Sentiment Distribution Analysis of News Headlines using Natural Language Processing and ANOVA Techniques

Final Project - May 31, 2023

Athan Zhang	1570738
Aidan Li	1562961
Amy Wang	1560918

Torbert Research Statistics,
Thomas Jefferson High School for Science and Technology

Abstract

Sentiment analysis plays a crucial role in understanding public opinion and sentiment towards various topics. This study presents an application of Natural Language Processing (NLP) techniques combined with Analysis of Variance (ANOVA) to analyze sentiment in news headlines. The objective is to gain insights into the overall sentiment and identify any significant differences in sentiment across different news categories and publishers.

Contents

1 INTRODUCTION

1.1 Rationale

Mean World Syndrome is a perceived cognitive bias in which people tend to see the world as cruel, a feeling that is amplified by repeated negativity in media. However, whether or not the media tends to trend more negatively in their reporting is unknown. This project aims to consider different variables that may affect the sentiment of news headlines in order to understand what really affects how positive or negative a news report can be.

We use Natural Language Processing (NLP) techniques in order to extract sentiment data from headlines. These headlines are gathered from large public datasets that contain information on news articles, their headlines, category, date, and other variables covered later in this report. We are able to extract information on polarity (negative or positive) and subjectivity (opinionated or factual) to understand the sentiments of news headlines.

To understand the statistical significance of the sentiment, we use Analysis of Variance techniques to see whether or not news headline sentiment and other variables, such as publisher and category, are homogenous in their distributions. This allows us to understand what variables play the largest role in news sentiment.

1.2 Related Works

People interact with news reports on various social media platforms every day. In an investigation conducted in 2022, researchers detected positive and negative sentiments, as well as basic emotions like anger, disgust, fear, joy, sadness, and surprise. They found increasingly negative sentiment over time, with right-wing outlets being more harmful than left-wing outlets on average. In terms of specific emotions, they observed an increased proportion of headlines with anger, fear, disgust, and sadness [Rozado et al., 2022].

Similarly, another article discusses the role of affect and emotion in successful viral diffusion, suggesting that effectively charged viral messages are more likely to spread than neutral ones. However, the article also contrasted these findings and established research on news factors, emphasizing adverse effects promoting propagation in news media. The methodology involves using machine learning techniques and applying them to different text corpora to classify messages as "news" and analyze sentiment [Hansen et al., 2011].

BERT is an excellent example of machine learning that is used for this purpose. It can be used to analyze the sentiment involved within phrases and words, leading to the examination of how people react to differently worded posts about the same incident on multiple social media platforms. The model identifies emotions involved in a sequence of words to output. BERT was able to successfully pass the GLUE benchmark, shedding credibility to the language interpretation model [Devlin, 2019].

In an article written in 2020, researchers tried using entity sentiment analysis that differentiates sentiment towards target entities and considers associated entities instead of using traditional methods. Their findings reveal the manner of sentiment expression and the specific aspect of the entity impacts sentiment strength. The study provides insights for improving entity sentiment analysis in news contexts by considering associated entities and sentiment expression and addressing the presence of biased entities [Luo and Mu, 2020].

Researchers attempted to classify news articles and categorize them with viewer engagement. They were able to successfully predict how viewers would react to different types of posts on various social media platforms through past comments and analyzing the sentiments toward corresponding posts [Aldous et al., 2022].

1.3 Research Question

Will the publisher or category of news headlines affect the polarity and subjectivity of them?

The goal of this research project is to find which variables such as publisher or category, have a relationship with the polarity and subjectivity of news headlines.

2 DATASET AND FEATURES

We use two of the most widely used datasets for news headlines: the *UCI News Aggregator* (NAD) and *All The News* (ATN) datasets. Table 1 provides an overview of the datasets. More information can be found later in this section.

Dataset	**Instances**	**Variables**
News Aggregator	422,419	4 (Headline, Publisher, Category, Date)
All The News	142,568	3 (Headline, Publisher, Date)

Table 1: Dataset Information

NAD will be used to analyze the effect of *Category* on Polarity, and thus the bar chart for categories can be seen in Figure 1. It will also be used to explore the relationship between Polarity and Subjectivity.

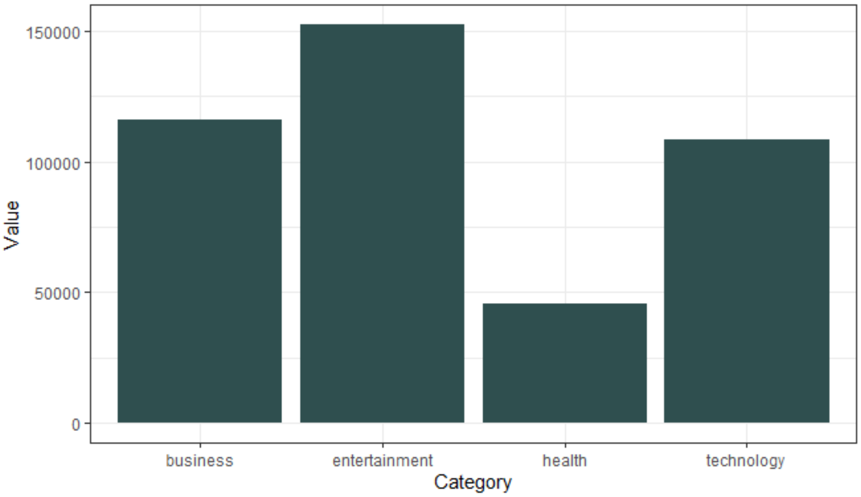

Figure 1: Bar Chart of Categories for NAD

ATN will be used to analyze the effects of *Publisher* on Polarity, and thus the respective bar chart is shown in Figures 2.

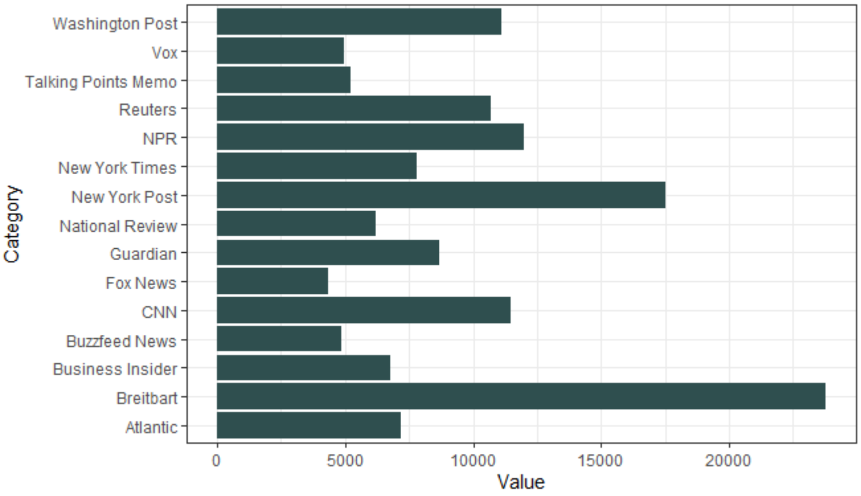

Figure 2: Bar Chart of Publishers for ATN

Each dataset was chosen for its respective strengths. The NAD dataset allows us to get a comprehensive overview of categories, while the ATN dataset provides a large yet concise analysis of different major media corporations.

More data analysis on Polarity and Subjectivity will be provided later in the Results section of this report.

2.1 News Aggregator Dataset

The News Aggregator Dataset (NAD) is provided by the University of California Irvine's Machine Learning Repository [Dua and Graff, 2017]. It is curated initially from understanding user interests based on web activity [Gasparetti, 2016].

The dataset is extensive and supplies a large amount of news article data. A primary variable of interest is the 'Category' variable, which allows us to explore disparities between different reporting focuses and their effect on sentiment. The main primary drawback of this dataset is its short timeframe, only pulling news data from 2014.

2.2 All The News

The All The News dataset is made publicly available by Kaggle. It is a web-scraped collection of news headlines from late 2015 to 2017. The data wasn't subject to random sampling and data collection was made at the curator's discretion [Thompson, 2017].

This dataset allows for a smaller range of different publishers but highlights major news corporations of interest such as The New York Times, CNN, Vox, Buzzfeed, etc. It also has a larger time frame which allows us to view changes in sentiment over time. However, this dataset did not group headlines by category, so there will be more data exploration within that variable.

3 METHODS

3.1 Analysis of Variance (ANOVA)

ANOVA, which stands for *Analysis of Variance*, is a statistical method used to analyze the differences between the means of three or more groups. It helps determine if there are significant variations among the group means and whether those differences are due to random chance or actual group differences. The formula for one-way ANOVA is:

$$F = \frac{\text{between-group variance}/(\text{number of groups} - 1)}{\text{within-group variance}/(\text{total number of observations} - \text{number of groups})}$$

The between-group variance measures the variation between the means of different groups, while the within-group variance measures the variation within each group. By comparing these variances, the F-ratio is calculated. If the F-ratio exceeds a certain critical value, it indicates that the group means are significantly different from each other, suggesting that there is a significant effect of the independent variable on the dependent variable.

3.2 Natural Language Processing

Natural Language Processing (NLP) is a field of study that combines linguistics, computer science, and artificial intelligence to enable computers to understand, interpret, and generate human language. It focuses on developing algorithms and models to process and analyze textual data. NLP encompasses various tasks, such as sentiment analysis, text classification, named entity recognition, machine translation, and question answering. Techniques like tokenization, part-of-speech tagging, syntactic parsing, and word embeddings are commonly used in NLP.

We used NLP techniques to analyze the headlines and generate polarity and subjectivity scores.

- Polarity: A range between -1 and 1 is used to measure positive or negative sentiment, positive being higher values.
- Subjectivity: A range between 0 and 1 and refers to personal opinions and judgments.

TextBlob is an open-sourced Python library that makes sentiment analysis simple. We tokenized each headline in the datasets and generated their respective polarity and subjectivity scores using TextBlob [Loria, 2018]. This was then used as our response variable of the study and a primary focus variable.

3.3 Implementation

The proposed approach was implemented on Python 3.6. and R 4.3.

- **GitHub Repository:**
  `https://github.com/athanzxyt/newsheadline_sentiment`

We used NumPy and Pandas Python data-processing libraries. TextBlob for Sentiment Analysis. Matplotlib and ggplot2 for Data Visualization. and Mosaic and RTools for statistical analysis.

4 RESULTS & DISCUSSION

4.1 Polarity & Subjectivity Calculations

The calculated Polarity and Subjectivity scores are calculated for the two datasets. The descriptive statistics are shown in Table 2 and the distributions in Figure 3.

	NAD		ATN	
	Polarity	Subjectivity	Polarity	Subjectivity
μ	0.043	0.224	0.030	0.256
σ	0.222	0.297	0.241	0.307
Max	1.000	1.000	1.000	1.000
Q3	0.080	0.455	0.100	0.455
Med	0.000	0.000	0.000	0.100
Q1	0.000	0.000	0.000	0.000
Min	-1.000	0.000	-1.000	0.000

Table 2: Descriptive Statistics for Polarity and Subjectivity Scores

We can see that the max, minimum, and first quartile are the same between NAD and ATN. Mean-wise, NAD has a greater mean for polarity compared to ATN, while ATN has a greater mean for Subjectivity than NAD. The opposite is true for the standard deviations. The quartile 3 for polarity at ATN is slightly higher than the one at NAD. For the median, the subjectivity for ATN is slightly more significant than NAD.

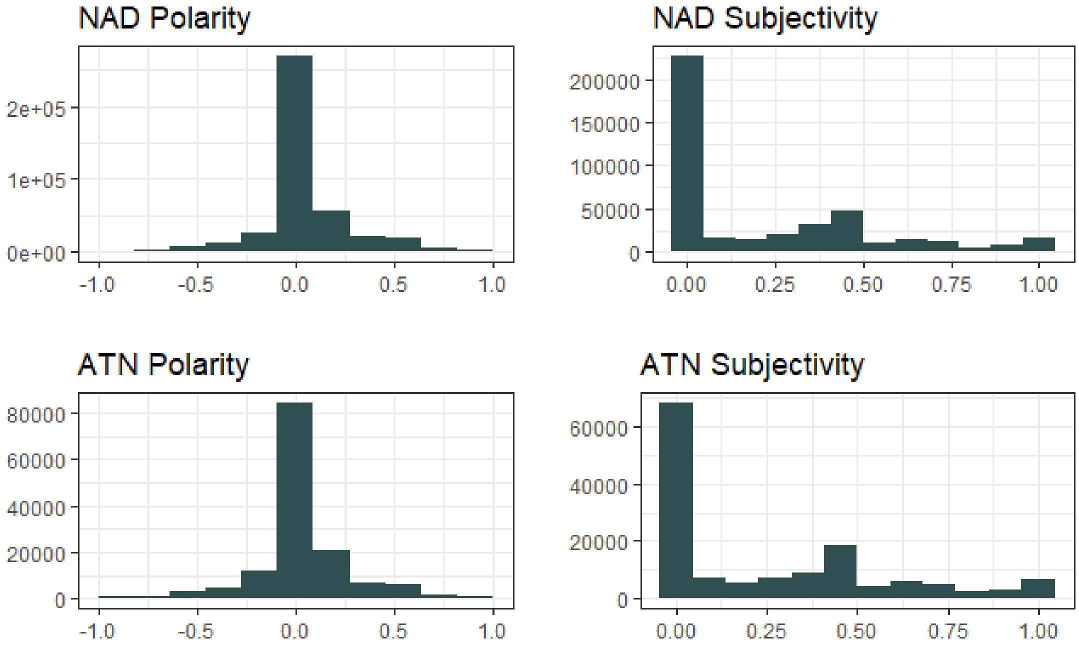

Figure 3: Histograms for Polarity and Subjectivity

We can determine that the histograms for polarity and subjectivity, each, are relatively similar. Between the NAD and ATN polarity, the histograms are centered around 0 and have a relatively normal distribution. However, for the NAD and ATN subjectivity histograms, the mean seems to be near 0 with both graphs being extremely right-skewed.

Further explanatory data analysis is conducted. We plot the scatterplots of Polarity versus Subjectivity and the absolute value of Polarity versus Subjectivity in Figures 4 and 5.

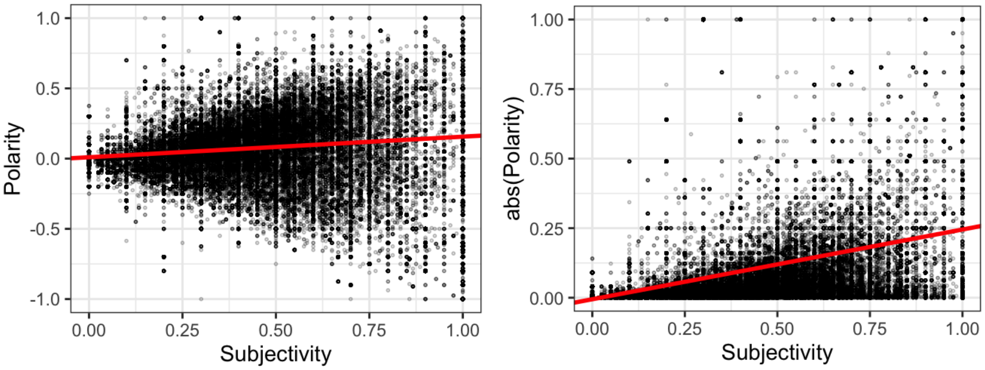

Figure 4: Polarity versus Subjectivity Scatterplot for NAD

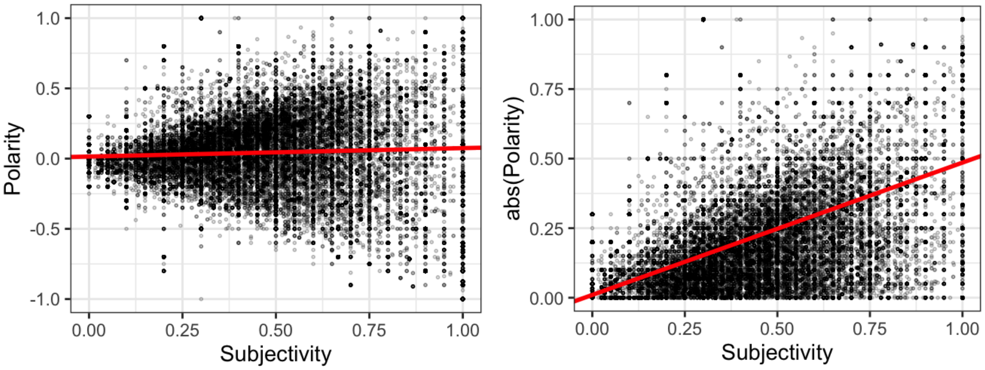

Figure 5: Polarity versus Subjectivity Scatterplot for ATN

As we can see, the relationship between Polarity and Subjectivity seems to be relatively apparent, with a moderate linear positive relationship. However, since the variance of Polarity increases with Subjectivity, we did not carry out regression tasks. Nevertheless, an interesting observation that as news headlines become more subjective, they tend to become more extreme in their sentiments.

4.2 Category and Polarity

We first explore the relationship between the category of the news headline on its polarity. We examined four categories: technology, health, entertainment, and business-related news headlines. The distributions of this can be seen in Figure 6.

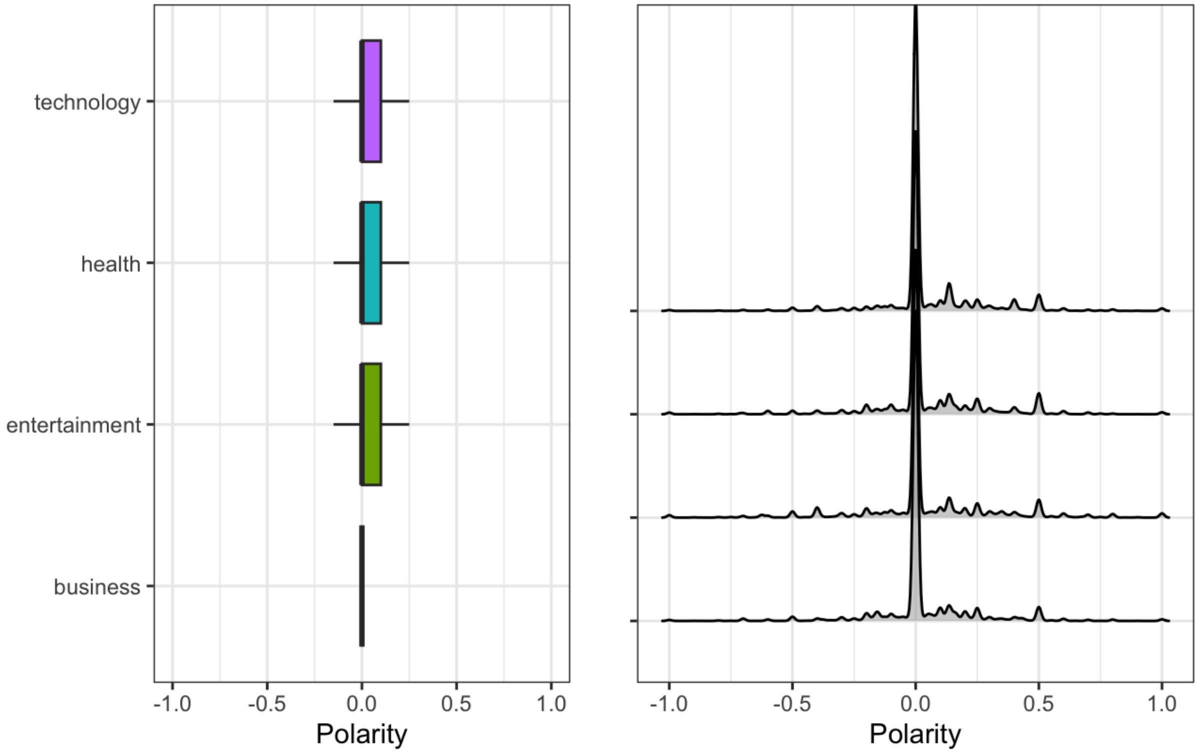

Figure 6: Polarity Distributions of each Category in NAD

All distributions appear to be symmetric and approximately normal. All distributions appear to have similar shapes, with the exception of the business category, which has its entire boxplot located at zero. All outliers were removed for the sake of visualization.

Since our dataset meets the independent condition (there are likely more than one million news articles from 2014), is generated from observation, and appears to be approximately normal (distributions and also generated from a significantly large dataset), we continue with a one-way ANOVA test of homogeneity. The hypothesis of the test can be found below.

$$H_0\text{: } \mu_t = \mu_h = \mu_e = \mu_b$$
$$H_a\text{: At least one of } \mu_k \text{ differs}$$

Where μ_k represents the true mean polarity for each category. The results of the ANOVA can be found below.

	DF	Sum Sq	Mean Sq	F-Value	P
Category	3	29.8	9.9473	202.52	2.2E-16
Residuals	422415	20747.5	0.0491		

As we can see, we have a relatively high F-value which suggests that the means are not necessarily all equal. This is furthered by the p-value of 2.2E-16, which is less than $\alpha = 0.05$; thus, we reject the null hypothesis. There is statistical evidence that one of the news headline categories does not have the same mean polarity.

Conducting a Fisher LSD procedure finds that technology and business are individual groups, while entertainment and health share a similar mean polarity.

4.3 Publisher and Polarity

The polarities are displayed graphically with side-by-side density and dot plots in Figure 7. We can see that all distributions appear to be relatively symmetric with similar spreads, with the exception of the New York Times.

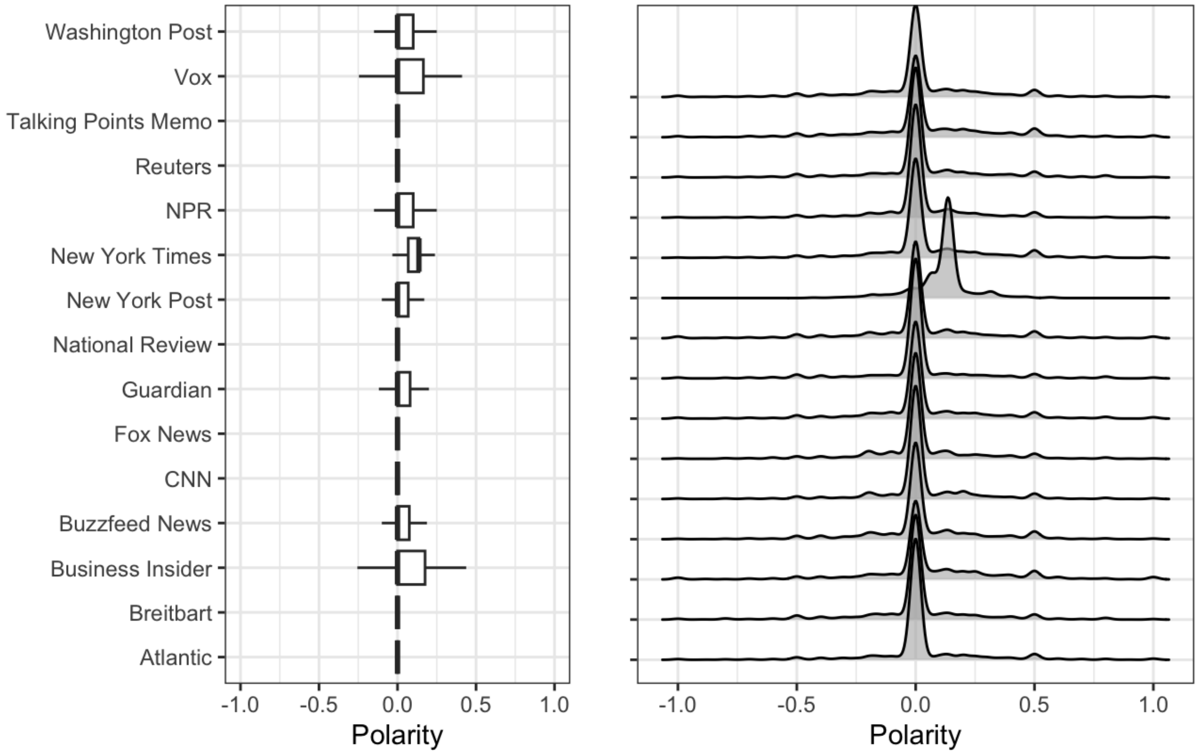

Figure 7: Polarity Distributions of each Publisher in ATN

Analyzing the density graphs, we can see that all of the graphs are centered around a polarity of 0 with a normal distribution, save for New York Times. New York Times has a slightly greater polarity mean as it is shifted to the right of all the other density graphs. As the box plots show, noting that all outliers have already been removed, the publishers all have varying variances. For example, Talking Points Memo, Reuters, National Review, Fox News, CNN, Breitbart, and Atlantic all are centered on a polarity of 0, with almost no variance at all. New York Times, New York Post, Guardian, and Buzzfeed News have a similar range, mostly around 0.25. However, the other publishers all have a relatively high variance, with Business Insider having a range nearing 0.75.

Since our dataset meets the independent condition (there are likely more than one million news articles from 2015-2017), is generated from observation, and appears to be approximately normal (distributions and also generated from a significantly large dataset), we continue with a one-way ANOVA test of homogeneity. The hypothesis of the test can be found below.

$$H_0\colon \mu_{WaPo} = \mu_{Vox} = ... = \mu_{Atlantic} = \mu_{CNN}$$
$$H_a\colon \text{At least one of } \mu_k \text{ differs}$$

Where μ_k represents the true mean polarity for each publisher. The results of the ANOVA can be found below.

	DF	Sum Sq	Mean Sq	F-Value	P
Category	14	78.4	5.5992	97.592	2.2E-16
Residuals	142551	8178.6	0.0574		

As we can see, we have a relatively high F-value which suggests that the means are not necessarily all equal. This is furthered by the p-value of 2.2E-16, which is less than $\alpha = 0.05$; thus, we reject the null hypothesis. There is statistical evidence that one of the news headline categories does not have the same mean polarity.

Conducting a Fisher LSD procedure aids us in identifying which publishers have similar means. Figure 8 shows a relationship map of these groups.

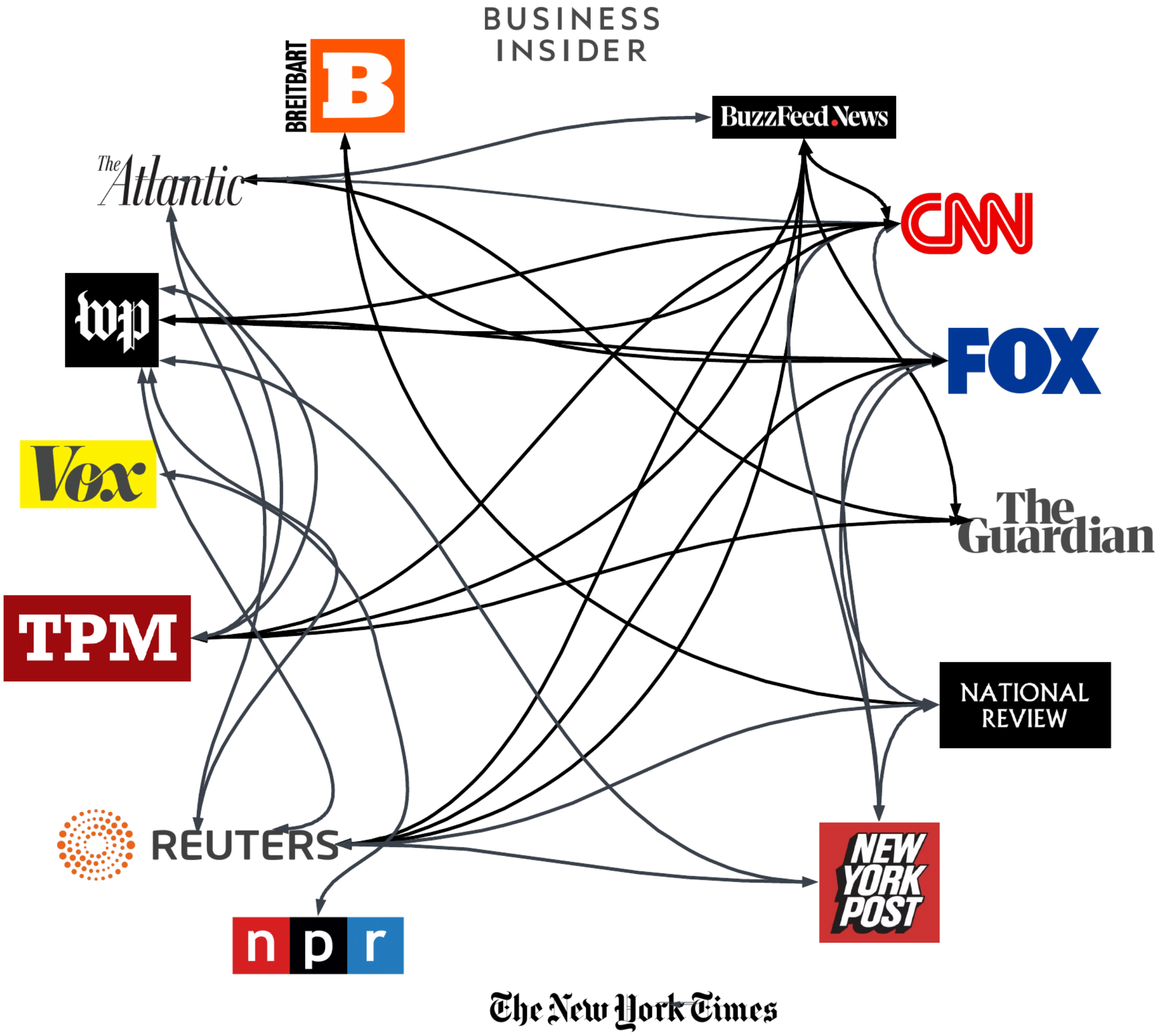

Figure 8: Shared Polarity Means among Publishers

Most notably, two news sources were not a part of any group: The New York Times and Business Insider. As seen from the previous boxplots, this is most likely indicative of a more positive polarity than the other news sources. Other groups, however, saw multiple news sources having similar average polarities. While traditionally right-leaning news sources were often grouped with traditionally right-leaning news sources and traditionally left-leaning news sources were frequently grouped with traditionally left-leaning news sources, there was a significant grouping of both left and right-leaning sources together. For example, while traditionally right-wing New York Post was correlated with other right-leaning sources like Breitbart, National Review, and Fox, it was also considered in the same group the Washington Post and CNN, both left-leaning news sources.

5 CONCLUSION

This report analyzes the polarity of news headlines on a variety of factors. We explored how the category of the news article, as well as the publisher of the news article, affects the sentiment of the headline.

We found a statistical difference in the polarity of news headlines across different news categories. This means that not all news categories are equally positive or negative. Most notably, we found that the business and technology categories are independent of health and entertainment, which share similar polarity scores.

We also found a statistical difference in the polarity of news headlines across different news publishers. This means that different news corporations tend to publisher different sentiment headlines than others. Most notably, both Business Insider and The New York Times shared no statistically standard polarity score with other groups. From the density distributions, we note that the New York Times tends to publisher more positive news headlines. We provide information on which of the top news publishers publish similar sentiments in their headlines.

We also noted an observational positive linear relationship between the polarity and subjectivity of news headlines. Most notably, as the subjectivity of a news headline increases, so does the extreme of its polarity. This reinforces the obvious idea that subjectivity results in bias, which results in extreme sentiments.

5.1 Connections to literature

Similar to the results found in Rozado et al. (2022), Hansen et al. (2011), Aldous et al. (2023), we found that, with machine learning, we were able to find correlations between polarity involved in the news headlines and the company that published the bar, as well as the correlation between polarity and the category the news headline was in. When comparing polarity and subjectivity, we noticed a positive linear trend between the two.

5.2 Further Research

Some future work for this project is to increase the data pulled and analyzed. Being able to extensively analyze news headlines over multiple decades with millions of headlines will increase our analysis.

Additionally, improving the sentiment classification algorithm could result in more variables and information to analyze. Building multi-layered neural networks or using pre-trained BERT models could allow for deeper analysis.

This work could lastly be improved by expanding to the entire news article instead of simply just the headline; however, this is very computationally expensive and beyond the scope of this project.

5.3 Limitations

There are a few notable limitations of this study that the authors want to note. First and foremost, the timeframes of the news articles collected span relatively short windows, reaching a timeframe of at max three years in some cases. The authors want to note potential biases and confounding variables that may be found in this window. However, nothing of notable significance occurred in these years, and sufficiently large amounts of data were pulled, mitigating some of these effects.

Secondly, the categories analysis did not include a political category, which is a typically charged sentiment domain.

Thirdly, the groups across both datasets were not necessarily balanced; for example, there were notably fewer health news headlines than others from categories.

Finally, the sentiment measurement of polarity and subjectivity could be naturally biased based off of the dataset that the TextBlob classifier was trained on, which may be reflected in our calculations.

5.4 Reflections

During our project on analyzing news headlines using sentiment analysis, we developed several skills that could significantly enhance our capabilities as researchers and budding statisticians, which could prove valuable in future research endeavors. The first skill, of course, is statistical analysis: This project necessitated the use of various statistical techniques to analyze the sentiment of news headlines accurately. We gained proficiency in applying statistical methods such as hypothesis testing and the four-step process, as well as identifying areas of inquiry. These statistical analysis skills are fundamental for any researcher and can be applied to diverse research subjects. Furthermore, the project demanded extensive coding in R for data manipulation, visualization, and statistical modeling. Applying R to real-world applications has equipped us with a powerful tool for future research projects in each of our respective fields. For instance, when studying the impact of environmental factors on public health, we can utilize R to analyze large datasets, visualize trends, and build predictive models. The ability to program effectively will allow me to handle complex research tasks efficiently and automate processes that may have taken ages manually. While we had many successes, our failures were numerous as well. Throughout the project, we encountered several challenges, such as dealing with errors in RStudio, managing outliers, and selecting appropriate statistical methods. Overcoming these obstacles required critical thinking and problem-solving skills, as well as sheer determination to not give up at times. Finally, we learned significant lessons during the interpretation and communication of research findings. We learned how to analyze and summarize the sentiment trends in news headlines and present the results in a clear and concise manner.

6 ACKNOWLEDGEMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

The authors would like to thank Dr. Shane Torbert of the Thomas Jefferson High School for Science and Technology (TJHSST) for his invaluable contributions in field knowledge and manuscript review.

References

[Aldous et al., 2022] Aldous, K. K., An, J., and Jansen, B. J. (2022). What really matters?: Characterising and predicting user engagement of news postings using multiple platforms, sentiments and topics. *Behaviour Information Technology*, 42(5):545–568.

[Devlin, 2019] Devlin, Chang, L. T. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Dua and Graff, 2017] Dua, D. and Graff, C. (2017). UCI machine learning repository.

[Gasparetti, 2016] Gasparetti, F. (2016). Modeling user interests from web browsing activities. *Data Mining and Knowledge Discovery*, 31(2):502–547.

[Hansen et al., 2011] Hansen, L. K., Arvidsson, A., Nielsen, F. A., Colleoni, E., and Etter, M. (2011). Good friends, bad news - affect and virality in twitter. *Communications in Computer and Information Science*, page 34–43.

[Loria, 2018] Loria, S. (2018). textblob documentation. *Release 0.15*, 2.

[Luo and Mu, 2020] Luo, M. and Mu, X. (2020). Identifying factors impacting entity sentiment analysis: A case study of sentiment analysis in the context of news reports. *Proceedings of the Association for Information Science and Technology*, 57(1).

[Rozado et al., 2022] Rozado, D., Hughes, R., and Halberstadt, J. (2022). Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with transformer language models. *PLOS ONE*, 17(10).

[Thompson, 2017] Thompson, A. (2017). All the news.