Final Project Report

Karim Zakir

4/23/2023

Introduction to the Project

Reading the news, whether in print or digital forms, has become a daily ritual for over 3 billion people (Admin 2019). The wide variety of news outlets help us stay up-to-date with the current events and provide us with a significant outlook on today's issues and the world in general. To provide its readers with the most accurate outlook, the news should be reported as neutrally as possible, ensuring that no bias occurs in the reporting. It is especially important that the writer does not let their personal biases impact their writing or how they report on different topics, unless it is an opinion piece. As the news has grown to be a global practice, however, certain news outlets have become very closely aligned with particular political parties and have allowed these alignments dictate how they report the news and what kind of topics they report on. This allows them to target specific groups of people who align themselves with the same or similar political parties or have the same political views as the news outlet.

The two most famous examples of this are CNN and Fox News in the United States (U.S.). CNN is considered to strongly align with the U.S. Democratic Party, while Fox News aligns more closely aligns with the Republican Party. These political beliefs strongly influence how these news outlets report on different topics, which, naturally, attracts different kinds of audiences. It could be argued that this kind of approach to news-making has possibly created a political divide in the country, since the two audiences have different and skewed viewpoints of the country and the world in general.

A newspaper that could possibly align themselves politically is the New York Times (NYT). Based in New York, U.S., the newspaper is one of the biggest news outlets in the world. It is considered to be a trustworthy source and is read by a large population. However, does the New York Times and its writers let their political beliefs influence how they write their articles? For instance, if in a given year, the Democratic party holds the most power in the country, do the NYT writers report more negative or positive news?

To investigate this question, I will consider all of the news headlines written in the NYT between 2013 and 2020 inclusive. Since it is hard to decide which political party had the most power in any given year, I will use the political party of that year's president as a proxy for that metric. Between 2013 and 2016, Barack Obama, who is a member of the Democratic Party, was the president of the U.S., so those years would be considered as the years when the Democrats had the most power in the country. Between 2017 and 2020, Donald Trump, who is a member of the Republican Party, was the president of the U.S., so those years would be considered as the years when the Republicans held the most power (Britannica, n.d.).

Throughout this investigation, I will answer the research question "does the NYT change how they write their headlines based on the party of the current United States president?" In particular, do NYT writers include more negative words in their headlines when the president of the U.S. belongs to a Democratic or Republican party? Similarly, when does the news outlet use more positive words in its headlines? I chose to focus on the headlines of these articles, since those are the first pieces of the articles that the reader sees and is, most-likely, what the reader skims when reading the news, so the reader's first impressions of the article and the current affairs are based on these headlines.

The dataset I will be using to answer these questions contains the headlines of all of the NYT articles written between 2013 and 2020 (inclusive), as well as other information, like their publication date, section name, and abstract.

Methods

Acquiring the dataset

The dataset was acquired using the publicly-available NYT API. In particular, I used the archive feature of their API, which returns an array of newspaper's articles for a given month and year combination. For this particular investigation, I have decided to focus on Obama's and Trump's presidencies, so I focused on articles published between 2013 and 2016 in order to analyze the sentiment during Democratic presidency, and on articles published between 2017 and 2020 to analyze the sentiment during Republican presidency. This would allow me to focus on the two different parties and their presidencies, while also having a compact dataset.

After retrieving the data for the articles for the given years and all their months, I extracted specific information from the API's results, such as the article's title, abstract, publication date, document type, what news desk wrote the article, its section, and, finally, its word count.

After obtaining the data, I pre-processed it by tokenizing each headline. Additionally, to ensure that the data we extracted provided us with useful insight into our questions, I have removed all of the stop-words listed in the tidyverse's stop-words from the data using Reg-ex. Furthermore, certain apostrophe's in the dataset were denoted using a different symbol than what is used in the tidyverse package. As such, all of those apostrophes were changed to match the tidyverse package.

Finally, it was also important to determine the sentiment of each word used in the headline. To do so, I have used bing's sentiment lexicon, which I retrieved using the tidytext package. It contains a large set of words that were deemed to either have positive or negative connotations or emotions.

Analysis

The dataset I retrieved contained information on 552669 articles, of which 58.75% were written during President Obama's time and 41.25% were written during President Trump's time. It is not clear outright why there's a difference in the amount of articles written during the two administrations; it is hard to argue that there were more news-worthy items in the early-mid 2010's than in mid-late 2010's.

Since we are focusing more closely on the headlines, I decided to look at the word and character counts of the overall headline variable, which includes both presidencies. This would provide us with some summary statistics, which could help us understand the headlines a bit better, and also allow us to check if there were any possible data errors when using the API.

Table 1: NYT Headlines Summary Statistics

Statistic	Min	Pctl(25)	Median	Pctl(75)	Max	Mean
Word Count	1	6	8	10	38	8.086
Word Count (No Stop-Words)	1	5	7	9	24	6.888
Char. Count	2	35	50	62	638	48.854
Char. Count (No Stop-Words)	2	32	45	57	638	44.993

Most of the headlines contain about 10 words, which results in 57 characters on average, including the stop-words. Since the article headlines need to be short, it is common practice for writers to omit articles, which are a subset of the English stop-words; hence, the difference between the word counts with and without the stop-words isn't large. The headlines with a single word did stand out to me, but after looking closer at them, I realized they were outliers or data errors. Those articles simply just contained a single word for its headline, such as "Par-TAY!" or "Jungleland." Similarly, the longer headlines were not data recording errors. This means we do not need to remove any observations from the data.

The articles with only a single character in its headlines raised some concerns as well. There was only one such article with just a dot in its headline. There was most-likely an error when retrieving the data or perhaps an error with the API, so that article has been removed from further analysis.

Another piece of insight we can gather from table 1 is that most of the article headlines are between 6-10 words, which means the writers are most-likely deliberate when selecting which words to use in their headlines, because they are quite limited. This suggests that they would be incredibly thoughtful when picking their language, suggesting that if there is any bias, it is most likely purposeful, further highlighting the importance of this analysis.

Although the overall headline summary statistics allowed us to check for any data errors, it did not help us answer our research question. As such, let's see if the headline word count distribution changed between the two presidencies.

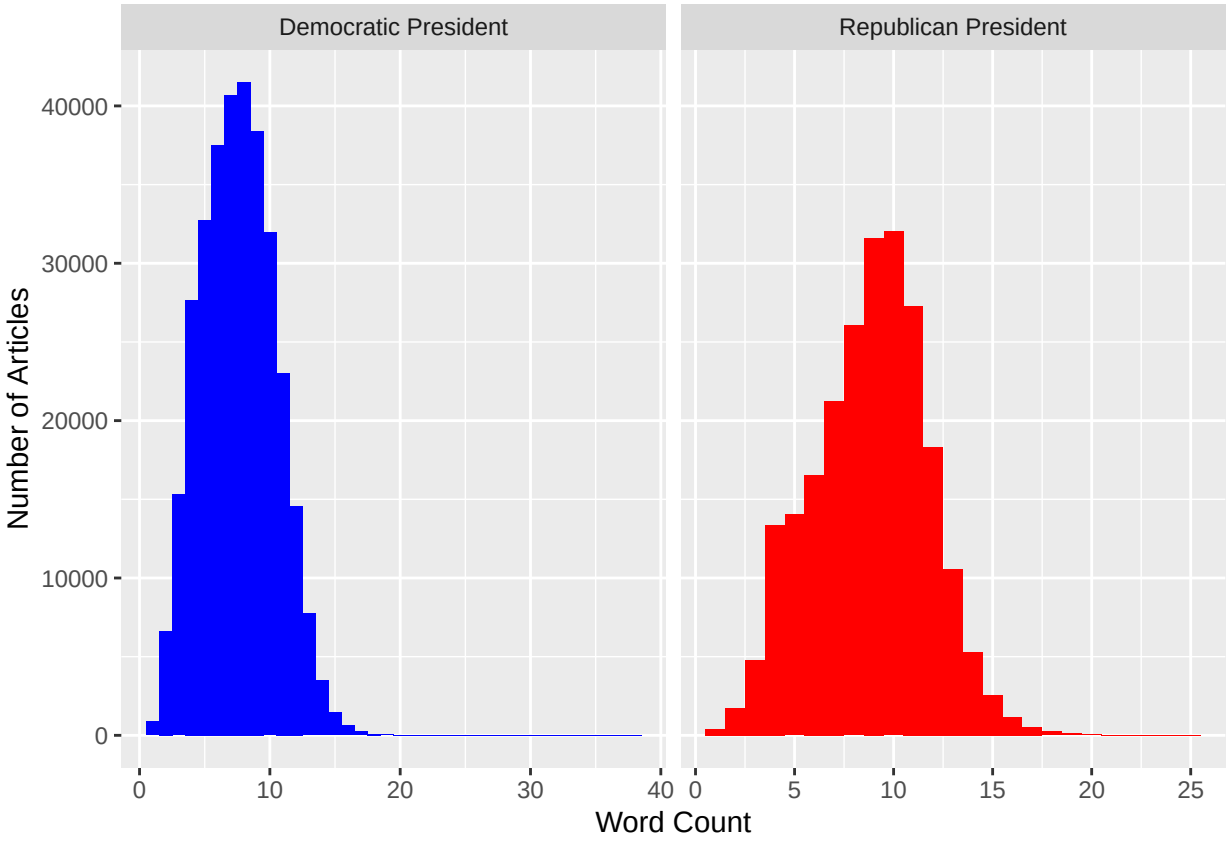

Figure 1: Headline Word Count Distribution During Democratic and Republican Presidents

In figure 1, we see that, although both distributions are uni-modal, the mean/mode of the right distribution is located closer towards 10 words, while the left distribution's peak hovers around 7-8 words, suggesting that the headlines written during Trump's presidency tended to be longer. This is further supported by the larger right tail on the right distribution. Overall, this means that there is some kind of a difference in how the headlines were written between the two presidencies, meaning the research question should be explored further.

Before diving into sentiment analysis, I first decided to look into what kind of words in general were used in headlines during each presidency using two word-clouds. This would help notice any possible errors or ambiguous words early-on.

Figures 2 & 3 highlight a few important insights about our data. The first insight we notice is how frequently "Trump" was used in relation to other words, based on the sizes of the words in the clouds, in both the clouds. Indeed, in figure 2, the cloud generated using headlines written between 2013 and 2016, the word "trump" was fairly prominent, even more prominent than the word "obama." This suggests that Donald

Figure 2: 50 Most Used Words during President Obama's Presidency in NYT Headlines

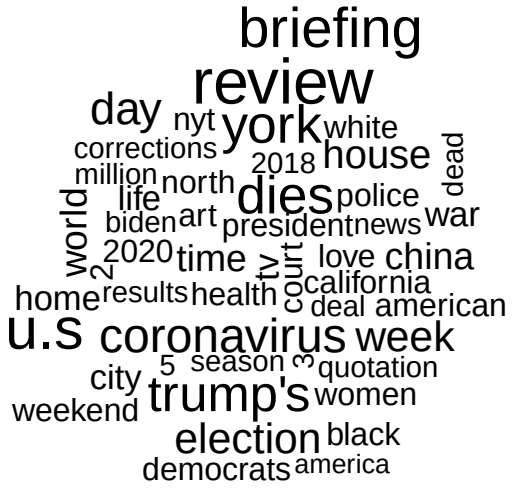

Figure 3: 50 Most Used Words during President Trump's Presidency in NYT Headlines

Trump generated a lot of news at the time and was perhaps a personality that drove engagement the most at the time. Indeed, in figure 3, the word "trump" further dominated the headlines, strongly outshining other words in the wordcloud, reinforcing the idea that Trump was a highly discussed and highly reported-on personality. This shows that NYT would report more on Trump's administration and campaign between 2015 and 2020, suggesting that there was a difference in the topics that NYT would report on at the time.

Another important consideration we need to make is that "Trump" in the context of NYT and U.S politics most-likely means the last name of U.S. president Donald Trump; however, the Bing semantic lexicon considers "trump" to be a positive word, since it also means "to win." This kind of confusion could make the analysis results inaccurate, so it was important to effectively differentiate between the two words. I differentiated the two words based on their capitalization: if "Trump" was capitalized, I would treat it as a last name, otherwise, I would treat it as the verb "to win." After drawing this distinction, I discovered that no article headline contained the verb meaning of the word "trump." Perhaps, this was by design to avoid any confusion in article headlines. Regardless, in further semantic analysis, the token "trump" will be filtered out, since it is a proper noun in all cases.

Secondly, the meaning/semantics of some of the commonly used words, like "Trump," "China," or "U.S." suggest that a majority of news headlines are related to the U.S. or World News. In the dataset, however, there are 72 unique categories, otherwise known as section names, which indicate what general theme an article is about. These range from U.S. politics to fashion to charity. The wide variety of categories suggest that these articles are written by different writers and target different readers, which means that there could be a difference in how positive and negative words are used in article headlines across these categories. To explore this further, I have decided to look into the 10 most popular sections between 2013 and 2020.

Table 2: 10 Most Popular Sections in NYT (2013-2020)

Section Name	Amount of Articles in the Section
U.S.	66772
Opinion	58224
World	57176
Arts	48285
Business Day	46388
Sports	37738
New York	35205
Fashion & Style	22883
Books	15868
Movies	13240

Table 2 shows that most headlines belong to U.S. or World news, but other categories, like Arts or Opinion articles, are also fairly prominent, meaning that they are also important to analyze and investigate. Throughout the rest of this analysis, I will focus on the top 5 categories: U.S., Opinion, World, Arts, and Business Day; these categories represent over 50% of all articles between 2013 and 2020, and they are most-likely to be impacted by U.S. politics, so these categories are the most appropriate for our analysis. To ensure that those categories do not contain any confusing words as well, such as "trump," it is important to look at the most common words in those categories.

Similarly to our previous findings, figure 4 shows that the news cycle has been reporting incredibly often on Donald Trump, with his name being in the top 10 most common words in four out of five top sections in the NYT. The only category which doesn't have "trump" as one of the most common words is the Arts section. Based on some expectations and figure 3, Arts seems fairly unaffected by U.S. politics, so it will be interesting to see how U.S. presidencies have affected this section later, when compared to other sections.

Finally, let's now look more closely into the semantics of the words used in NYT headlines. In particular, let's start by looking at how often, on average, positive and negative words appear in each headline as a proportion of the headline's word count. This will give us a general sense of how often words with positive or

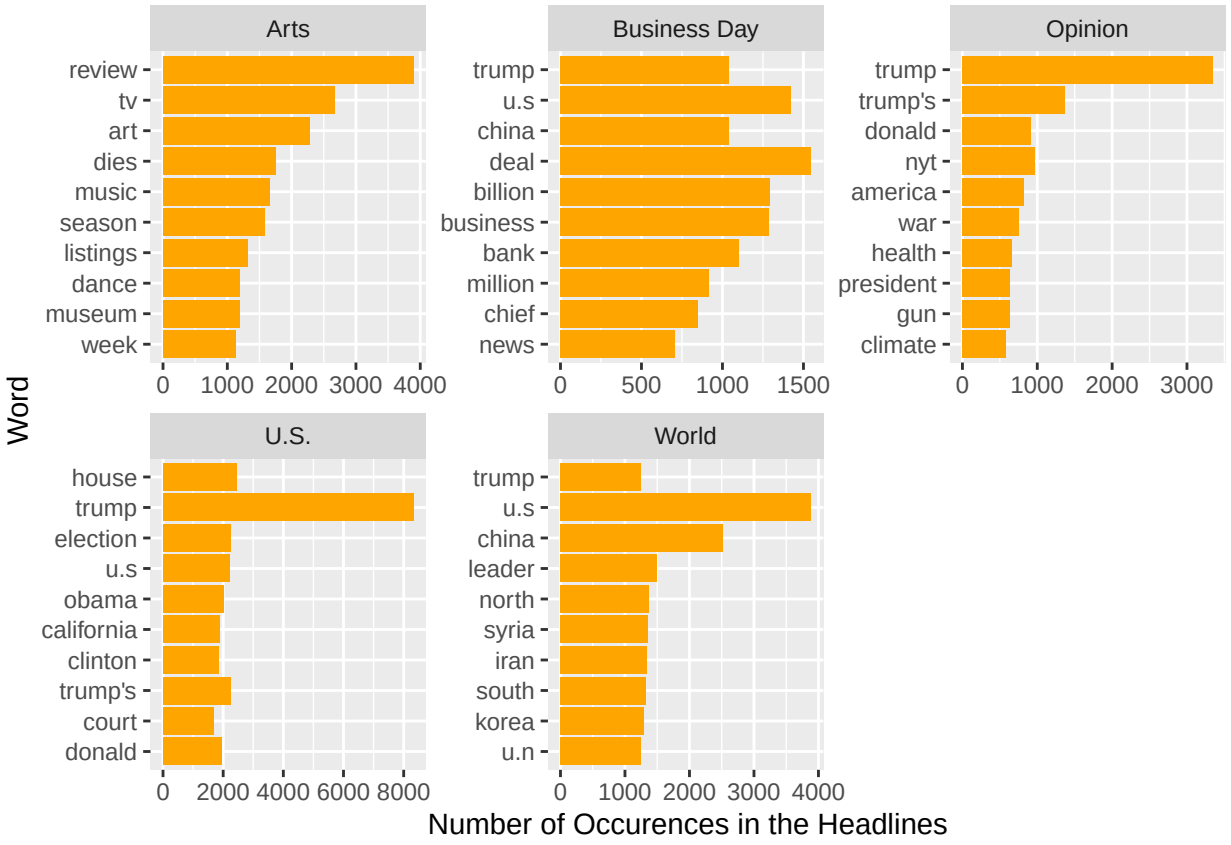

Figure 4: Top 10 Most Common Headline Words Across 5 Most Popular Categories in NYT between 2013 and 2020

negative connotations appear in headlines and how that usage varies across sections. Additionally, by using proportions, we will be able to avoid the impact of the changing word count length that we saw in figure 1 on our analysis.

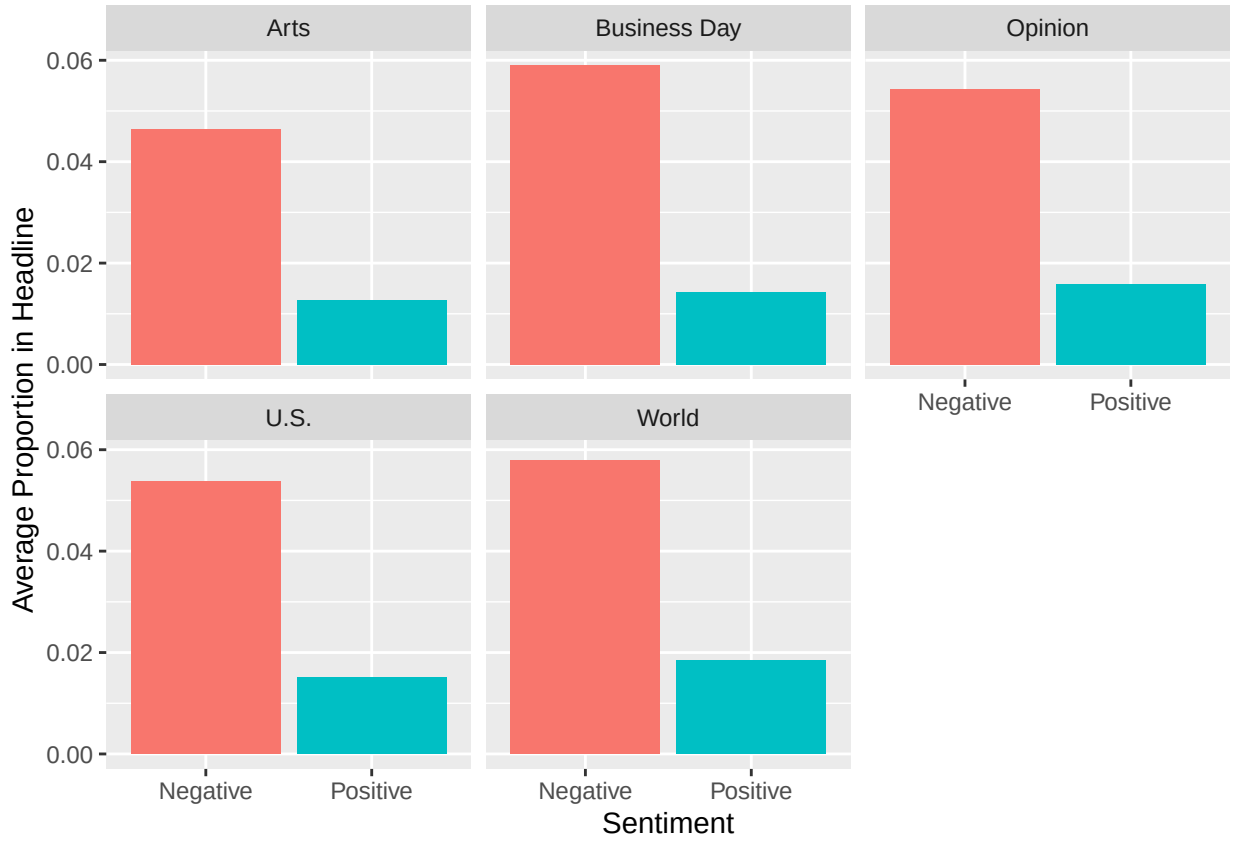

Figure 5: Average Proportion of Positive/Negative Words in NYT Headlines between 2013 and 2016

Firstly, a prominent insight that we can gather from figure 5 is that during Obama's presidency across all categories, words with a positive sentiment are used less than half as much as words with a negative sentiment, suggesting the news has a heavier focus on more negative topics, which is fairly expected, since news are most often used to highlight concerns. Additionally, we can see that certain categories like "Business Day" had more negative words than others, like "Arts," suggesting a difference in how different categories get reported or how their headlines are written.

Figure 6 supports the insights we have gained above. As seen in figure 5, during Trump's presidency, positive words were used less than half as much as negative words, with certain categories having more negative words on average in general. Finally, we can also see differences across the presidencies in the usage of positive and negative words. For instance, if we look at the U.S. section in figures 5 and 6, we can see that during Obama's presidency, negative words would make up about 5% of the headlines on average; this number increased to over 6% during Trump's presidency, suggesting a larger negative sentiment during the time when the president was the member of the Republican party.

It's not clear, however, from figure 3 whether these differences are significant or not. As such, it is important to conduct a statistical test to verify whether certain categories use more negative words than others. Furthermore, it is interesting to check the same for positive words, although it seems that for all sections, positive words were used similarly in frequency during the two presidencies.

The statistical test that I will be using is the Welch Two Sample t-test. This test aims to compare the

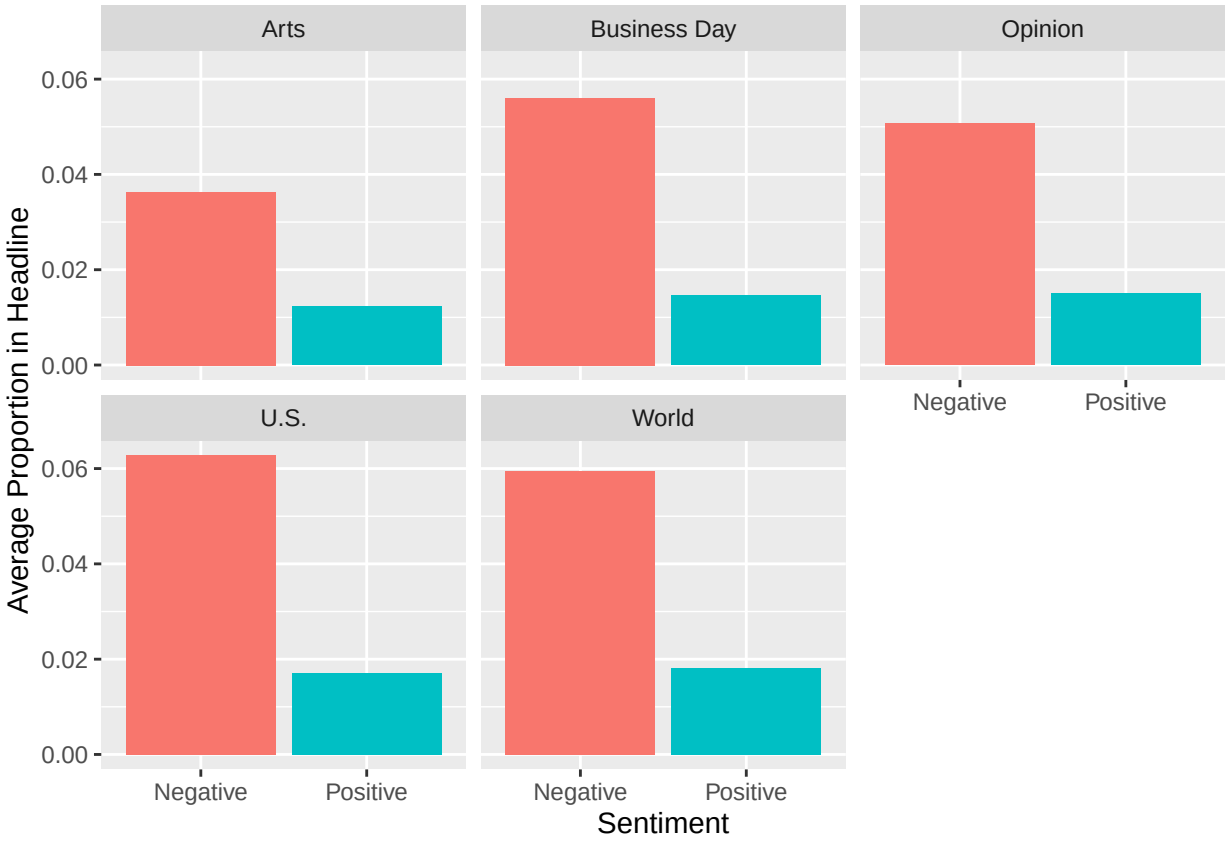

Figure 6: Average Proportion of Positive/Negative Words in NYT Headlines between 2017 and 2020

means of two independent groups. In our case, the two independent groups will be the articles written during Obama's presidency and the articles written during Trump's presidency. The reason that I am using this specific variation of the t-test is because the two groups do not have the same variance, meaning I have to use the test that adjusts for this fact. To analyze the results of these tests, I will plot the point estimate and the confidence interval for these differences and see if 0 falls within the confidence intervals.

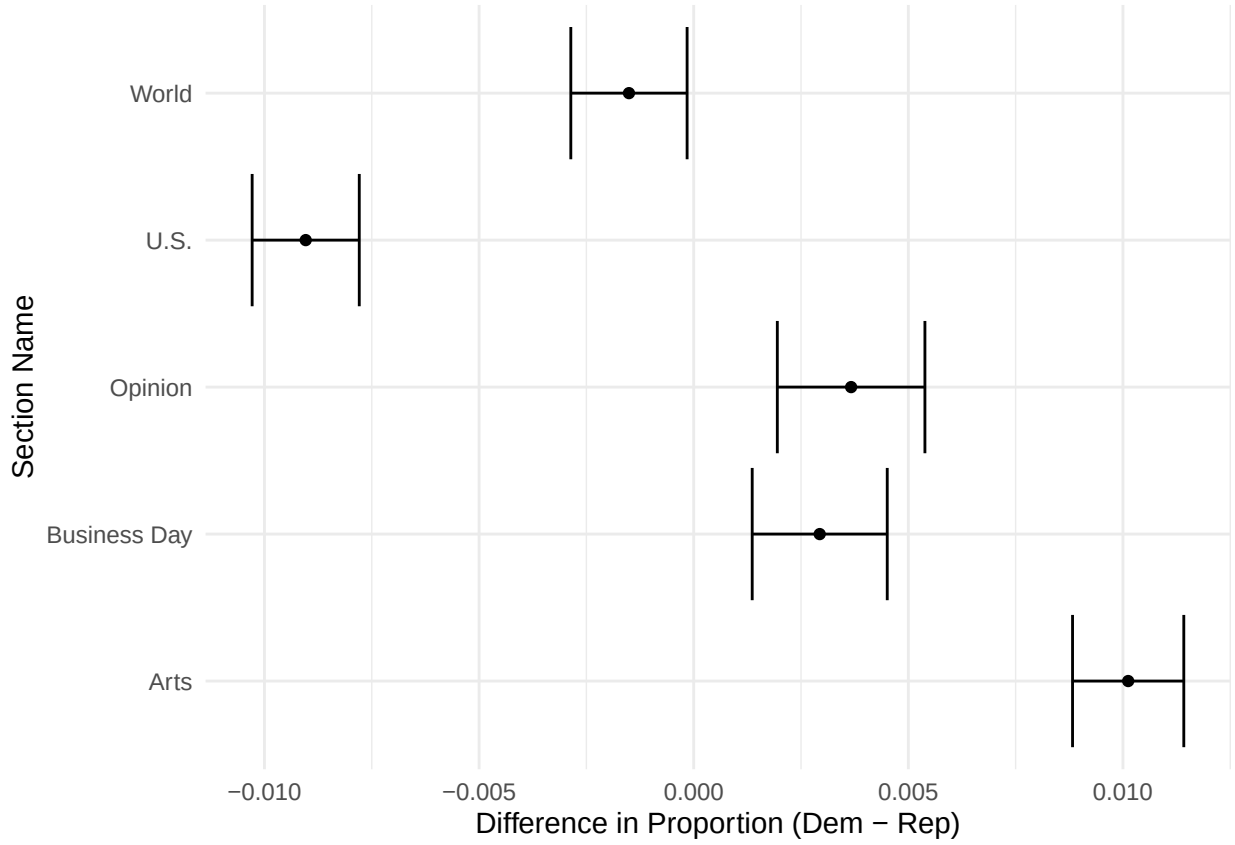

Figure 7: Difference in Negative Word Usage in NYT Headlines between Obama's and Trump's presidencies

Figure 7 shows that, for negative word usage, all of the differences between the two presidencies are statistically significant, since none of the confidence intervals include zero. The two biggest differences in terms of the point estimate are for the U.S and, interestingly, Arts sections. Figure 7 shows that, on average, the NYT writers used around 1% less negative words during Obama's presidency than during Trumps' presidency, suggesting that NYT writers felt more positive sentiment during Obama's presidency. Indeed, this difference stands out the most, as it is quite large, and would naturally be most impacted by U.S. politics, suggesting that the presidency had a large impact on the section. Interestingly, however, some sections, like Arts, Opinion, and Business Day, used less negative words during Trump's presidency. It is unclear why a category like Arts might have had such a reduction in negative word usage during Trump's presidency; however, for Business Day, it could be argued that due to president Trump's fiscally conservative politics, business and the markets experienced a more positive four years.

Although all of the differences are statistically significant, and some could even be explained by real-life events, those differences are not large, with the largest ones being only around 1%, when taking into consideration the extreme bounds of the confidence intervals. As we saw previously, on average, each headline consists of around 7-10 words, so 1% does not even make up a word. This suggests that perhaps the NYT writers did not use different amounts of negative words during the two presidencies, and even if they did, the difference is not large enough to be noticeable by a casual reader. It could be argued that these small differences compound

over the years, and possibly lead to skewing a reader's perception of affairs, but it would more likely depend on the reader, how much media they consume, and how analytical they are when it comes to the news.

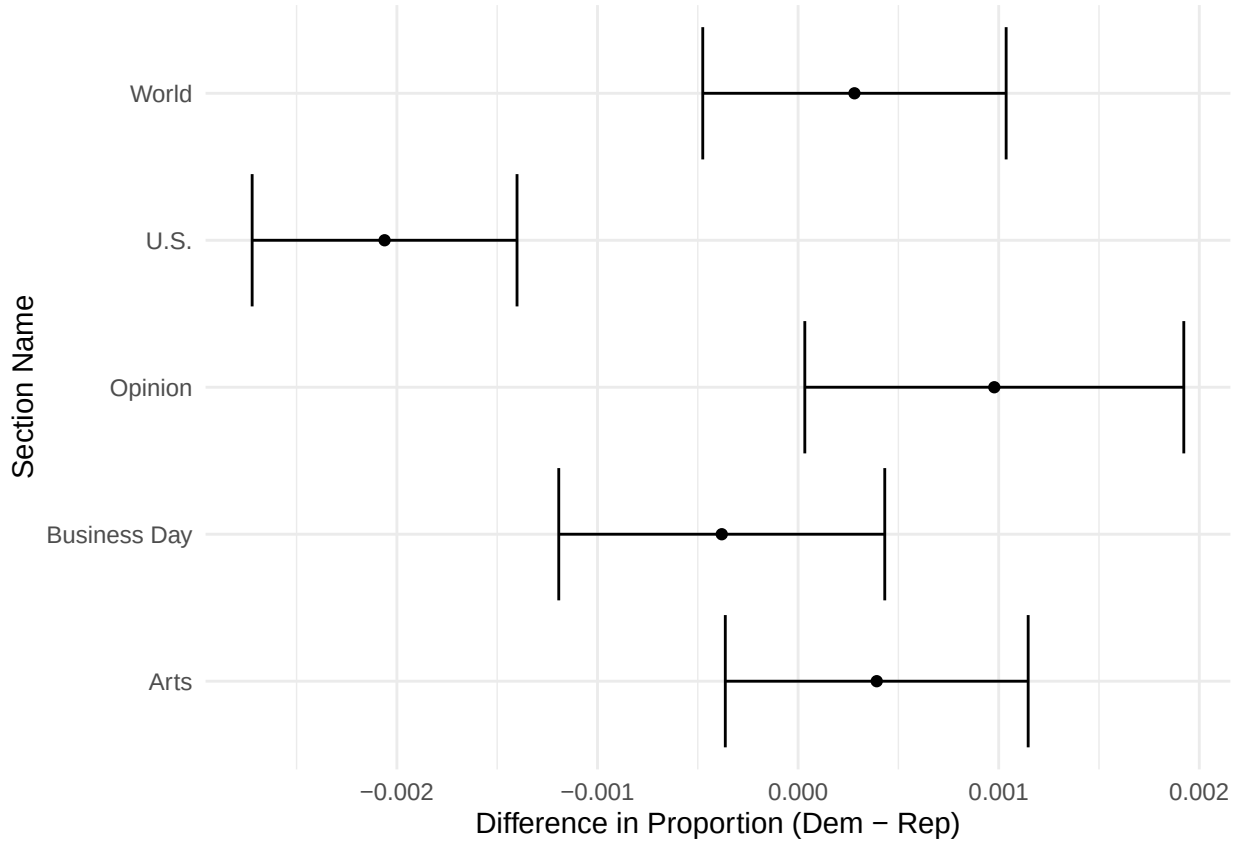

Figure 8: Difference in Positive Word Usage in NYT Headlines between Obama's and Trump's presidencies

Similarly to negative word usage, figure 8 shows that the differences in positive word usage are also fairly small, with the largest differences hovering around 0.2%. Unlike negative word usage, most of the differences are insignificant, since zero is in most of the confidence intervals, with only U.S. and Opinion sections having a significant difference. It is interesting that for the U.S. section, positive word usage has increased during Trump's presidency. Based on figure 7, we could have made the conclusion that NYT writers were biased towards president Obama, but figure 8 suggests that that might not necessarily be the case, since the writers used less positive words during Obama's presidency. Perhaps, this shift in language suggests that the reporters started using more emotional language in general, or perhaps the differences stemmed more from the real-life events that happened during the two presidencies, like the pandemic. Similarly, the Opinion category has actually seen a decrease in both positive and negative word usage, which could reflect that how the Opinion section is written has changed over the years. This change is slightly surprising, since Opinion sections are usually expected to be less objective, and so would perhaps contain more emotive language, when compared to other categories.

Since the statistical tests above did show that the differences in how often the positive or negative words are used in article headlines are fairly small, perhaps, it is the words themselves that are different. It is possible that certain categories employed different words during the two presidencies, which is not reflected in the analysis above. As such, it is important to also determine the most used positive or negative words in each article section and investigate if they are different.

Figures 9 and 10 show that there was some variation in the words used between the two presidencies;

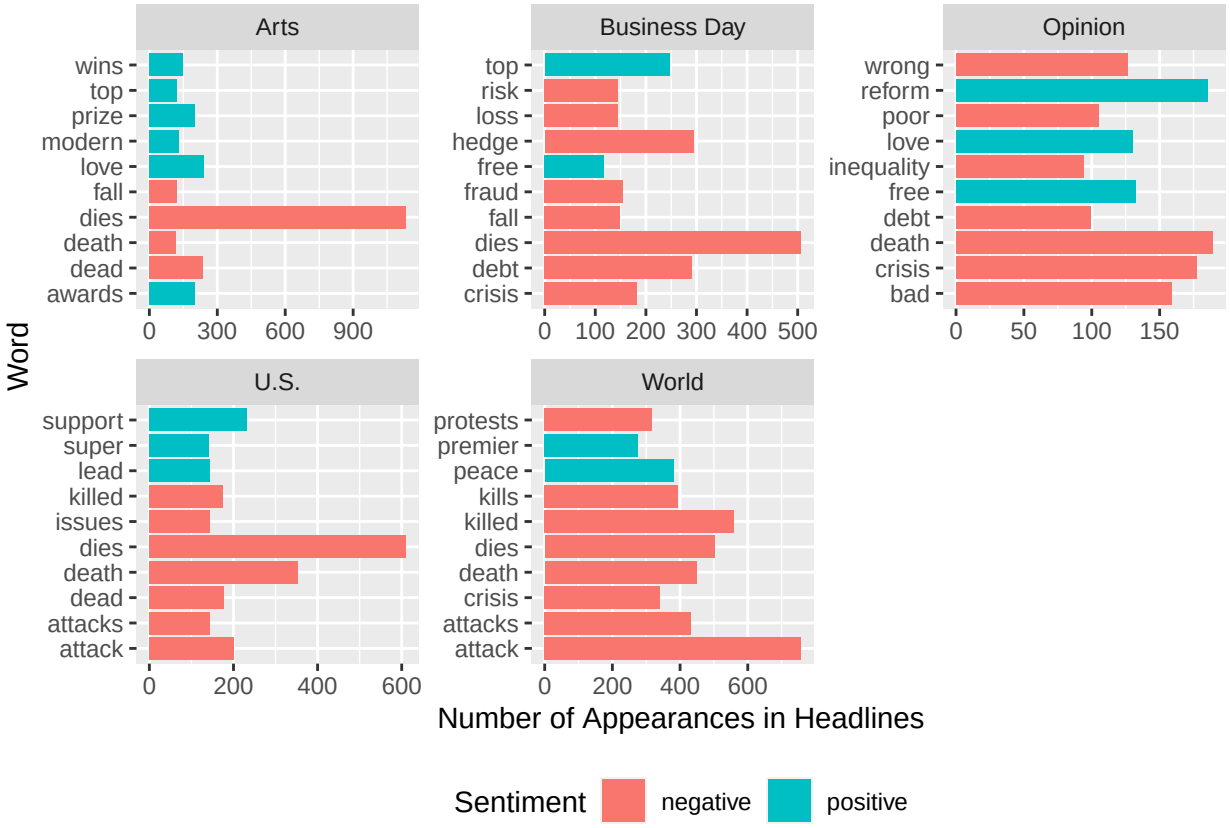

Figure 9: 10 Most Commonly Used Positive or Negative Words in NYT Headlines during Obama's Presidency

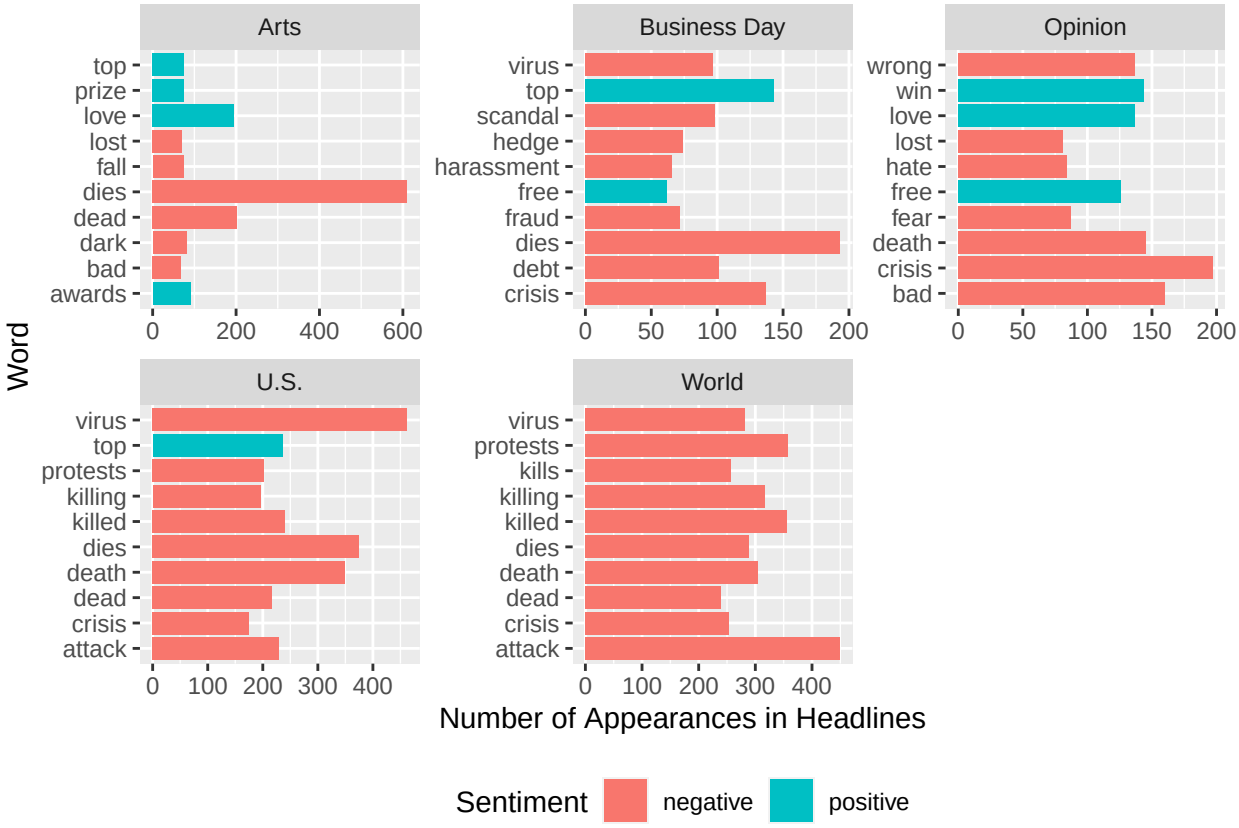

Figure 10: 10 Most Commonly Used Positive or Negative Words in NYT Headlines during Trump's Presidency

however, it is not clear if these variations can be attributed to the fact that the presidents during the two periods belonged to different political parties. For instance, in the U.S. category, the word "virus" was the most common headline word during Trump's Presidency, despite not being used frequently during Obama's Presidency. It is fairly evident that this is because a global pandemic happened during 2020, during Trump's Presidency, making it understandable why it would be such a common word. For certain sections, such as the "Arts" section, we see very little difference in the words used between the two presidencies. This suggests that U.S. politics had very little impact on the Arts or that the presidencies in general do not have a large impact on the exact language used by the writers of NYT.

Conclusion

There are a few differences in how NYT writers structured and wrote their headlines during Obama's and Trump's presidencies. Firstly, overall, the amount of articles written during Trump's presidencies was less than those written during his predecessor's time. Additionally, the headlines tended to be longer when a member of the Republican party was in office.

For some categories, the writers tended to use more negative language during Trump's presidency. This difference was significant for the top 5 categories of NYT, but, nevertheless, the difference was quite small, hovering at around 1% at most. This suggests that, perhaps, the difference was not noticeable to the reader and would likely not impact their world perception, if the reader was unbiased and exclusively read headlines. This assumption, of course, is unrealistic, so the reader's viewpoint would most-likely be impacted by other factors, like how much media the reader consumer or their already-existing political beliefs.

NYT writers used a similar amount of positive words across the two presidencies, with most of the differences being statistically insignificant. It is interesting that for some categories, the writers used more emotive language, positive and negative, during Trump's presidency than they did during Obama's presidency, while for others, the writers used less emotive language. This suggests that perhaps the presidency had little to no impact on how the language that the writers used in their headlines.

Indeed, this is further supported by the exact words used by the writers. The words more closely reflect the events that happened during the two presidencies, rather than any political belief or opinions about the two presidents.

In conclusion, although there were some differences in how article headlines were written between Obama's and Trump's presidency, it is hard to attribute this difference to the change in the president of the U.S. This, in turn, means that the political party of the president and the political landscape in general of the country seems to have little to no effect on how NYT writers write their headlines and what kind of language they use.

Limits

There are several limitations with this analysis. Firstly, I have only looked at two presidencies in total, one for each political party, both of which happened in the last decade. This means that my analysis is greatly susceptible to recency bias. It could be argued that perhaps the media outlet's views on a specific president during the last decade were particularly harsh when compared to past presidents who belonged to the same party. This limitation makes it difficult to generalize these findings to all presidents and political parties in general. Perhaps, by getting more data from other presidencies, it would be easier to generalize the findings of this study to other presidents and the political parties of the country in general.

Another possible limitation of this analysis is the fact that it simplified U.S. politics into just the presidency, although there are other branches of government, which also serve an important role in the country, like the Congress or the Senate. Even if a president belongs to the Republican party, for example, the House or the Senate could still be majority Democratic, which could influence how the NYT reports on different issues. When we try to incorporate these factors in our analysis, it becomes quite difficult to define which years Democrats held more power and which years Republicans held more power. As shown by (Aschwanden 2015) and the FiveThirtyEight team, different factors could be included and discluded to skew the results one way or the other. As such, for this analysis, the presidency seemed like the most appropriate and straightforward position to explore, since it is the one that is most discussed in the media, which is what we looked at today.

Another limitation is that we considered the meaning of the words in the headline at face-value. This is ineffective for multiple reasons. Firstly, as we saw with the word "trump," the word could have multiple meanings. Although we were able to detect this word early on, it is possible that other words with a similar issue were not detected. This skews the results, and possibly over-counts how many emotionally-charged words there are in the headlines. Additionally, writers might intend certain words with sarcastic meaning. For instance, positive words like "amazing" could be intended sarcastically, meaning they would actually have a negative connotation about them. It could be argued that a writer might not do this to avoid confusion, but it is possible, nevertheless, especially when it comes to more opinion-based articles, which were considered in this analysis. A possible way to avoid both of these issues is to have a human go through the dataset and interpret each word in context and indicate its possible meaning. This, however, would take a long time and would not guarantee perfectly accurate results, since it might be hard to perfectly accurately understand the context from 8-11 words.

Future Steps

In this analysis, I have exclusively looked at headlines; however, political bias might be reflected in other ways. For instance, in future analysis, it would be interesting to look at the topics that the writers reported on, instead of just the language that the writers used to do so. Perhaps, NYT writers would highlight topics that would emphasize achievements of a political party, instead of its failures. This kind of analysis would especially be interesting if compared to other newspapers and media outlets to see if their biases are different. Perhaps a neutral foreign media outlet could serve as the control variable, since that outlet would most-likely be netraul and would not prefer one U.S. political party over another.

This analysis could also be extended by considering the article body itself. While headlines are the first thing the reader sees, article bodies are also quite important, and tend to be much longer, providing the writer with more space to report on the events, and, possibly, let their bias or personal political belief seep through. Additionally, the extended length would also allow us to be more confident in the results of this analysis, and would perhaps show larger differences in how the language has changed over the presidencies.

Bibliography

Admin. 2019. "How Many People Read the Newspaper in the World?" *MassInitiative*. https://massinitiative.org/how-many-people-read-the-newspaper-in-the-world/.

Aschwanden, Christie. 2015. "Science Isn't Broken." *FiveThirtyEight*. FiveThirtyEight. https://fivethirtyeight.com/features/science-isnt-broken/.

Britannica, The Editors of Encyclopaedia. n.d. "List of Presidents of the United States." *Encyclopaedia Britannica*. Encyclopædia Britannica, inc. https://www.britannica.com/topic/Presidents-of-the-United-States-1846696.