggle[[5]](#footnote-5), but the larger, updated dataset was published on Components’ website[[6]](#footnote-6). The data was made available in an SQLite .db database file. This database contains the table ‘longform’ which was exported as a CSV file. This CSV file was then read to support the analysis.

The dataset contains 204,135 news articles from 18 American publications. It includes attributes such as date of publication, the title of the news article, the name of the publication, the full article text, year, month, and in some cases the article URL. Although the full timespan of the date range was from 2013 to 2018, the bulk of the data resided from 2014 to 2018. An additional field labeled ‘digital’ is also included in the dataset; however, its meaning is not defined and cannot be inferred from its contents, leading to it being excluded from this analysis.

Table 1



Table 2



The data contained 37,309 complete cases of observations, with the number of remaining incomplete cases at 166,826.

Missing values were assessed by publication. Observations with the publication missing contained mostly missing values across the other variables. Eleven publications were missing ‘section’ entirely.

A screenshot of a cell phone

Description automatically generated

Figure 1

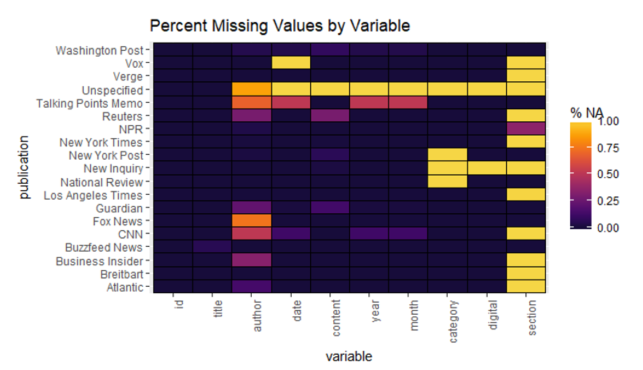


Figure 2

The data was filtered down to observations with complete cases in key variables that were deemed relevant to support this analysis. These included ‘publication’, ‘content’, ‘date’, ‘year’, ‘month’, ‘category’, and ‘digital’.

Once filtering was complete, the data showed a much more robust landscape for analysis.

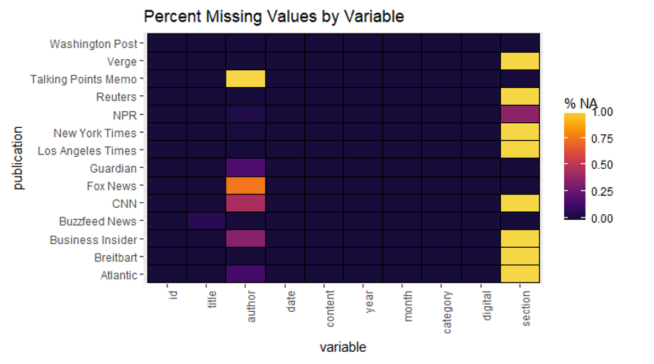
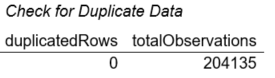


Figure 3

Table 3



The assessment did not find any duplicated information. All documents are mutually exclusive from one another.

### Incorrect and Unusable Data

The title and content variables contained HTML tags and non-UTF-8 string characters which were subsequently cleaned into a useable text format. UTF-8 encoding was first specified for the variables and then converted to latin1 format. HTML tags were removed along with carriage return and new-line characters using global substitutions.

Nominal and potential class labels were converted to type factor. These included the variables ‘year’, ‘month’, ‘publication’, ‘category’, ‘digital’, and ‘section’. The publication date field was converted to date format. The URL field, although not incorrect, was removed from the dataset as it was deemed irrelevant for this analysis.

### Data Transformation

Title and content were combined as they were more useful as one string element than separate.

### Subset Data

Data had been filtered for complete cases in the key variables ‘**publication**’, ‘**content**’, ‘**date**’, ‘**year**’, ‘**month**’, ‘**category**’, and ‘**digital**’. The data was then further filtered to remove ‘section’ given the prevalence of missing values and filtered down to only the most prolific authors across all the publications as there were over 15,000 authors. This was done both for feasibility and to retain the core substantive messages from each publication. For each publication, just the top 80% of authors by the number of articles produced were kept.

Once the data was subset, what remained were 108,687 articles written by the 4,532 most prolific authors across 13 publications.

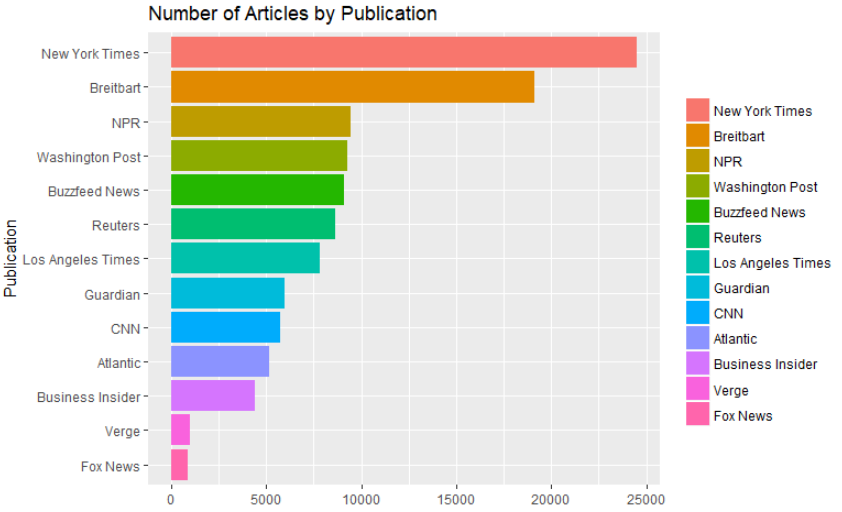


Figure 4

### Text Mining and Document Term Matrices

Document term matrices (DTM) were created to support supervised and unsupervised learning methods. Stop words included any explicit mention of the publication. As stemming was deemed useful for topic-oriented classification, the primary document term matrix applied this technique. A normalized version of this DTM was also made available by applying TF-IDF weighting to the term frequencies.

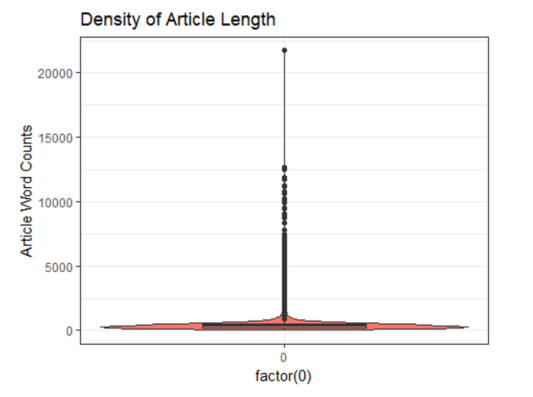


Figure 5

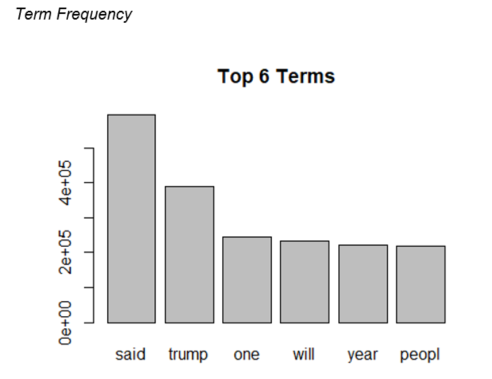


Figure 6

75% of the terms had frequency counts less than 4. The size of the DTM was reduced to only include terms with term totals greater than 99.2th percentile (i.e., a minimum of 2000).

### Sentiment Analysis

Articles’ sentiment was measured across all 100,000 news articles using the AFINN lexicon. This was made available to enable the assessment of trending sentiment and overall publication sentiment. As stemming inherently affects sentiment extraction, a separate DTM was created solely to support this activity. The following plot displays top terms that contributed the most to positive and negative sentiment based on the data.

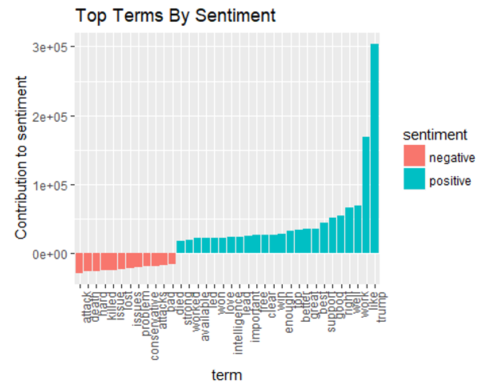


Figure 7

Overall numeric sentiment was net-positive for the 108,687 news articles used in this analysis. However, the distribution of positive and negative articles shows a clear picture of an approximately normal curve that is slightly left-skewed (negative). Boxplots indicate that there are far more extreme points of negative sentiment than there are positive.

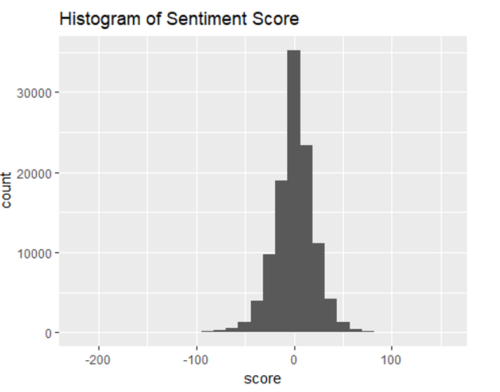


Figure 8



Figure 9

This section will explore sentiment across the news articles in the repository. Assessment will include trends, word clouds, and overall sentiment by publication. Trending will look at the years between 2014 and 2018 given a sufficient number of observations in each of these years.

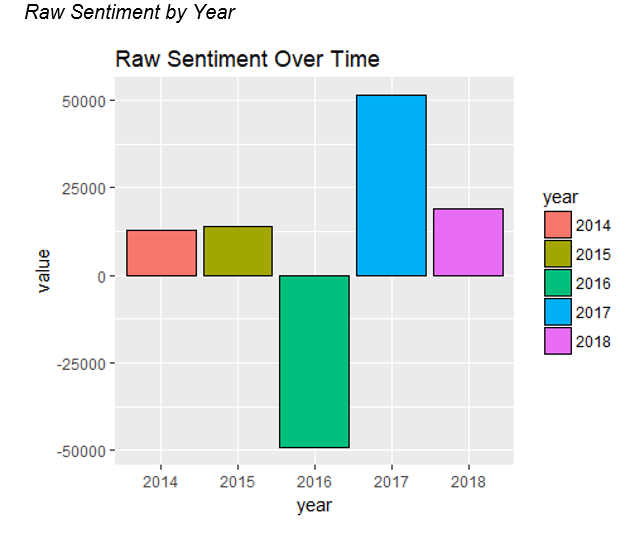


Figure 10

#### Normalized average sentiment by year

For each year, sentiment scores were normalized using Min-Max transformation. These values were then grouped by year and aggregated by the mean.

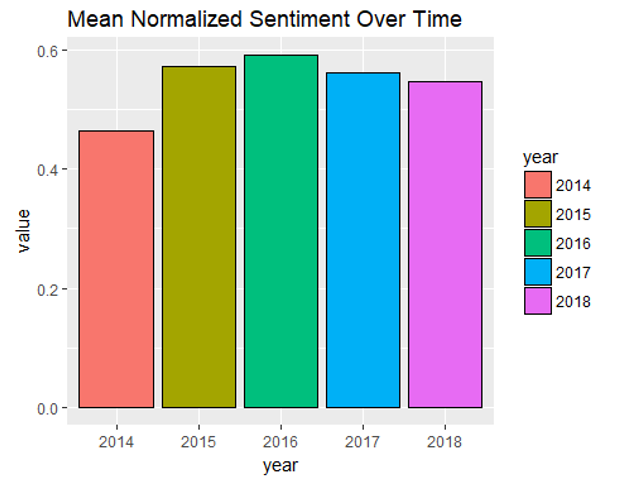


Figure 11

On average, sentiment increased from 2014 to 2015 and began to decrease slightly from 2016 to 2018.

#### Sentiment Word Clouds of Articles: 2014 - 2018

****

Figure 12

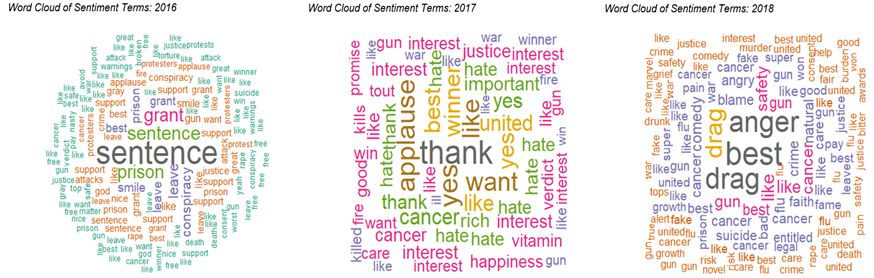
****

Figure 13

#### Sentiment by Publication

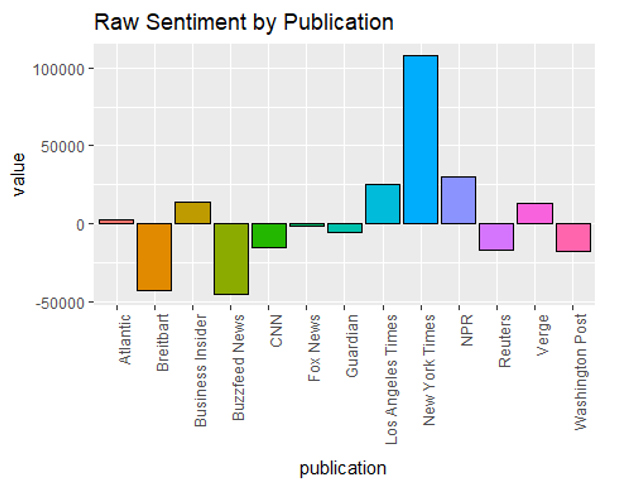


Figure 14

Measured by raw sentiment score, it appears overall that NYT has the highest positive sentiment in their articles. Breitbart and Business Insider have the lowest negative sentiment.

#### Normalized average sentiment by publication

For each publication, sentiment scores were normalized using Min-Max transformation. These values were then grouped by publication and aggregated by the mean.

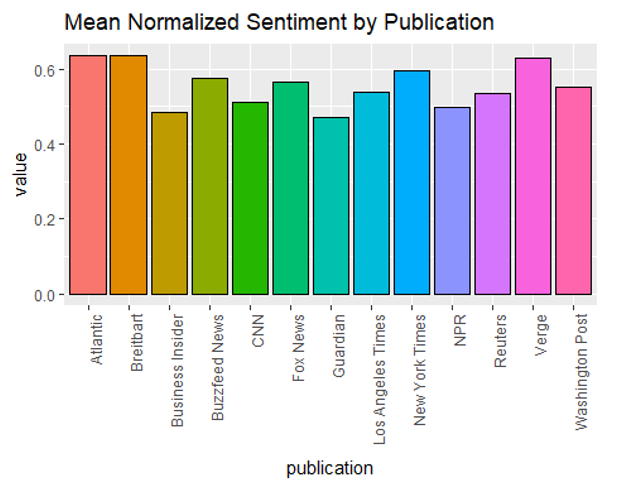


Figure 15

By normalizing the sentiment scores by publication, it appears **on average** that Atlantic, Breitbart, and Verge have the highest positive sentiment. The lowest negative sentiment on average seems to be from articles by Business Insider, Guardian, and NPR.

### Item Matrix for Association Rule Mining

In this analysis, association rule mining would be done on the article text. To support this activity, a third DTM was constructed specifically not to apply stemming and filter only on articles published in 2018. This was again due to feasibility and to focus on substantive and interesting terms. Ideally, the resulting terms would not include common function words, extremely rare words, and overly common words.

Once the DTM was constructed, respective terms were combined for each document and converted to a transactions dataset.

## Models

### Association Rule Mining

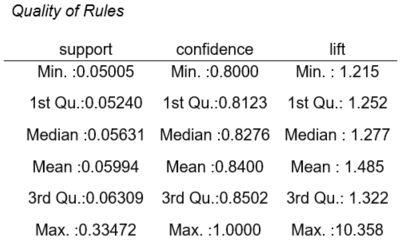
What patterns and relationships can be learned from a repository of news articles? The objective of this analysis will be to use association rule mining (ARM) to discover interesting and important patterns from news publications in the first quarter of 2018. The ARM technique is used in many text-mining applications, including web search engines and other text information retrieval systems. Similar benefits of ARM are realized in the medical field where patterns in symptom-related free-form text are connected to recommended treatment options, given the information available in electronic health records. Exploration of the news is the key objective of this analysis. Patterns in the co-occurrence of terms can help users navigate the vast collection of documents for relevant articles as well as give a high-level overview of the top news stories in that specified time period. What terms are most relevant in describing the overall news of 2018? What terms are most relevant for targeted searches about a specific topic? These questions will be addressed in this report.

One fundamental limitation of this analysis is the memory-intensive process of the apriori algorithm. Unlike some other data-mining techniques applied to text data, the method used in this analysis uses a stopwatch to prevent exceeding a certain threshold for running this algorithm. There is potential to generate thousands, hundreds of thousands, or millions of rule itemsets from a relatively small dataset of 3,836 observations, depending on the parameters used: results come at an increasingly higher cost of memory. These considerations are taken into account within this analysis so that the rule generation completes from start to finish without a truncation.

Rule generation was first simulated over predefined parameter settings of minimum support and minimum confidence values. Minimum support was tested across the values of 0.05, 0.10, 0.15, and 0.20. Minimum confidence was tested over the values of 0.8, 0.9, and 1.0. Based on the results, it was determined that specifying minimum support of 0.05 and minimum confidence of 0.8 for the apriori algorithm would be a good baseline to generate a sufficient number of rules with reasonable average quality and size of rule itemsets.

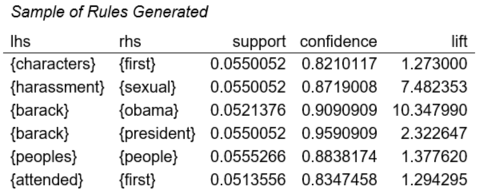
#### Generate Rules with minimum support of 0.05 and confidence of 0.8

Table 4

****

A total of 43,603 rules were generated. The average size of rule itemsets was k = 4.2. On average, the rules exhibited a frequency of 6%, 84% of the time they were found to be true, and exhibited an acceptable average lift measure of 1.5.

Table 5



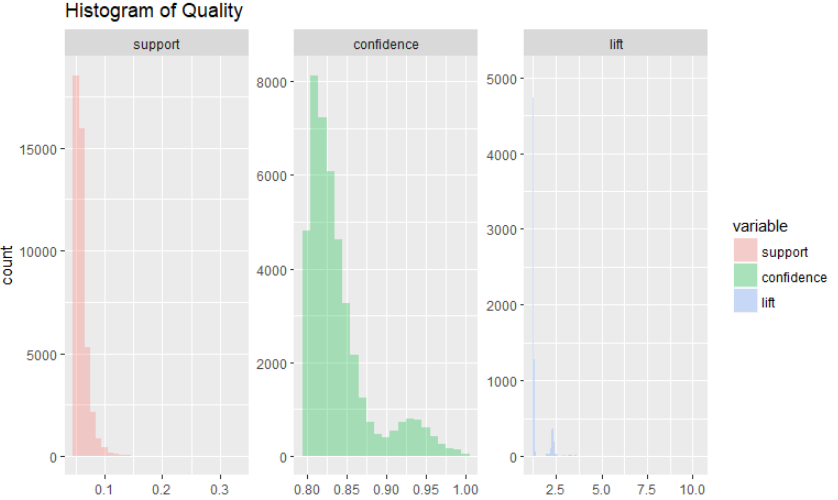


Figure 16

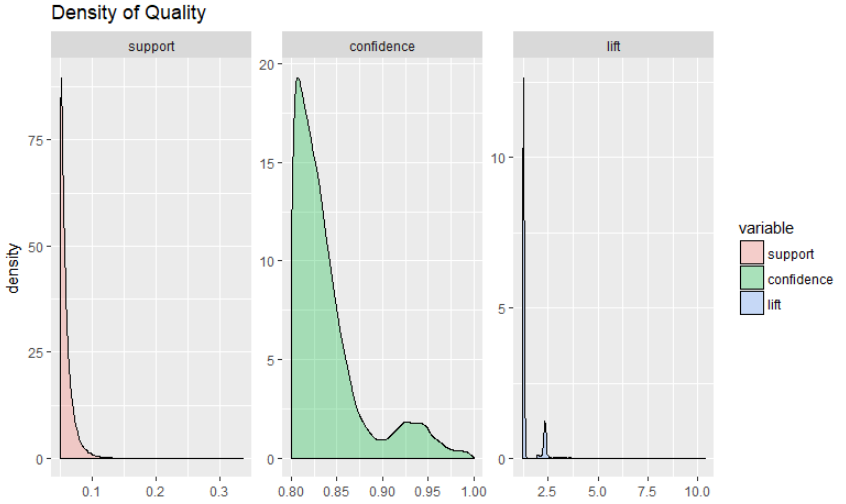


Figure 17

All distributions of quality display a right-skew and include values in high ranges. There is a small population of rules with support upwards of 30%, confidence at 100%, or lift of 10.

The following plots graphically display all of the rules.

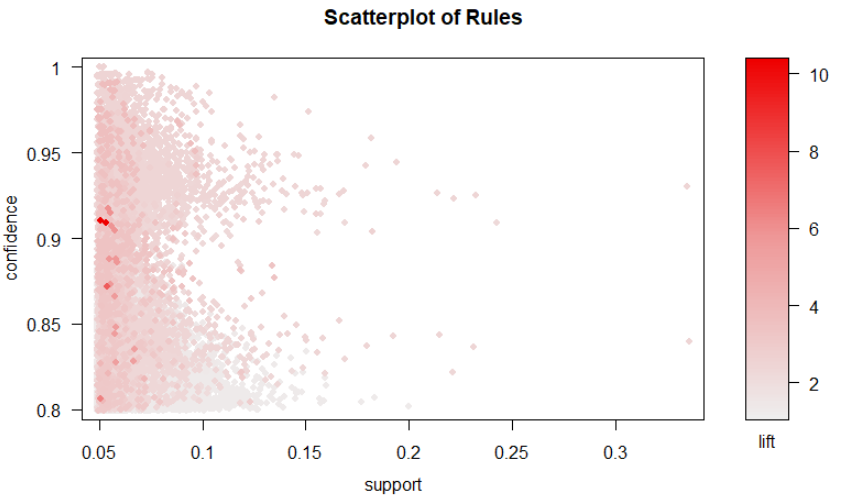


Figure 18

It can be seen that rules with very high lift values of 10 only appear 5% of the time in the dataset.

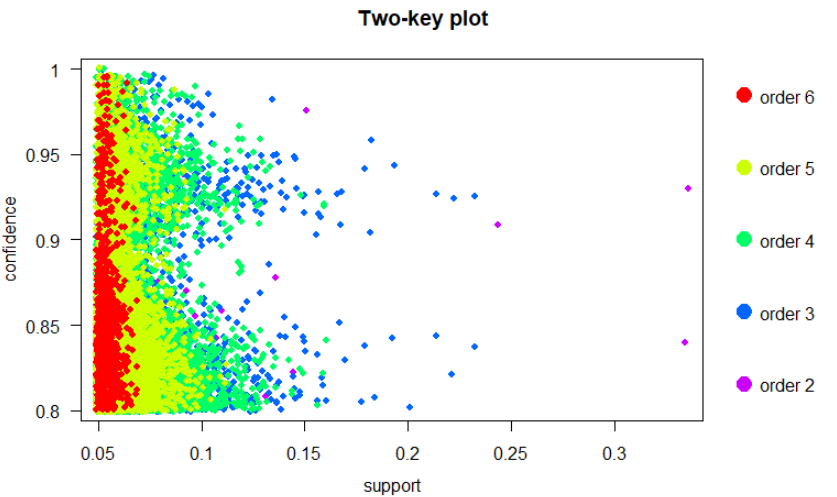


Figure 19

Rule itemsets of size 6 generally appear 5-6% of the time in the data. Any rule over the 15% support threshold can expect only to contain sizes on the order of 3 or lower. Interestingly on this plot, there appear to be very few rule itemsets with a size of 2. This is initially promising because, with text data, one can find many uninteresting and general terms, and larger rule itemsets are expected to yield more interesting associations (i.e., combinations of terms).

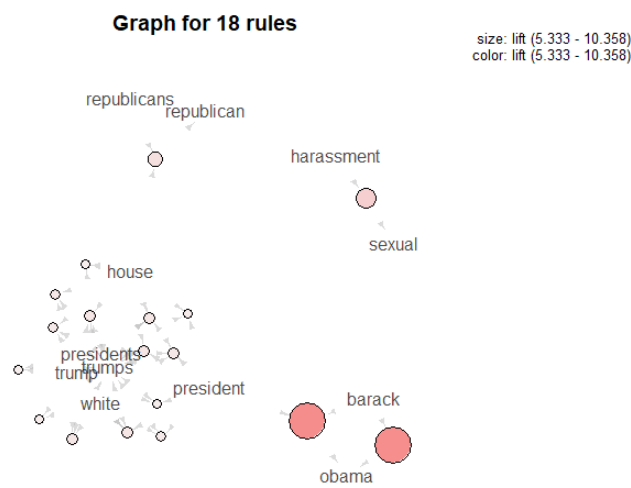
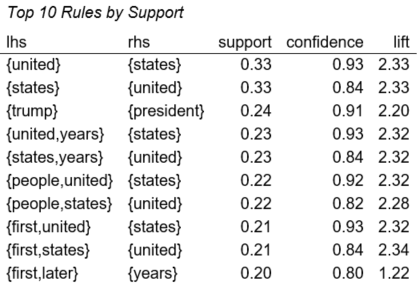


Figure 20

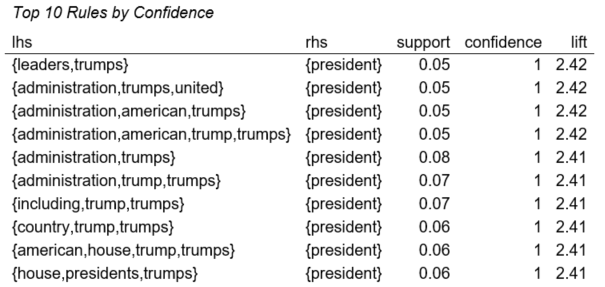
The above graph displays rules with the highest lift (greater than 4.5). These elude to 18 of the most interesting rules from the specified time period which include references to former president Barack Obama, sexual harassment (Me Too movement began in Q4 2017), and House Republicans (major defeat in the House of Representatives for Republicans). Interestingly, sexual harassment was not a consequent of politics, whereas the other consequents have antecedents tied closely with politics.

Table 6



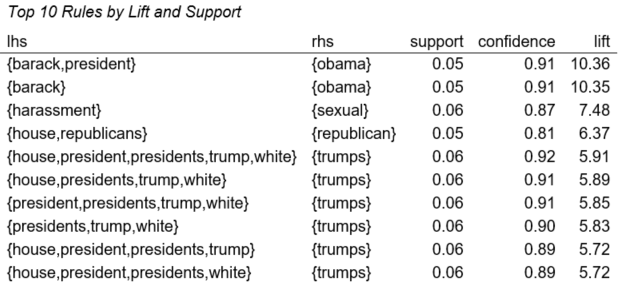
The rules that appear most frequently in the data are generally small and focus on the country and the current president. These rules also display reasonably high confidence and lift.

Table 7



These high-confidence rules, which are true 100% of the time, display keywords that always associate with discussion about the president.

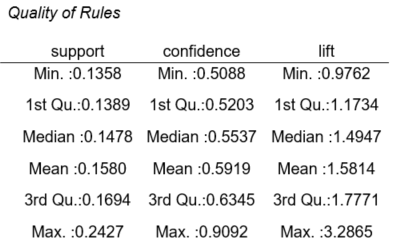
Table 8



The most interesting rules, however, discussed terms that related back to Obama, sexual harassment, and matters related to Trump.

#### Targeted Rules: LHS set to Trump Generate rules based on min. support 0.001 and min. confidence 0.5

Table 9



A total of 14 rules were generated. The average size of rule itemsets was k = 2. On average, the rules exhibited a frequency of 16%, 60% of the time they were found to be true, and exhibited an acceptable average lift measure of 1.6.

Because of the small size of the resulting rule itemsets and relatively low lift (ranging from 1-3), the expectation was not to discover too many interesting associations. If one were to see the term ‘trump’ mentioned in a news article, the content would likely cover topics related to his administration, such as the White House, and comparisons between him and the former president.

Table 10

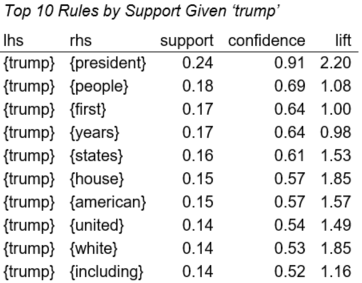


Table 11

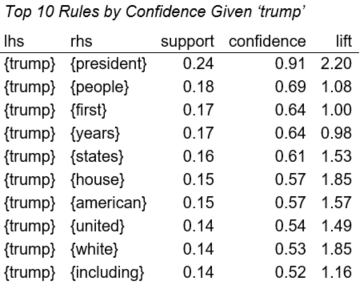
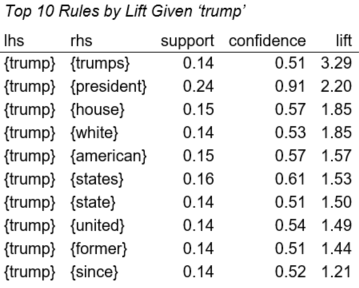


Table 12



The rule with the highest lift contained the possessive term ‘trumps’ on the RHS, signifying that the most interesting news articles given ‘trump’ had to do with the actions he has taken as an individual.

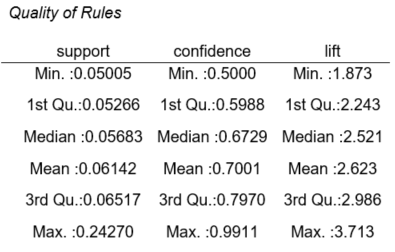


Figure 21

The above plot shows the words that were most associated with articles that contained the term ‘trump’.

#### Targeted Rules: RHS set to Trump Generate rules based on min. support 0.05 and min. confidence 0.5

Table 13

****

A total of 2,084 rules were generated. The average size of rule itemsets was k = 4. On average the rules exhibited a frequency of 6%, 70% of the time they were found to be true, and exhibited an acceptable average lift measure of 2.6.

Table 14

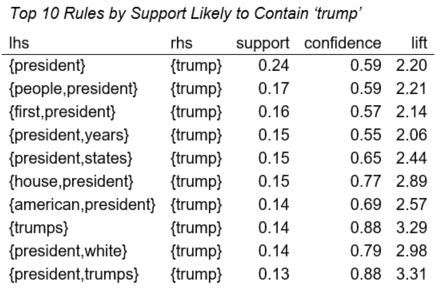
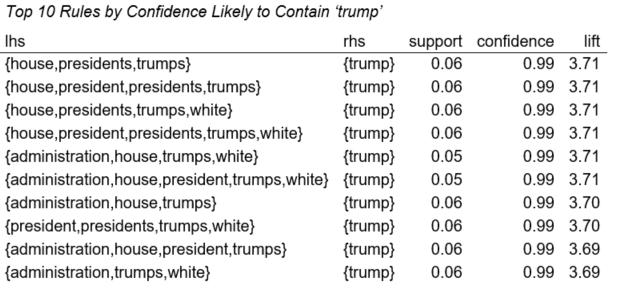
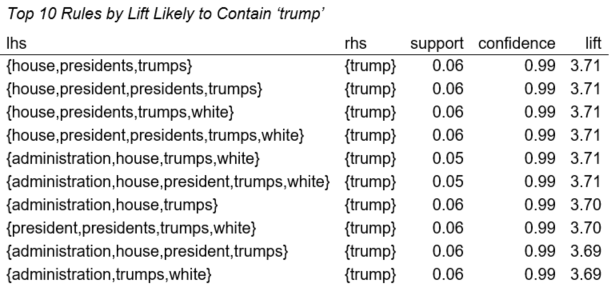


Table 15



Topics having to do with house, presidents, trumps, and administration were almost always about Trump.

Table 16



The most interesting rules that determined whether Trump was mentioned in a news article were almost always, once again, concerning the details of the White House and his administration.

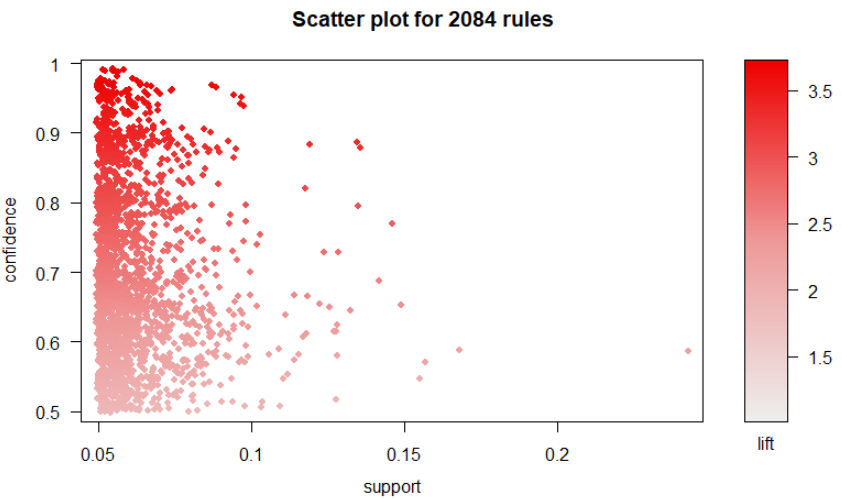


Figure 22

The most interesting rules were also found to be true more than 85% of the time.

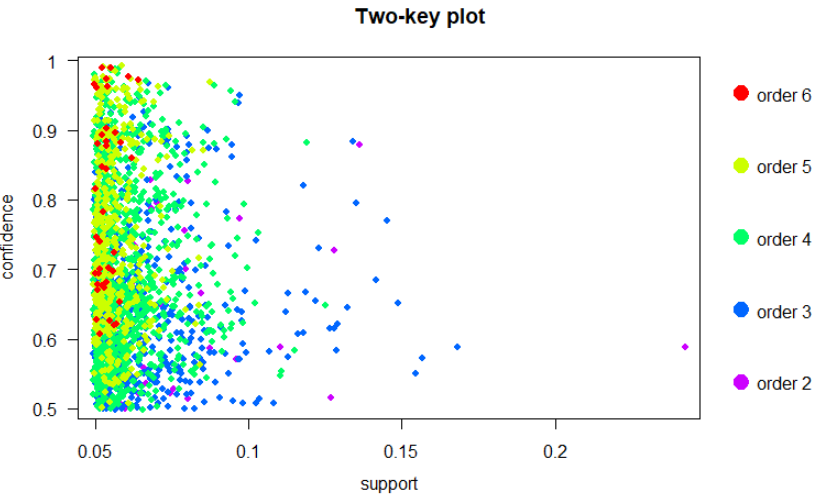


Figure 23

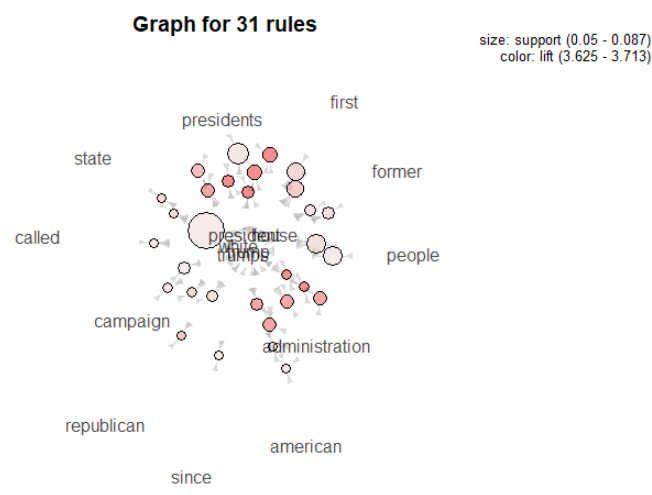


Figure 24

The most interesting associations that lead to articles mentioning trump include the terms ‘**presidents**’ and ‘**administration**’. These rules were found to be true on average 98% of the time and occurred in 6% of all news articles by the New York Times in Q1 of 2018.

### K-Means Clustering

In an exploratory effort to understand similarities between publications, a K-Means clustering model was used to group publications based on several aggregate statistics. One cluster was executed using mean article length and mean normalized sentiment. Another was created using those same inputs, but with the addition of normalized vocabulary standard deviation. This input is a measure of the spread of the distribution of word frequencies summed from that publication’s articles. Both models sought to split the publications into two clusters. Both used the Euclidean distance measure.

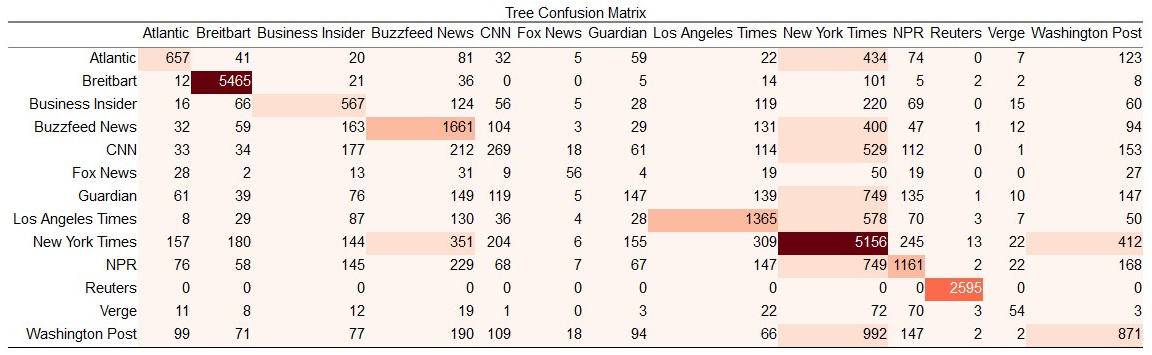
### Hierarchical Clustering

Hierarchical clustering can be performed using articles’ publishers as document names – this may be able to show if publishers’ articles are similar in the language they use. To accomplish this, the publication name is set as the document names in the document term matrix by altering the dtm.pubs$dimnames$Docs element of the object. After converting the document term matrix to a standard R matrix, a random set of rows is sampled due to computational limitations. The euclidean distance was calculated for the sample, and then hierarchical clusters created using the ward.D method. Dendrograms for the hierarchical clusters are provided in the Results section.

### Decision Trees

Trees and forests are quite robust in their ability to sort articles by publication using only a corpus of their lexicon and their length in characters. Trained on this data, a default decision-tree model from the package “ctree” was able to match 60% of the test data to the correct publisher among 13. Unfortunately, being unpruned, the tree is labyrinthine and illegible. See the figures below.

Table 17



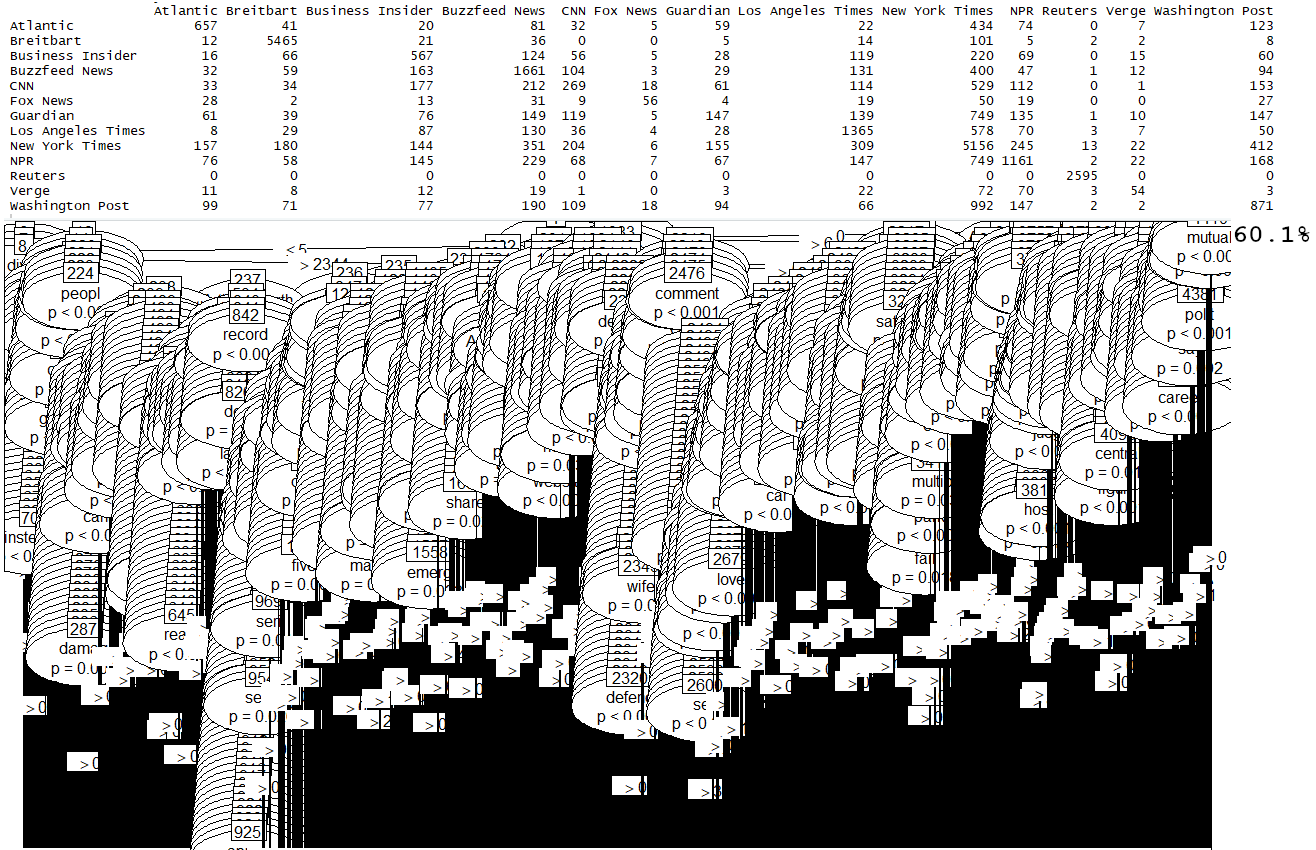
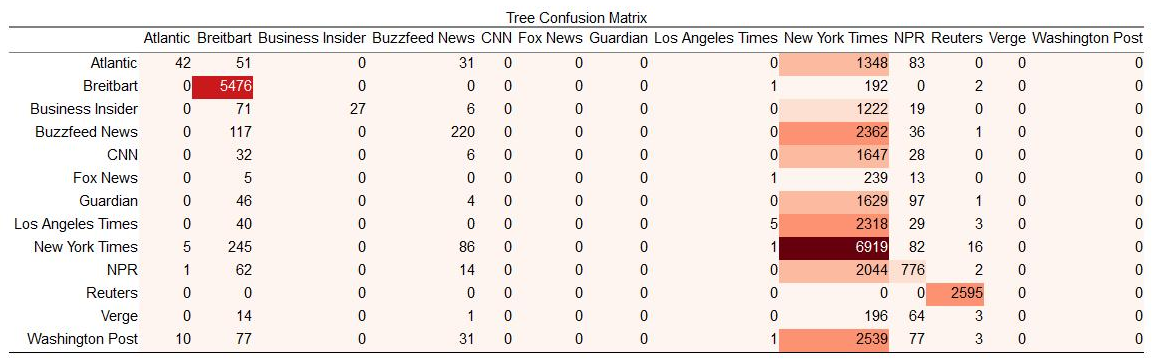


Figure 25

A version of the tree pruned to five levels is only 48% accurate but provides insight into the model’s reasoning. See the figures below. In the leftmost branch, “said” and “say” apparently have much sway in the model. In other circumstances, these words might be dismissed as “stop words,” but apparently publishers use these words with different frequencies. Similarly, “your” in the third layer belies some writers lapsing into second-person.

Table 18



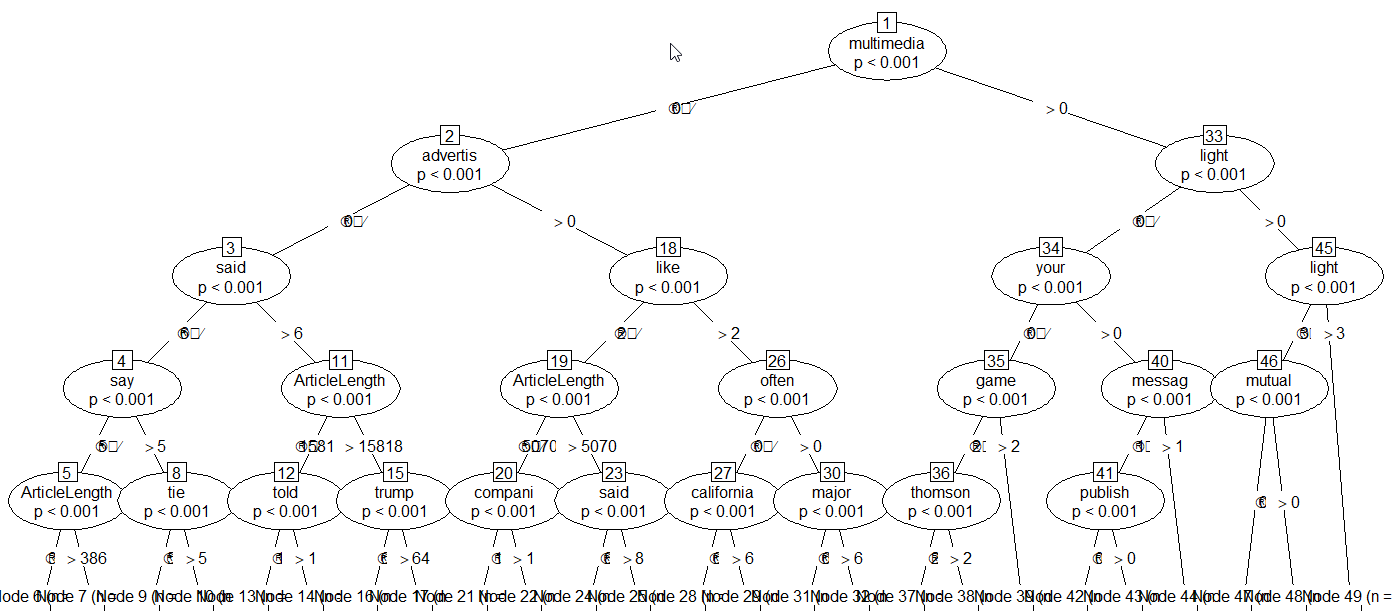
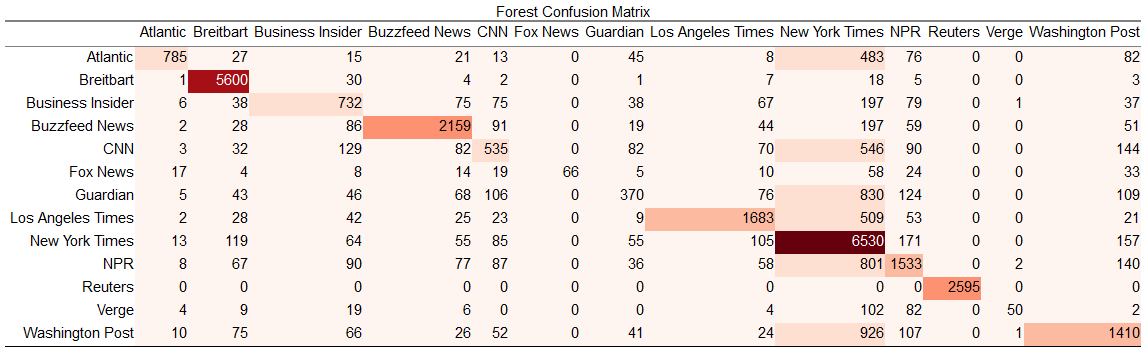


Figure 26

A random forest is more accurate than a single tree: see figures below for a confusion-matrix and list of words in order of Mean Decrease GINI for a forest with 500 trees sampling 500 random variables each step. The most valuable words for sorting are, somewhat suspiciously, “advertis” and “thomson”. Are some articles about a person named Thomson? Why is Thomson so much more influential than “Donald” and “Trump”? Did some publications include advertisements in their articles, letting the forest take a shortcut?

Table 19



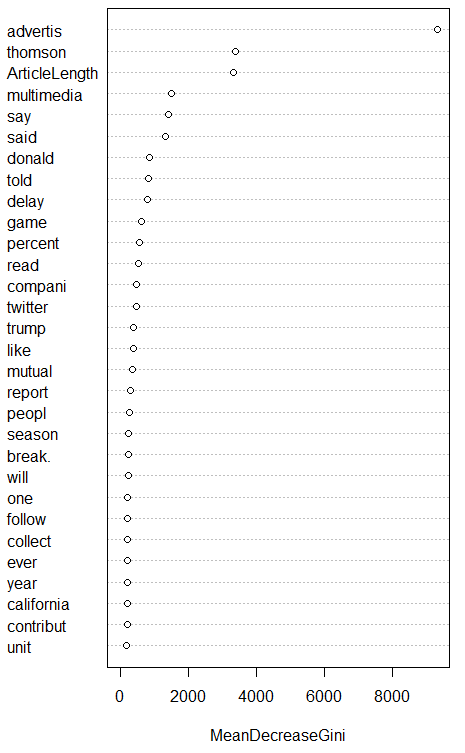


Figure 27

The confusion matrix has errors that demonstrate similarities between publishers’ word-usage. For example, an article from CNN is likely to be mistaken for the New York Times (understandably, as there are far more New York Times examples than CNN). Regarding word-choice, CNN and the New York Times are apparently similar. The Guardian and Verve have the same problem, perhaps because they are underrepresented compared to the Times.

Reuters was correctly identified 100% of the time with no false positives. Perhaps there was a telltale word missed in cleaning, or Reuters naturally has an eccentric vocabulary.

### Naïve Bayes

The default settings on the Naive Bayes model from R package “e1071” are quick to implement, but poor in output. The confusion matrix below shows only 32.2% of articles correctly sorted by publisher.

Table 20



The model is ineffective, but the tables of particular words are still insightful. See the figure below for five such tables regarding variables with intuitive meaning, their mean (aggregated by publisher), and their standard deviation (aggregated by publisher). The publishers are distinguishable by the length of their articles and their use of the words “trump” “obama” “republican” and “democrat” among others.

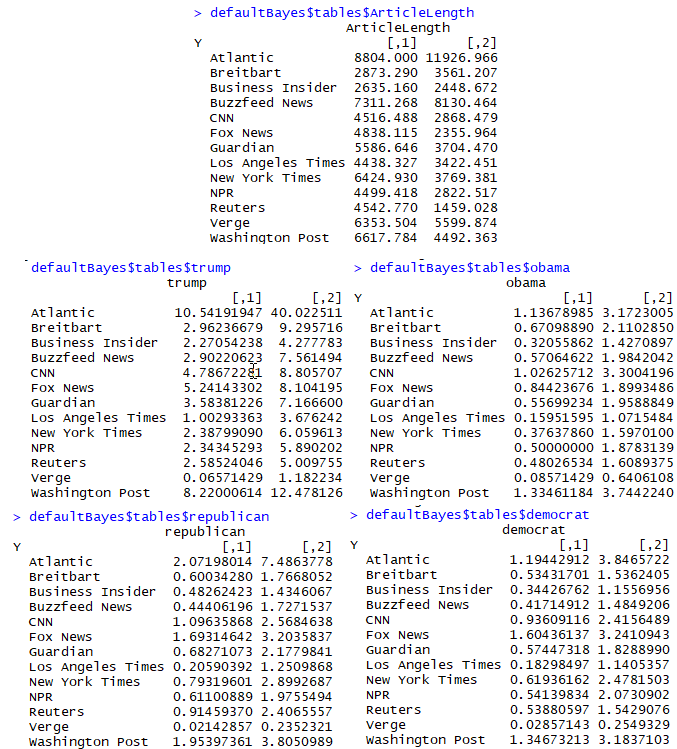


Figure 28

### Support Vector Machines

In the effort to develop a model that could classify articles by publication, Support Vector Machines showed promise. Three distinct approaches were undertaken, each based on different inputs. The first SVM model sought to differentiate between the lexicons of the different publications through the use of term-frequency analysis. The second model attempted the same task without the use of term frequency; instead, it utilized only article length, article sentiment, and whether the article was written before or after a momentous political event. The third model experimented with the efficacy of using a support vector machine trained on features selected by gain ratio to conduct binary classification on articles from a subset of two publications.

#### Term Frequency Model

To classify articles by their publication’s perceived lexicon, a document-term matrix was created for word counts. The counts were then normalized by the Term Frequency-Inverse Document Frequency method. These normalized frequency scores (without the article length variable used in the decision tree models) were used as inputs for this model.

Compute power and memory became serious issues with such a massive, multidimensional matrix. For this reason, the dataset used for the final training of the model was only a randomized 20% of the cleaned data. The remaining 80% was used for testing. Initial tuning on 5% and 10% training sets showed significantly lower accuracy, which strongly suggests that using more than 20% of the data would also have a positive effect on accuracy. After several iterations at comparable costs, the radial, sigmoid, and polynomial kernels were determined to be poor choices, and the linear kernel was chosen. A high cost of 100 was used to ensure significantly greater accuracy than those that were observed at lower costs. The high dimensionality of the data seems to have necessitated this.

#### Sentiment Model

The second model was different from the first not only through the inputs it used but through the scope of its application. Differentiating between 13 publications, many of which are similar in subject matter and political lean, was simply too unrealistic of a task for an SVM using only article length, article sentiment, and a binary-discretized date variable. Instead, three publications of varying political lean were chosen to train and test the model upon. Breitbart was chosen to represent the far right-wing, the New York Times was chosen to represent the left-wing, and Reuters was chosen to represent a centrist position. There is no accepted way to quantify political lean, but these three papers are solid choices to achieve a wide range of political motivation and dialogue. The inclusion of a centrist position also serves to challenge the SVM not simply to distinguish between left and right, but to also begin to understand politics as a more complex and subjective array than a binary choice. The political event that was chosen to discretize publish date to a binary before-and-after variable was the United States Federal Election of 2016, the results of which were announced on the night of November 8th, 2016.

Using the AFINN 111 sentiment dictionary, each word in each article was scored with an integer value between 5 and -5, representing positive and negative sentiment respectively. The total sentiment of the words in each article was summed. The bar chart below represents the average sentiment (normalized for article length) of all articles in each given publication. Clearly some publications use more positive words and some more negative words. Most fall relatively close to the median and fairly close to neutral. The rationale for any disparity in sentiment between publications may be explainable by the different segments run by each (various types of editorials and columns might be much lighter in tone than, say, serious news coverage or opinion political pieces). However, there is also the possibility that political lean itself plays a role. The very positive publications are generally left-leaning, while the right-leaning publications tend to be somewhat negative and staunchly centrist publications appear neutral or slightly negative. The sampling of articles from publications of different sides of the political spectrum is not what it could be, particularly due to shortcomings in the data’s collection process.

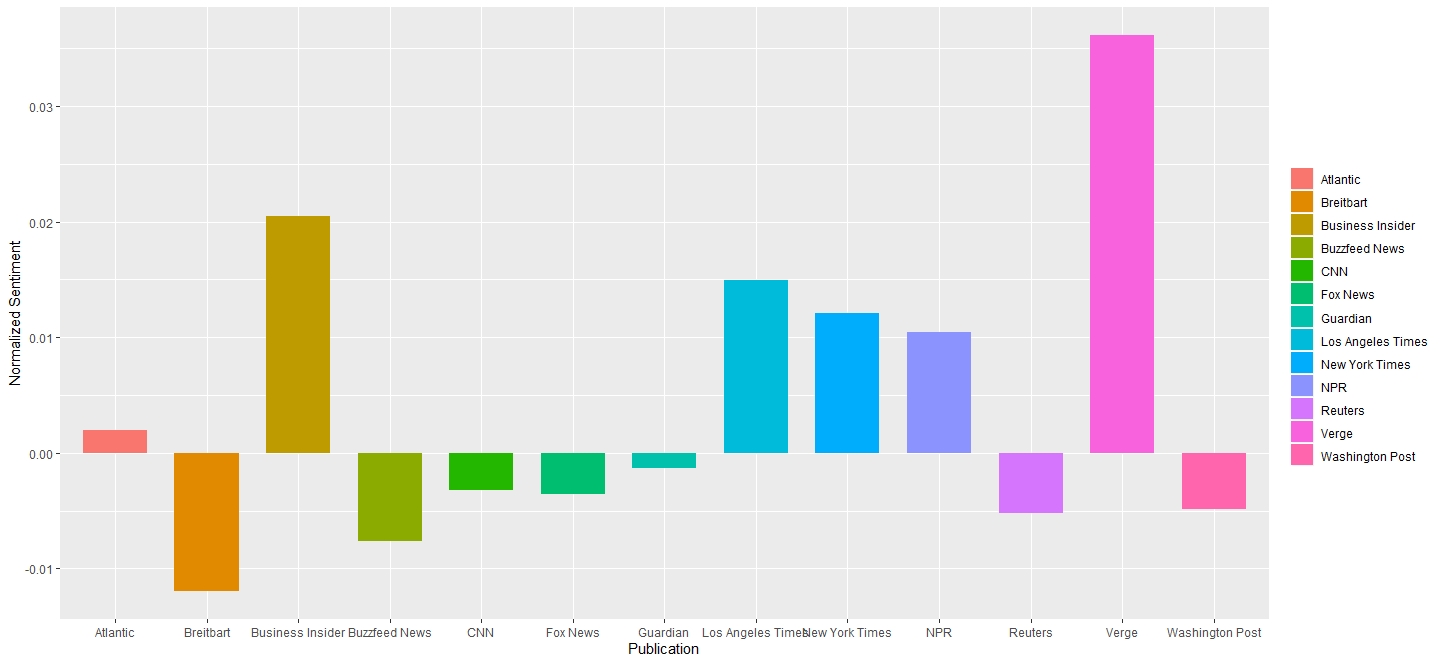


Figure 29

Since the vast preponderance of articles in the dataset date from 2014 to 2018, most of the articles were published (relatively) soon before or soon after the election of 2016, which is ideal for this sort of model. The grouped bar chart below shows the breakdown of mean normalized sentiment for the three publications selected for this model. Each publication’s sentiment is shown before and after the election. These results appear somewhat counterintuitive. The right-wing Republican party won the contentious presidential election and retained control of both houses of congress, yet Breitbart’s sentiment remained relatively unchanged. More surprising still is the left-leaning New York Times’ obvious increase in positive sentiment after the election. The centrist Reuters grew slightly more positive. These results seem very counterintuitive, as one might reasonably expect to see an increase in positivity after a victory for those a publication is politically aligned with or a corresponding increase in negativity after a loss. This data suggests that the lead-up to a contentious election is itself a more significant cause of negativity than the outcome thereof.

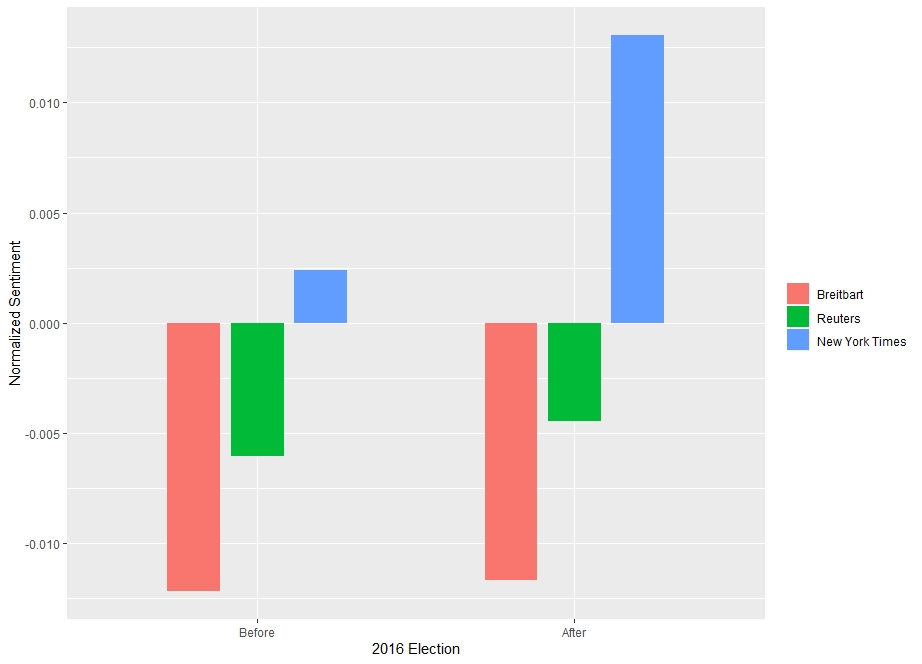


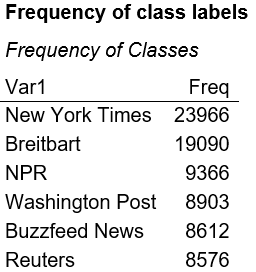
Figure 30

In contrast to the first SVM model, the best kernel for this analysis was found to be the radial kernel. On this data of radically smaller dimensionality, a lower cost of 10 proved effective without leading to much overfitting at all. Compute power and memory were not an issue, so 50% of the subsetted data was used for training, with 50% used for testing.

#### Binary Classification Experiment

The objective of this analysis is to demonstrate the capabilities of the support vector machine as a binary classifier. For this purpose, the binary classes will be the top two publications by the number of observations in the dataset, “New York Times” and “Breitbart”. Sentiment analysis performed in this report has shown these two publications to be very distinct in sentiment as well.

Table 21



NYT and Breitbart have a roughly even number of observations. The SVM will be a binary classifier that predicts these two classes. For modeling feasibility, 5,000 observations will be sampled from both classes.

Table 22

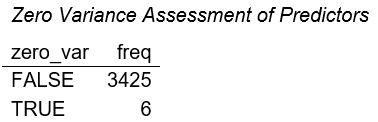
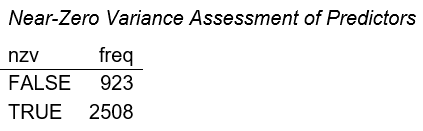


Table 23



Six variables displayed zero variance, and these were removed from the data.

#### Normalization and Scaling of features

Matrix was normalized by dividing values by their row sums. Scaling was done by centering and standardizing the values.

#### Chi-squared test for independence on discretized features

From the Chi-square test, there were 2,594 significant features.

#### Feature ranking using Gain Ratio



Figure 31

Table 24

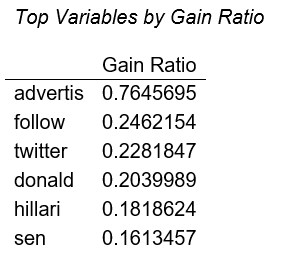
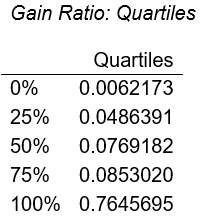


Table 25



#### Training and Test Split

70% of the data was used for training, while 30% used for testing. In the training set, there was an approximately equal distribution of classes.

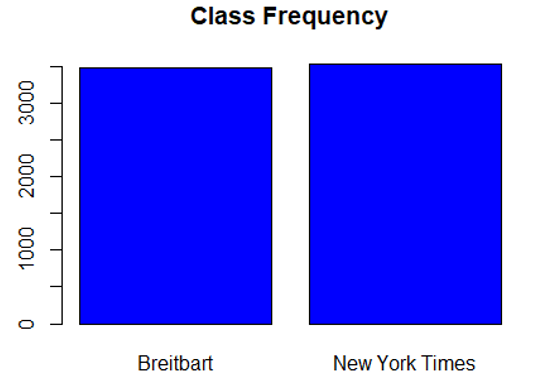
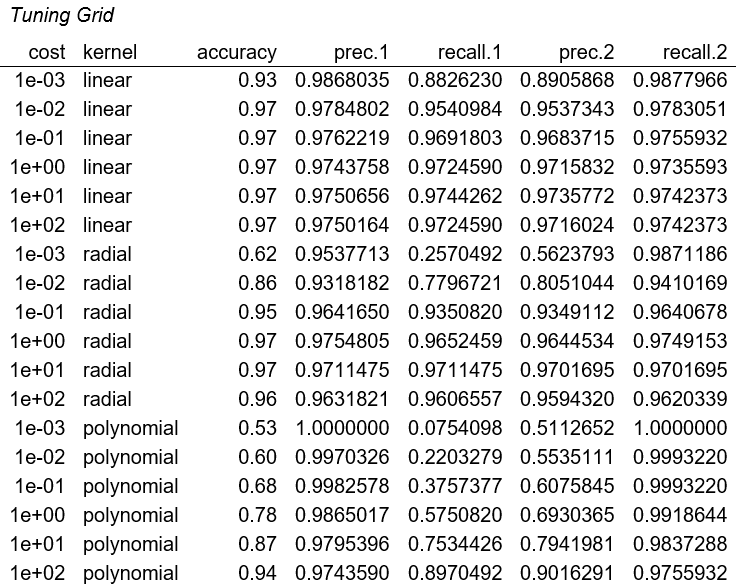


Figure 32

#### Modeling SVM: Binary Classifier

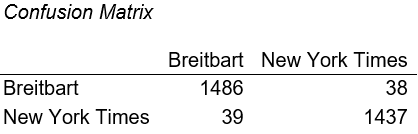
For the SVM, a tuning grid was used to evaluate the effect of parameter settings on prediction results. 62 top-ranked features (based on Gain Ratio) were used to train the model.

Table 26



The table above shows the best results were from specifying a cost parameter of 10 using a linear kernel and trained on a subset of 62 features with the highest gain ratio.

Table 27



The final model achieved an overall accuracy of 0.97.

## Results

### Association Rules

Prior to rule generation, 12 simulations of the apriori algorithm were performed with minimum support values ranging from 0.05 to 0.2 and minimum confidence values ranging from 0.8 to 1. Based on the results from simulation, an optimal combination of 0.05 for minimum support and 0.8 for minimum confidence was chosen by taking into account the number of rules generated, the mean size of the rules, and the mean quality of the rules. Based on these parameter specifications, an average quality of 0.06 for support, 0.84 for confidence, and 1.5 for lift could be expected in rule generation. Limitations with the apriori algorithm would not allow for reducing the minimum support threshold much further below 0.05. These simulations would cause a premature stopping of rule generation and would limit returning rules of substantial size and quality.

Results from non-targeted, holistic rule generation using the above parameters yielded a total of 43603 rules. The actual average quality came to 0.06 for support, 0.84 for confidence, and 1.5 for lift across all the rules. The most interesting rules demonstrated the highest lift and describe some of the most interesting news articles for 2018 Q1 overall. These included antecedents and consequents pertaining to Barack Obama, sexual harassment, House Republicans, Trump, and South Korea.

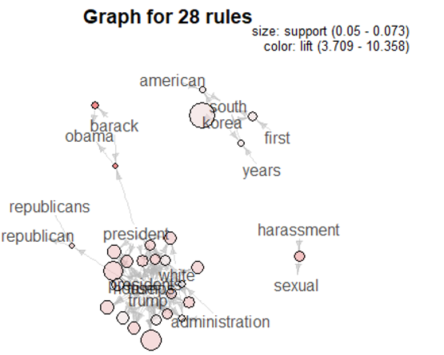


Figure 33

Results from targeted rule generation for Trump as the antecedent (LHS), using minimum support of 0.001 and confidence of 0.5, yielded 14 rules. The average size of the rules was on the order of k=2 rule itemsets. The average quality of the rules came to 16% for support (i.e. the proportion of the rule appearing in the data), 59% for confidence (i.e., the proportion of how often the rule was true), and 1.6 for lift (i.e., the importance of the rules). These rules were not that interesting and did little more than describe Trump in single terms. Setting Trump as the consequent (RHS), using slightly stricter support of 0.05 generated far more rules, on the order of 2084. However, the most important itemsets also did not yield anything particularly new and insightful for understanding the news of this period. The same items, in different combinations, would appear across the top 50 rules and be very repetitive and uninteresting, despite having an average itemset size of 4. This indicated that the news about Trump itself was mostly repetitive and displayed largely inter-mixed terms that were merely high-level associations with a person who is president.

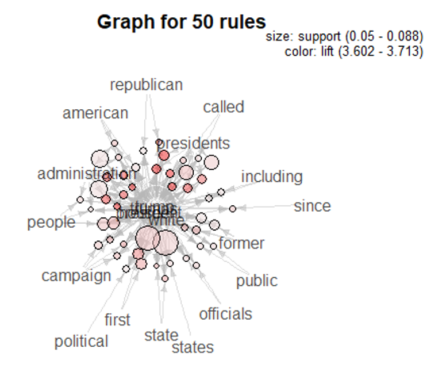


Figure 34

### K-Means Clustering

The first K-Means model created one cluster containing Breitbart, Fox News, Reuters, CNN, the Guardian, NPR, the LA Times, and Business Insider, and another cluster containing the Atlantic, Buzzfeed News, Washington Post, the New York Times, and Verge. The second cluster is composed of generally left-leaning and investigative publications. The first cluster, by contrast, has a composition far more right-leaning and centrist. It is interesting to note the inclusion of a publication like Reuters, known for generally objective and unbiased reporting, in the same cluster as a publication like Breitbart, known for the opposite. However, they are still an appreciable distance from one another in the graph space.

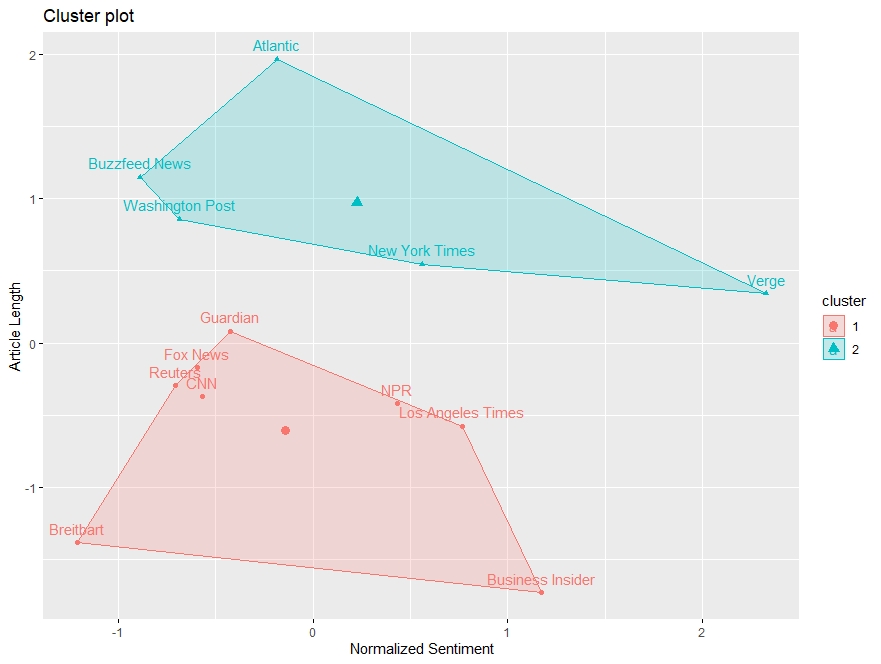


Figure 35

In the second clustering model, the clusters were identical to the first. As can be seen in the 3-dimensional images below (both of the same model) the same clusters were still fairly close together even after the addition of normalized vocabulary standard deviation as a variable. This similarity between models suggests that this measure as well serves to distinguish between these two groups of publications. However, it is worth noting that as average article length increase, normalized vocabulary standard deviation generally does as well. It is possible that a relationship between the two variables exists, and therefore it would not be surprising that the addition of one to a model containing the other would not change the clustering composition.

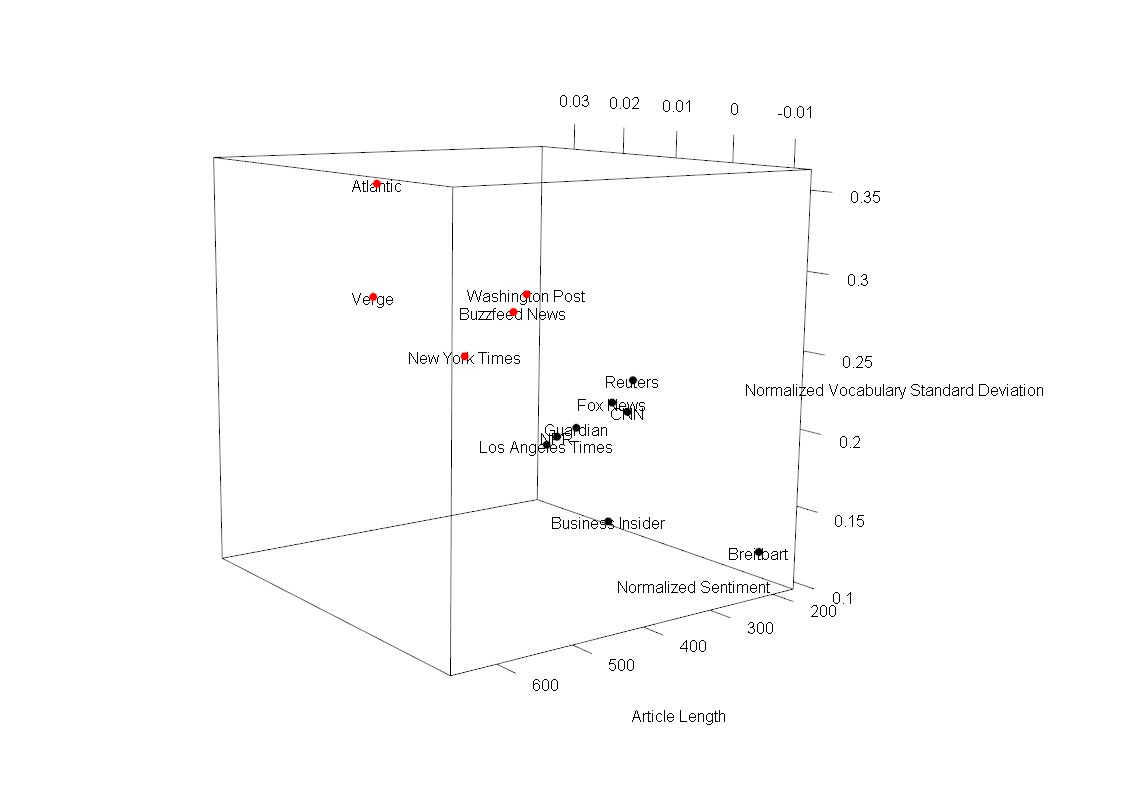


Figure 36

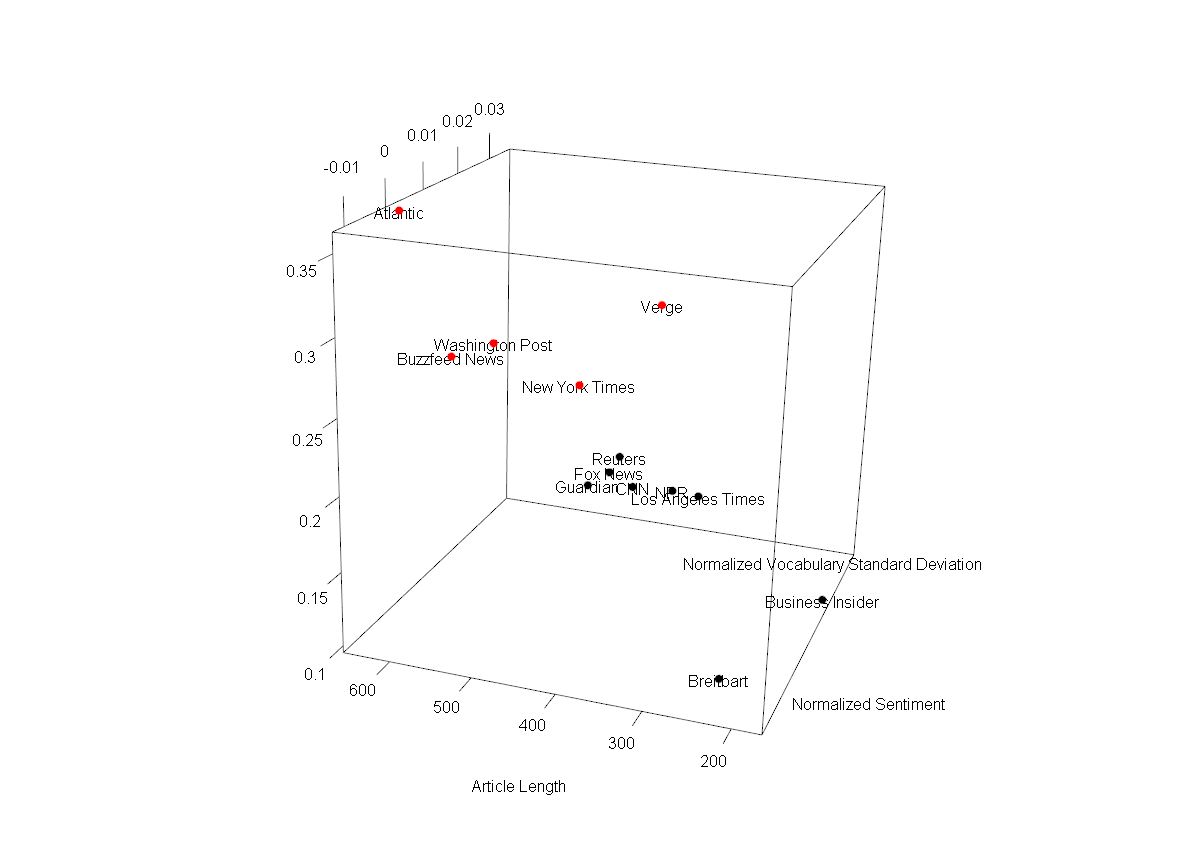


Figure 37

### Hierarchical Clustering

Hierarchical clustering on this dataset was very computationally expensive. With about 110,000 articles in the cleaned and transformed dataset, approximately 6 billion distances would have to be calculated – with R using 8 bytes per distance, this would require over 48 GB of memory to work with. Sampling is required to make it feasible to generate the hierarchical clusters object with standard computing resources. Even when sampling only 1% of the data, the dendrogram is too complex for any useful analysis, as can be seen in the figure below.

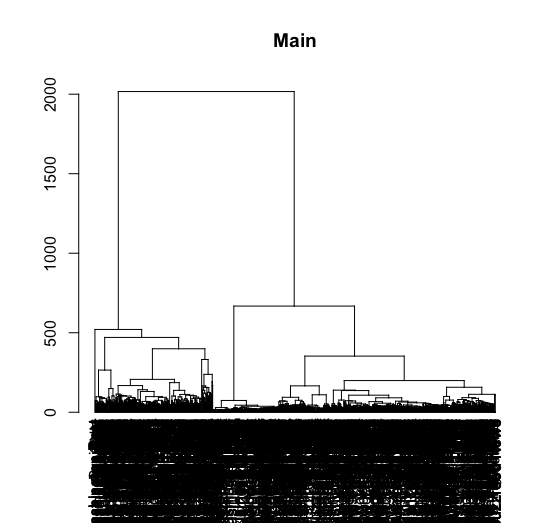


Figure 38

When reducing the sample size to only 0.5% for analysis, it is more feasible to analyze the clusters. As can be seen in the following figure, focusing on an individual branch of the dendrogram shows that with the sampled documents Reuters articles greatly outnumber articles from other publications. It can be concluded that based on words used in the sampled articles, articles’ publications may be inferred by their content.

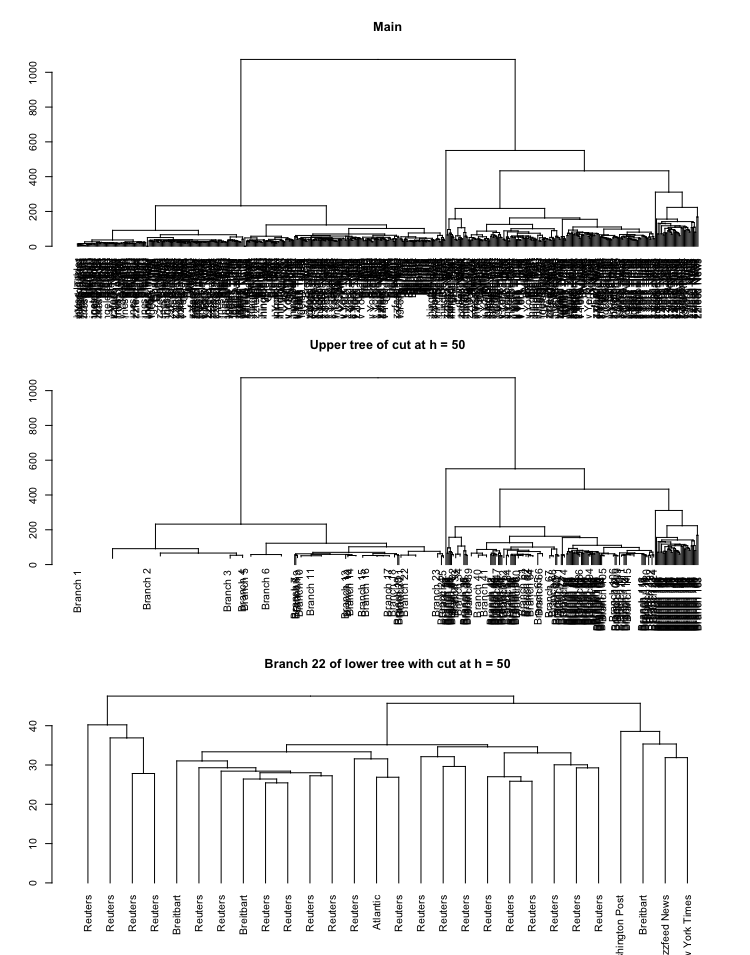


Figure 39

### Decision Trees

Decision trees and the ensemble random forests are excellent at detecting which publisher produced an article using only the article’s length and the words included. While a human might be able to guess the publisher based on the article’s content and context, it is unreasonable to expect a human to have half the accuracy of the forest.

These models also provide insight into the most insightful. With this a human might see broad patterns that distinguish one publisher from the others, using the model to enhance their own understanding.

### Naive Bayes

Naive Bayes was hardly accurate at all in sorting articles by publisher, but the tables associated with the model are still worth considering. For example, an article with over 20,000 characters is almost certainly from The Atlantic, based on its average length and standard deviation.

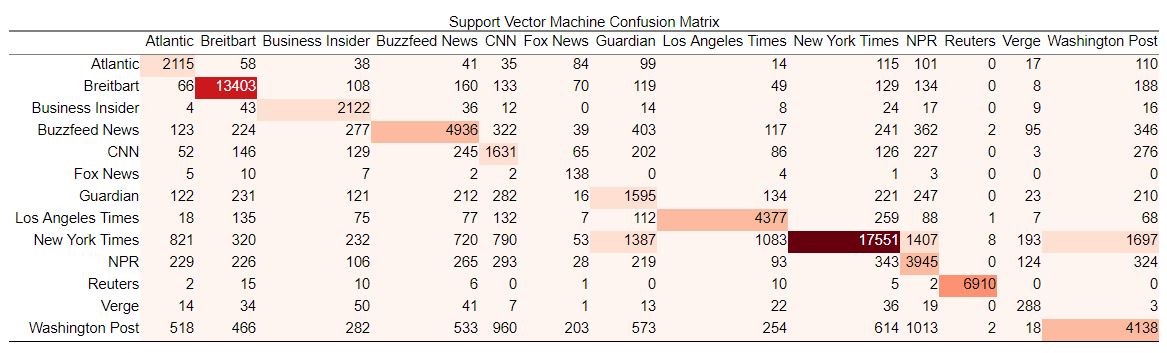
### Support Vector Machines

#### Term Frequency Model

The model produced a 71.08% testing accuracy. Although this was not an overtly impressive result, it was promising considering the classification was between 13 different levels. The training accuracy was 87.12%, suggesting high overfitting. Efforts to curb the overfitting, however, led invariably to significant reductions in testing accuracy. Future models executed on stronger computers may show that using a larger training set may increase testing accuracy while closing the gap between testing and training accuracy.

The confusion matrix below shows the predicted publication on the vertical axis and the actual publication on the horizontal axis. The largest clusters of error occurred in false predictions of the New York Times. This is most likely due to the sheer number of articles by the New York Times found in the dataset, but also may indicate a similarity between the lexicon of the New York Times and other publications like NPR, The Guardian, and The Washington Post. The political leanings and formats of these publications make their confusion with the New York Times understandable. Politics seemed to play a role in other areas of error or lack thereof, as there was very little misclassification of the right-leaning Breitbart with any of the left-leaning publications in the dataset.

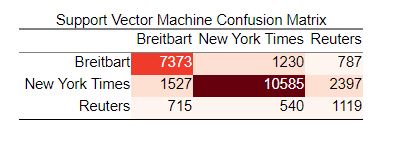
Table 28



#### Sentiment Model

This model returned a 72.61% testing accuracy. The training accuracy was 73.47%, suggesting only minimal overfitting. The confusion matrix below shows that the model had the most trouble classifying Reuters articles. Although this shortcoming is intuitive given the centrist position of the publication compared to the other two publications, it is noteworthy because it suggests that political lean is having a tangible impact on the model’s ability to discern between the publications. Overall, it seems that this simplistic model was reasonably effective at discerning between publications when the political leanings of these publications were very different.

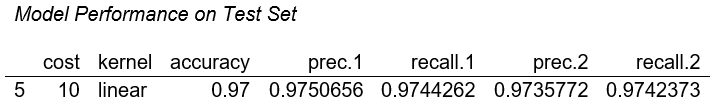
Table 29



**Binary Classification Experiment**

The support vector machine model showed promising results as a binary classifier of the publication label. In this experiment, the top 2 publications (i.e., New York Times and Breitbart) were sampled down to 10,000 observations and used for training. During pre-processing, a small subset of zero-variance predictors was removed from the data, and features were normalized and scaled. In the assessment for important variables using gain ratio, the term “advertis” had the highest value of 0.75, which represents the ratio of the information gain to split info (i.e., the number of branches generated). Overall, 50% of the predictors had a gain ratio of less than 0.08. The proportion of data used for training was 70% with a roughly equal representation of classes, and only the top 62 features (based on gain ratio) were kept. A tuning grid was used to look across different kernels and cost of constraints, and the best results came from a linear kernel with cost = 10. The resulting performance of these parameters are displayed below:

Table 30



### Text Mining: Sentiment

Article sentiment was measured across the years 2014 - 2018. The overall raw sentiment of news articles appeared to dip in 2016 and spike in 2017. However, the mean normalized sentiment score (i.e., using min-max) showed a different picture by displaying a spike in average sentiment from 2014 to 2015 and a downward trend from 2016 to 2018. In 2014, the top terms that had a high impact on sentiment included “death”, “gift”, “penalty”, “defeated”, and “big”. In 2015, the top terms included “applause”, “flu”, “want”, “thank”, and “like”. In 2016, the top terms included “sentence”, “prison”, and “grant”. In 2017, the top terms included “thank”, “applause”, “yes”, “like”, and “want”. In 2018, the top terms that impacted sentiment included “best”, “drag”, “anger”, “safety”, “gun”, “cancer”, and “like”.

Raw sentiment by publication displayed overall that New York Times had the highest positive sentiment in their articles. Breitbard and Business Insider had the lowest negative sentiment. In contrast, the mean normalized sentiment score by publication displayed on average that Atlantic, Breitbart, and Verge had the highest positive sentiment. At the lower end of sentiment were Business Insider, Guardian, and NPR. Some results may appear to be contradictory to one another, but the average score is sensitive to outliers that can sway the score in either direction. These outliers can include very low (close to 0) or very high (close to 1) normalized sentiment scores.

# Conclusions

Articles, publications, sentiment, and high-level topics and themes can help readers to understand and effectively navigate through a sea of news stories. It is clear from the results that publications have a distinct voice by which consumers of the news can use as a “yardstick” to distinguish credible from non-credible news sources. Although the purpose of this report was not to discredit certain news stories or statistically validate in-depth sentiment, it does, in fact, show that at a high-level some key messages do emerge out of this repository of articles. The preliminary steps of finding associations in news topics, grouping articles based on their content, and the identification of publication are necessary strides towards bringing objective and meaningful insight to a population overloaded with information. To those who find this information valuable, better means and mediums for recommending news content can be developed through methods such as these, so that information about current events are neither obfuscated or misrepresented but delivered in a consistent, yet robust way.

The exploration in this report revealed that some stories that are typically sensationalized in the news are not all that interesting. In particular, news related to President Trump was repetitive and mired within the same framework of political themes that go back and forth on similar terms and topics. It is every once in a while that unique stories stand out from the noise of politics, such as topics about the Me Too movement tackling the issues of sexual harassment. The overall positive and negative sentiment in the news appeared balanced for the most part, without it being overly negative as some might expect. Overall trends of positive and negative news appeared balanced before and after the 2016 election, but it tells quite a different story at the publication level. In some publications, positive news stories became more prevalent over time. Publications have different stories to tell, measured by the words they use and the message conveyed. Some are more like-minded in their tone and cadence than others. “Words offer a great many opportunities for discrimination; there are so many of them. Some vary considerably in their rates of use from one paper to another by the same author; others show remarkable stability within an author.”[[7]](#footnote-7)

“In an age in which the digital attention economy is shoveling more and more clickbait… and fragmenting… focus into emotionally charged shards, the right response is to become more mindful in… media consumption.”[[8]](#footnote-8) On an individual level, this is hard to do. An effective way is to deal with and understand things in the aggregate. This report addressed the reality faced by millions of readers of the news, the many outlets and mediums through which people access the news, and concerns over validity. People want to know where they should be getting their information, but in most cases, a conscious decision is not made about each site and newsfeed visited. In the digital attention economy, media organizations can place more considerable attention towards obtaining more views and clicks than ensuring thoughtful and accurate reporting for cultivating an informed audience. Now and then it is good to take a step back and look at the big picture, go above the noise, and understand the overarching themes.

1. https://www.businessinsider.com/these-6-corporations-control-90-of-the-media-in-america-2012-6 [↑](#footnote-ref-1)
2. https://www.ssim.ac.in/blog/how-social-media-impact-on-newspapers/ [↑](#footnote-ref-2)
3. https://www.wired.com/story/free-speech-issue-tech-turmoil-new-censorship/?CNDID=50121752 [↑](#footnote-ref-3)
4. https://www.vox.com/2018/4/18/17252410/jordan-peele-obama-deepfake-buzzfeed [↑](#footnote-ref-4)
5. https://www.kaggle.com/snapcrack/all-the-news [↑](#footnote-ref-5)
6. https://components.one/datasets/all-the-news-articles-dataset/ [↑](#footnote-ref-6)
7. Inference in an Authorship Problem, Mosteller and Wallace, Journal of the American Statistical Association, 1963. http://ptrckprry.com/course/ssd/reading/Most63.pdf [↑](#footnote-ref-7)
8. https://ideas.ted.com/love-the-news-but-hate-clickbait-and-fluff-how-to-be-a-slower-and-more-thoughtful-news-  
   consumer/ [↑](#footnote-ref-8)