Exploring Data
On
News Articles

Stat 3355.501 Project

David Munoz, Shawn Kennedy, Ajay Palankar
April 8, 2020

Introduction

The data used for this project helped identify some interesting trends seen in both media outlet and media consumer habits. The group explored one dataset for this project and it was sourced from the UCI database. The dataset accounts the popularity of news articles from November 2015 to July 2016 on the three different social media platforms Facebook, Google+, and LinkedIn. The dataset includes data on the positivity or negativity of the news headlines and titles as well as the popularity across all three of the social media platforms. It covers 93239 articles with 11 variables accounted for through the eight month period. The dataset variables are a unique ID(IDLink) assigned to each article, the title of the article(Title), the headline of the article according to the official media source(Headline), the original source of the news article(Source), the topic of the article obtained by the specific query used to obtain the article(Topic), the date of publication(PublishDate), a sentiment score calculated using the text from the title(SentimentTitle), popularity of the article on Facebook based off the final score given by Facebook(Facebook), popularity of the article on Google+ based off the final score assigned by Google+(Google+), and finally the popularity of the article on LinkedIn according to the final score assigned by LinkedIn(LinkedIn). Our reasoning behind choosing this dataset to analyse is because we believed it would reveal some interesting and important trends found in media outlets and the way consumers act in the media market.

The trends that our group was most interested in uncovering were which variables in the dataset would affect the popularity of a given news article. The viral nature of news is often unpredictable but certain trends can be observed given enough data. This dataset's most interesting variables are the sentiment and the popularity variables so most of our questions posed are focused on those two and how the other variables may affect them.

Data Cleaning

- Time_of_day - The time of day that the article was posted with three possible values “Morning”, “Evening”, and “Night”.
- Bias - The bias of the source of the article obtained by outside sources [*]
- combpop - The combined popularity across all three social media platforms.
-

Questions and Findings

1. **Does the time of day published affect the popularity of the article?**

Looking into the data to answer this question, we do not care about an article’s popularity on any one social media platform in particular, so to explore the question, the combined score of all three platforms was used. As seen in figure 1, the time of day with the most popular articles is the evening and the least popular articles are posted in the morning. Due to the viral nature of some news articles, it was important to use a boxplot to explore this question as there were plenty of outliers in the data for every section of the day.

This question is important to ask because while news outlets have a duty to report the news they are at their core a business seeking to make a profit so it is important to know this kind of information. Interesting or important news could be overlooked or buried based on the time of day it was posted.

This trend has been artificially created through having a large majority of articles posted in the evening and a smaller number posted in the morning. This could in turn inflate or understate the medians shown in figure 1. Given this concern a boxplot was again the obvious choice as we can see the amount of outliers in each dataset and easily compare each time of day's boxplot.

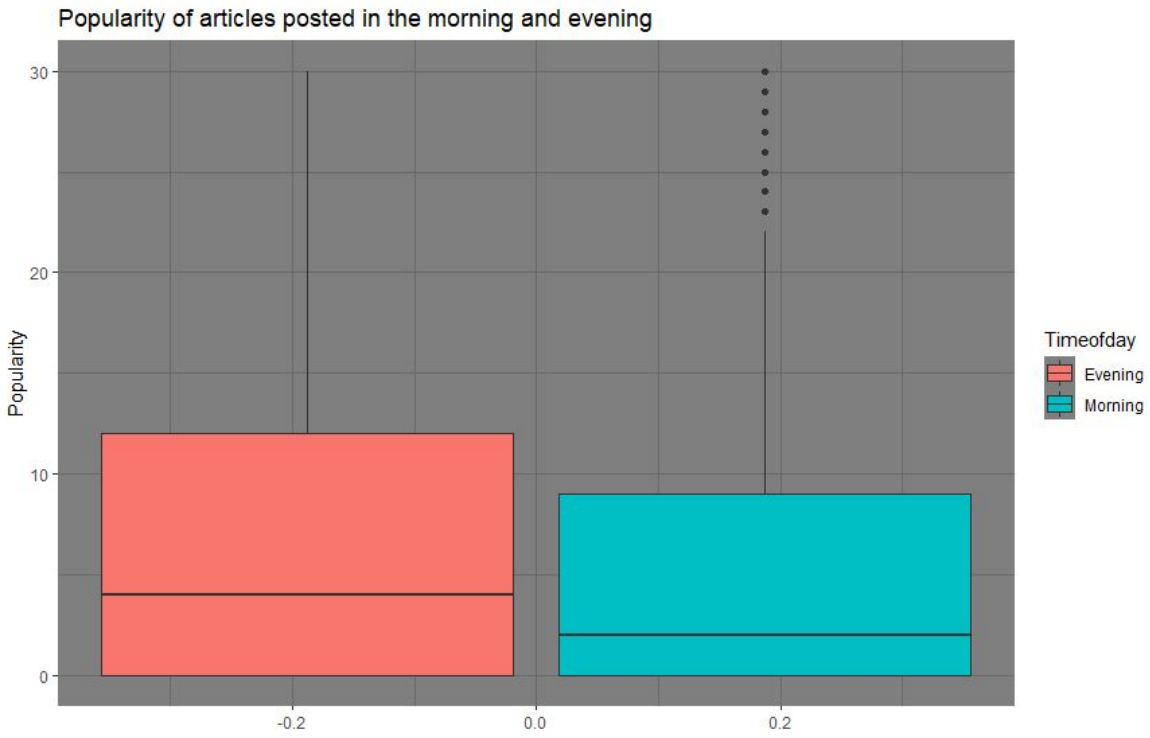


Figure 1. Boxplots of the popularity of news articles by time of day

2. Which news sources are the most popular?

The dataset contains 5755 unique news sources. In order to simplify our results, we only considered the 30 most popular sources. Like the previous question, we chose to look at the combined popularity of each article across the three social media platforms. From there, we then averaged the popularity scores of each news source's articles together, and then compared them from greatest to least, subsetting the results by the first 30 most popular news sources.

The resulting comparison yielded some interesting results; Liberty News Now, LinkedIn blogs, and Global Grind were much more popular than the proceeding twenty-seven news sources (Global Grind in third place having an average popularity of a little under 6000, versus American Kennel Club's average of just over 3000).

These results, while interesting, are also perhaps inaccurate. The popularity scores use a method that is best described as "mysterious" seeing as there is no information provided by the authors of the dataset. Several values for articles are "0" or "-1", perhaps denoting an NA value. In any case, the simple averaging method employed to answer this question may not have been the best

way to handle these scores, but given the lack of information, we decided to proceed with it. For future questions, we will pare down media sources by the amount of articles published rather than averaged popularity scores.

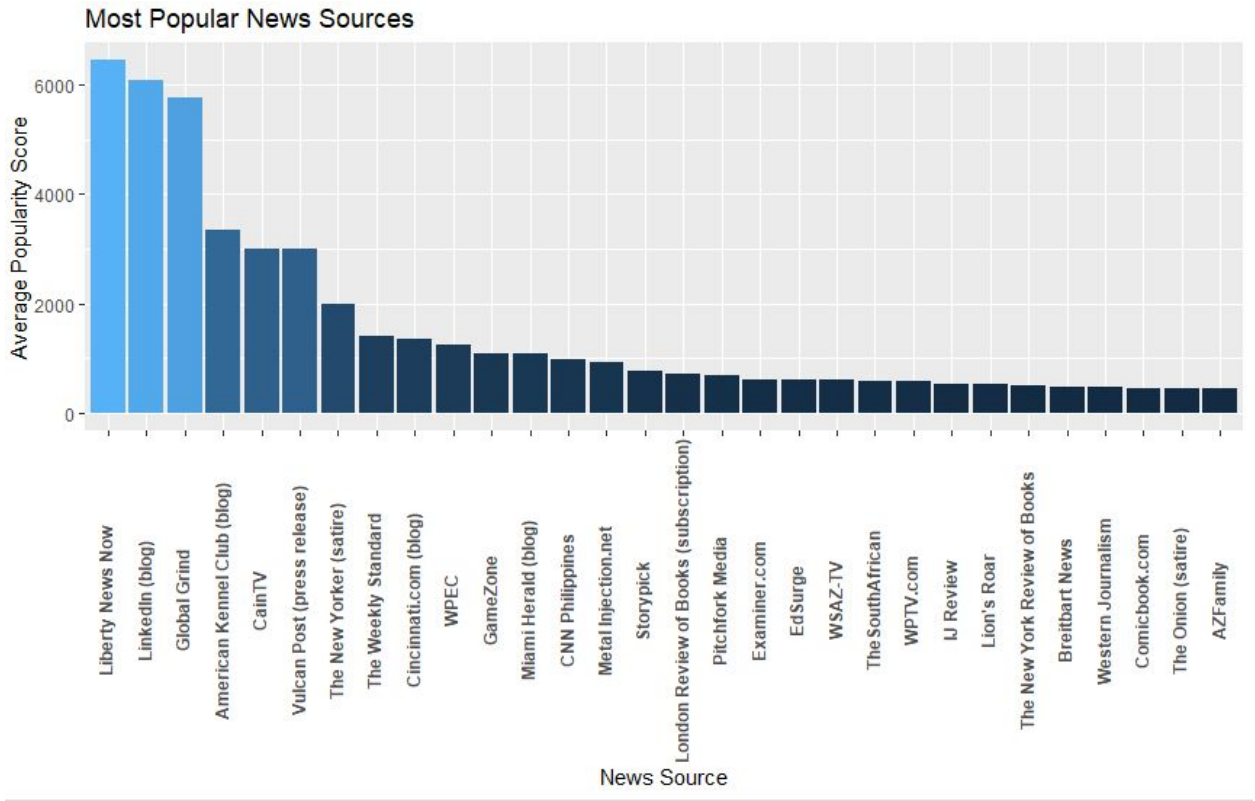


Figure 2. Comparison of news sources by average popularity score. Using this method, Liberty News Now, LinkedIn blogs, and Global Grind are by far the most popular news sources in the data, with disparities beginning to even out from Examiner.com onwards.

3. **Do certain popular news outlets receive better headline sentiment scores on the same topic than other popular news outlets?**

As stated in the previous question, we will be using the amount of articles published within the dataset to subset news sources based on popularity. The resulting 10 most popular news sources ranked in this way can be seen in figure 3 below.

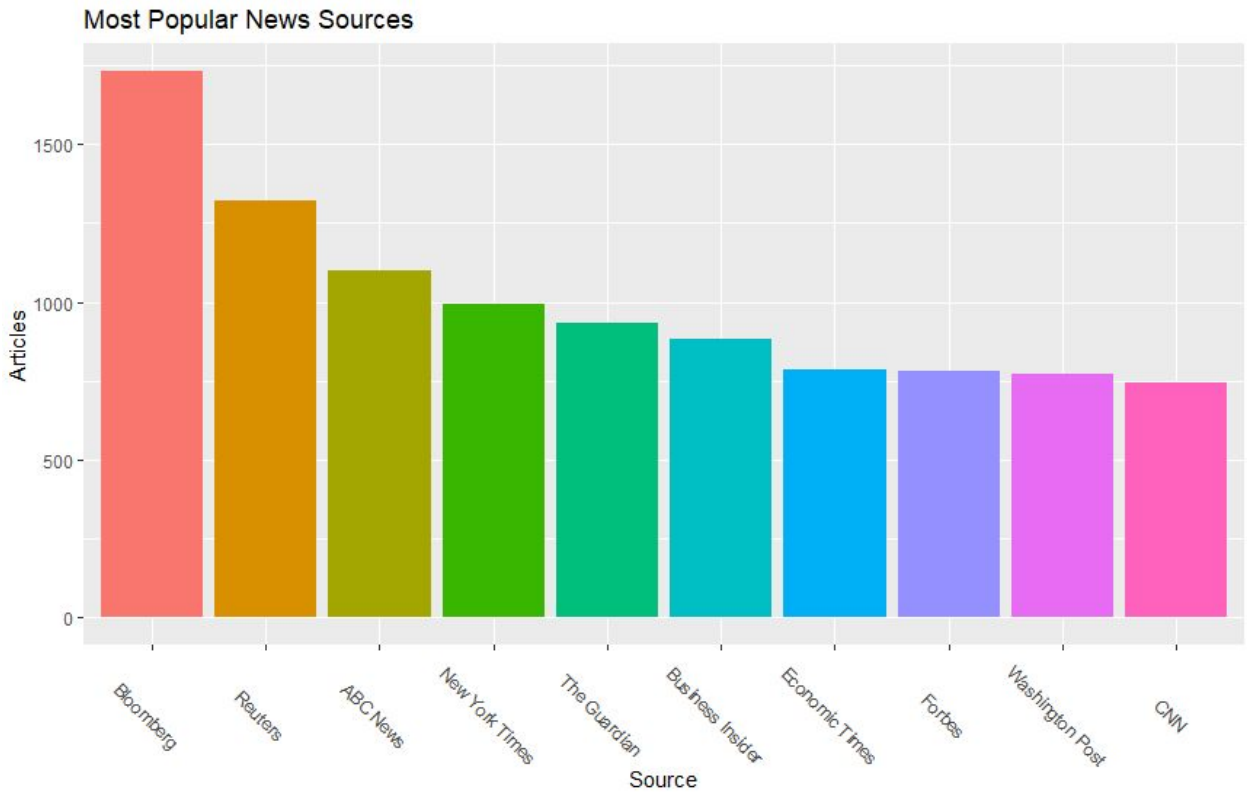


Figure 3. Comparison of news sources by total articles included in the data. Using this method, Bloomberg is by far the most published news souce.

From here, we then examined headline sentiment scores for each news outlet by each topic within the data (“economy”, “Microsoft”, “Obama” and “palestine”) in figures 4-7.

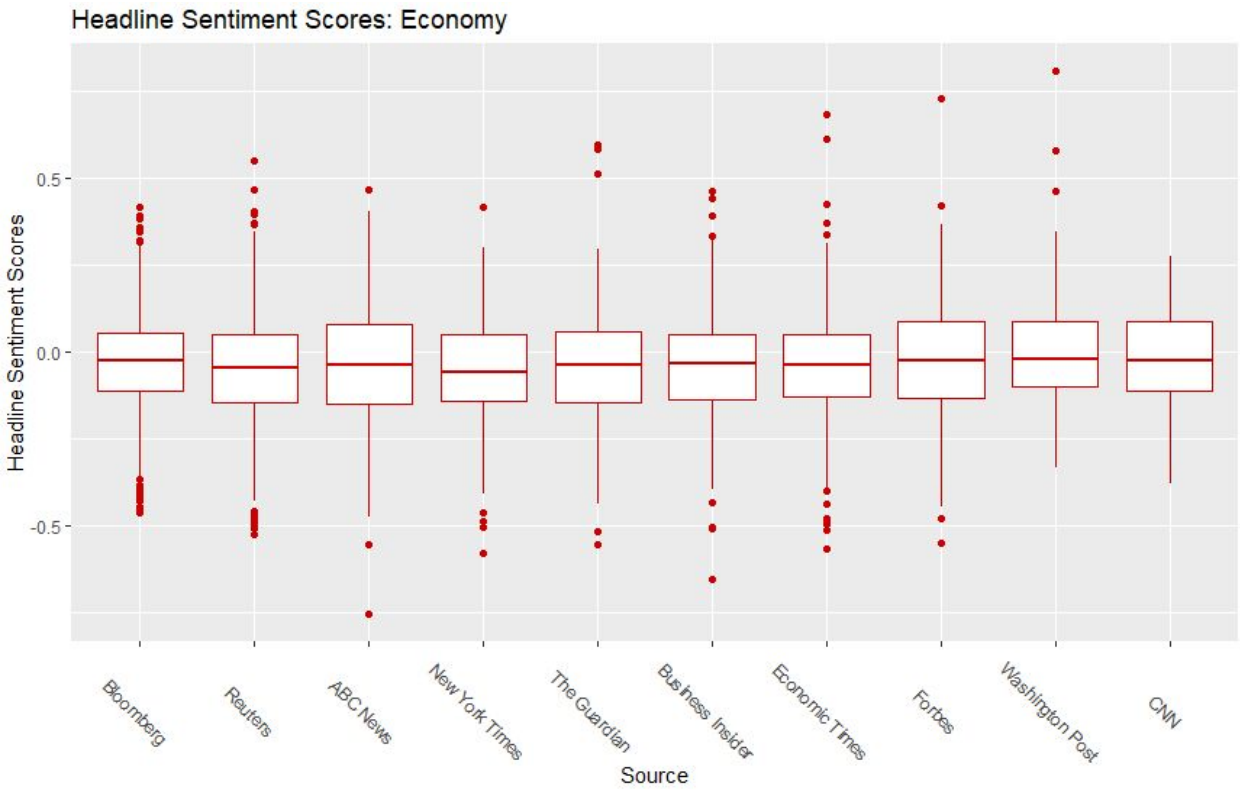


Figure 4. Comparison of headline sentiment scores by news source. For the economy topic, the medians and spreads of each news source's sentiment scores are very similar, all being the majority slightly negative, each source except CNN having its share of outliers.

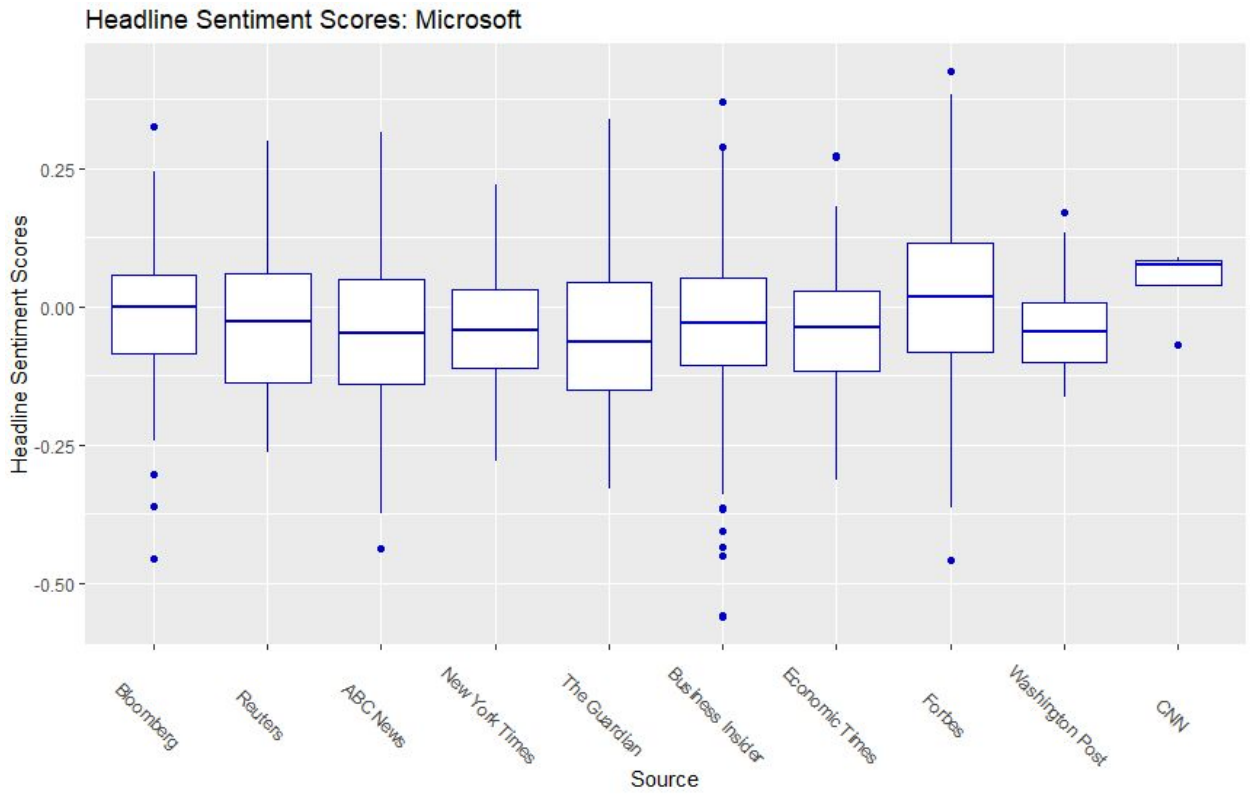


Figure 5. Comparison of headline sentiment scores by news source. For the Microsoft topic, the medians and spreads of each news source's sentiment scores are very similar, all being the majority slightly negative, with the exception of both CNN and Forbes. In the case of Forbes, its Median value is actually slightly positive, and the range of scores eclipses that of the other sources. All of CNN's scores are slightly positive, save for one outlier.

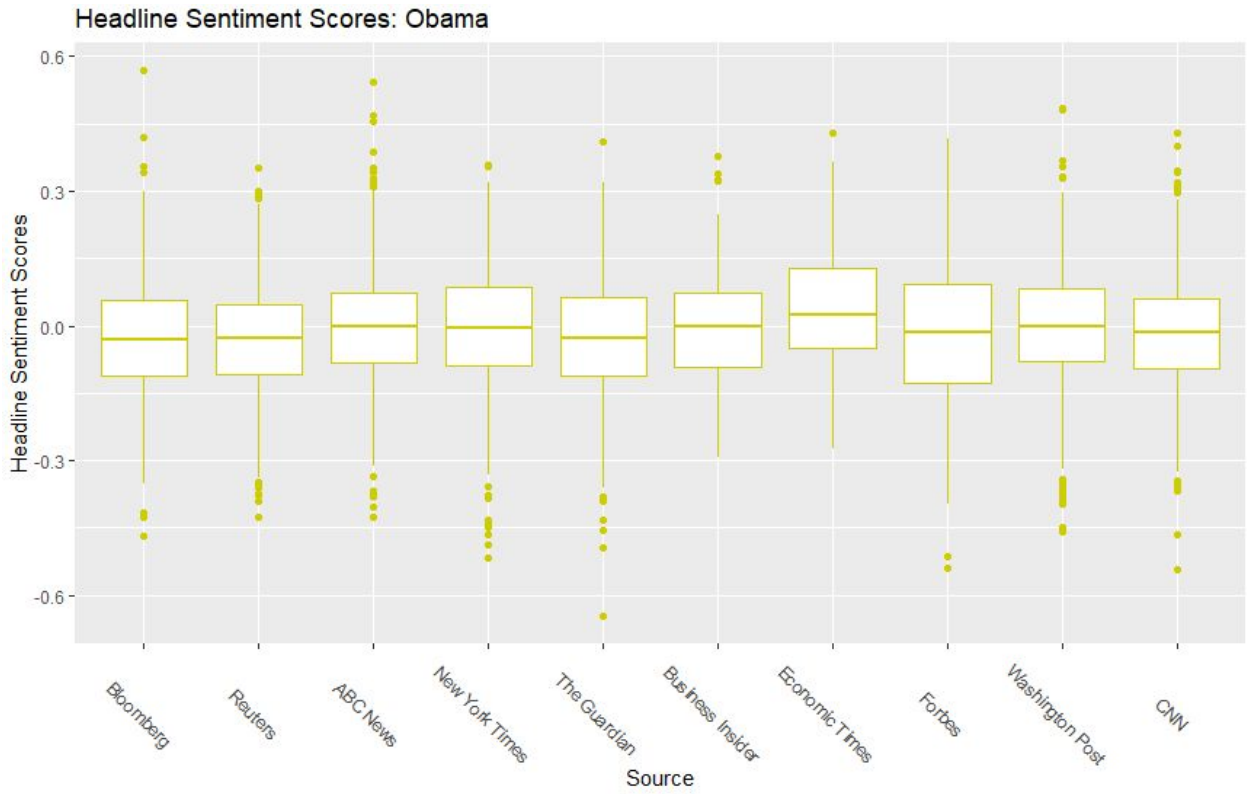


Figure 6. Comparison of headline sentiment scores by news source. For the Obama topic, the medians and spreads of each news source's sentiment scores are very similar, most being the majority slightly negative, though ABC News, New York Times, Business Insider, and Washington Post are centering neutral, and the Economic Times being slightly positive.

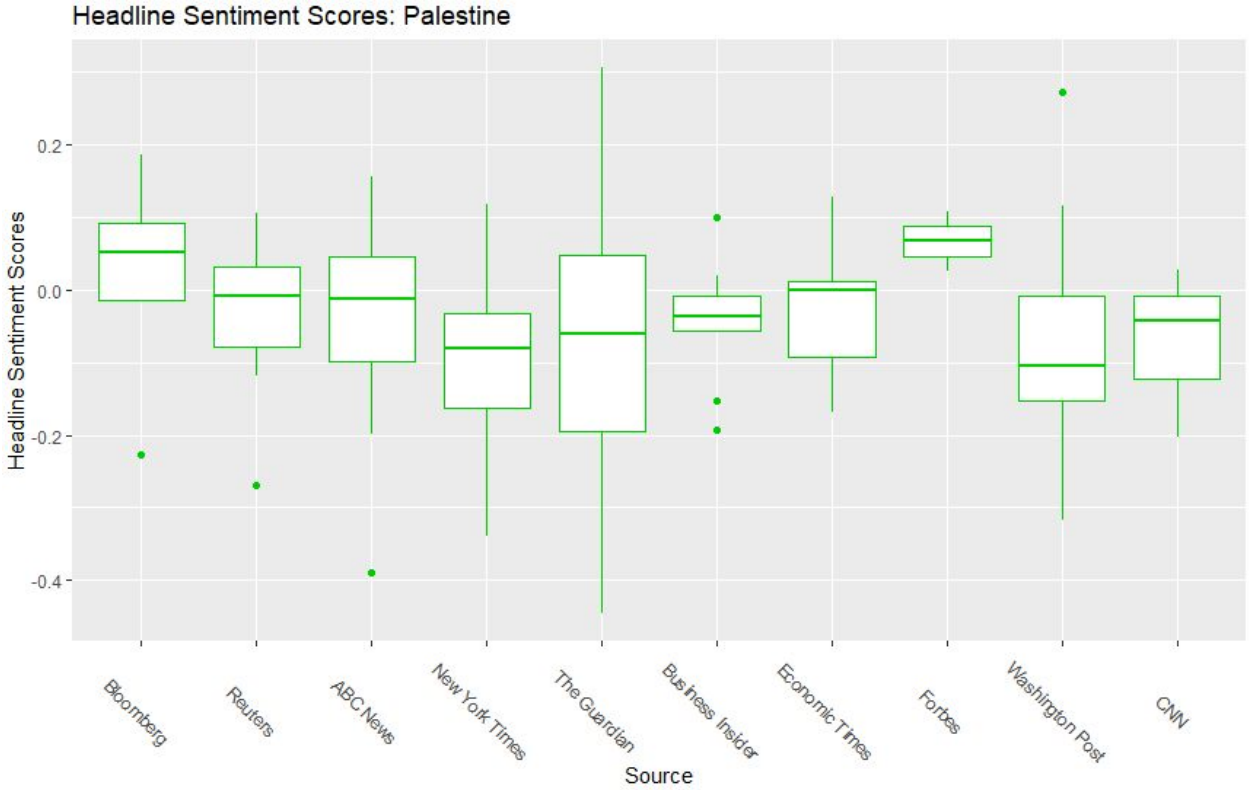


Figure 7. Comparison of headline sentiment scores by news source. For the Palestine topic, news sources do not agree on a trend. Bloomberg's articles are the most positive in sentiment score, whereas CNN's are mostly negative. The Guardian, also being mostly negative, has the greatest spread of all the news sources, with no outliers.

4. Are negative news articles more popular than positive articles?

This dataset contains various data on 93,329 articles that pertain to various topics (Economy, Microsoft, President Obama, and Palestine) across multiple social media platforms . In addition, there are variables (SentimentTitle and SentimentHeadline) which refer to the positivity and negativity of all the articles . Now, to determine the popularity of the negative news articles in relation to the sentiment headline variable, our group decided that for each of the social media platforms, we would take the sum of all the final popularity scores of the articles whose sentiment headline score was lower than zero. Likewise, to determine the popularity of the positive news articles in relation to the sentiment headline variable, our group would take the sum of all the final popularity scores of the articles whose sentiment headline score was greater than zero. Moreover, the same processes were followed to determine the popularity of the negative and positive news articles in relation to the sentiment title variable.

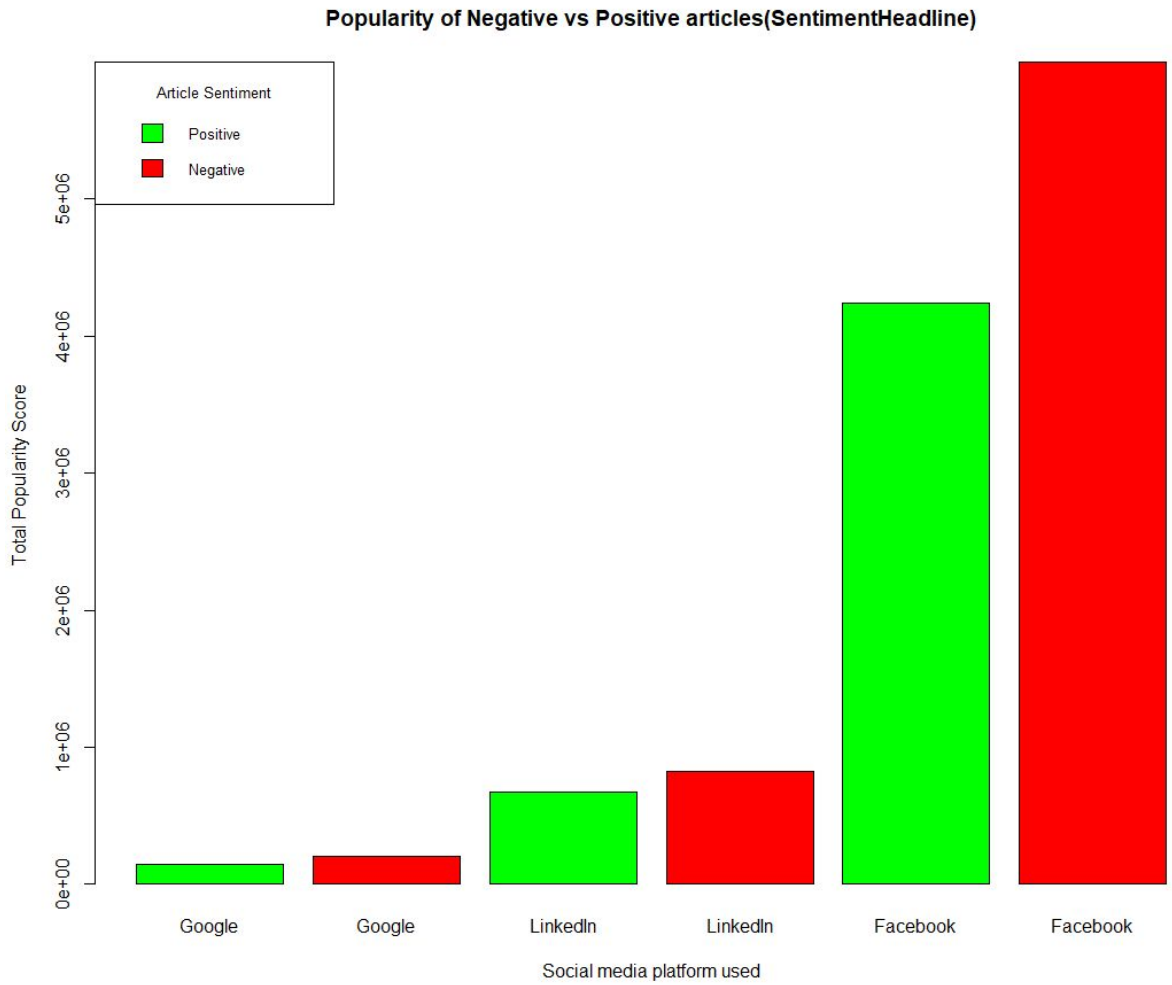


Figure 8. Comparison of total popularity scores of positive and negative articles among different platforms by means of sentiment headline value. For each of the social media platforms, the negative articles have a higher total popularity score than the positive articles. Facebook in particular has a great increase in total popularity scores in both positive and negative articles in comparison to the other social media platforms.

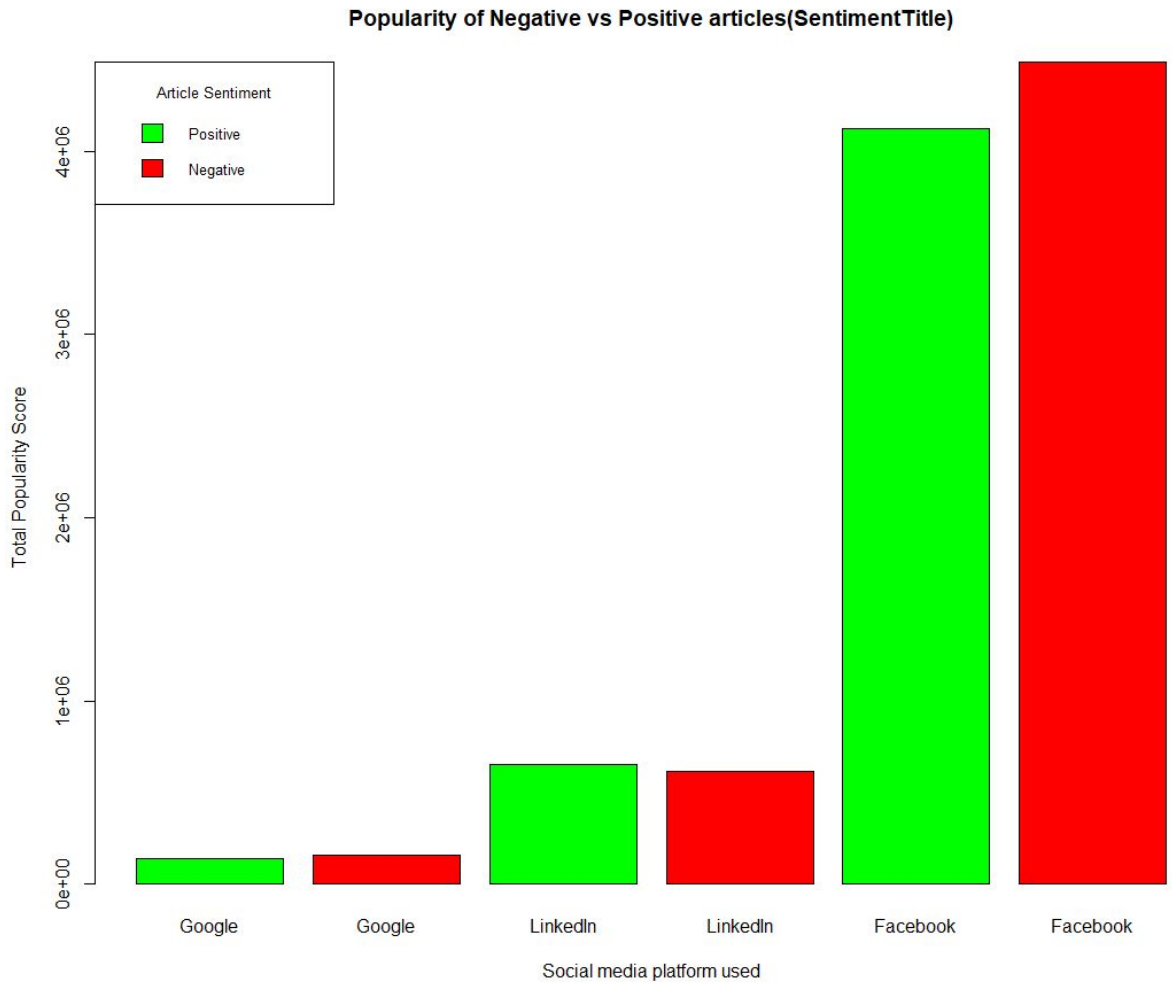


Figure 9. Comparison of total popularity scores of positive and negative articles among different platforms by means of sentiment title value. For each of the social media platforms, with exception to LinkedIn, the negative articles had a higher total popularity score in comparison to the positive articles. Similar to the previous bar plot, Facebook had particularly high total popularity scores in comparison to the other social media platforms.

Based on the results of the 2 bar plots above, it is clear that for the majority, negative news articles have a higher total popularity score among all social media platforms used (in exception to the LinkedIn negative news articles bar from the SentimentTitle graph).

Now one may wonder why we decided to incorporate both sentiment title and sentiment headline variables into our process of discovering whether negative articles are more popular than positive articles. The reason for doing so was because we did not want to leave any variables unused and thought it best

to ensure that our findings were supported/proven by more than one piece of data utilized.

Another finding our group found intriguing was why there was such a big gap in popularity score among the different social media platforms, particularly with Facebook in comparison to LinkedIn and GooglePlus. Now, one thing to keep in mind is that while all these articles are published on different platforms, there is one common ground each of these platforms and platforms share: Advertising. To be clear, while Facebook, LinkedIn, and GooglePlus may all distribute the same articles, there are multiple differences between all the platforms such as target audience, targeting options, cost per click, and ad types.

To begin, one can agree Facebook has a different audience in comparison to LinkedIn and GooglePlus. For example, according to “Battle of the Ads: Facebook vs Google vs LinkedIn” by Lana Macagapal of AdSpark, Facebook content is readily available and is targeted towards the typical ‘day - to day’ consumers who enjoy spending time with their friends and family , while GooglePlus content has to be manually searched by the consumer despite having the same target audience. On the other hand, LinkedIn, primarily focuses on professional news topics/articles and is mainly targeted towards people in the professional setting. In addition to target audience, Facebook has a lower cost per click and cost per ad among all three platforms. To be clear, according to Stephanie Mialki of Facebook Advertising, Facebook charges $0.51 per ad while LinkedIn charges $5.61. Moreover, the average cost per click for Google Ads are $1 - $2, slightly higher than Facebook and LinkedIn at $0.27 and $0.80 respectively. To conclude, one can see part of the reason Facebook has a higher total popularity rating in comparison to LinkedIn and GooglePlus is due to difference in target audience and inexpensive costs per click and costs per ad.

Now given the analysis of this question, there is still a possibility that the data may be inaccurate as there are numbers in the data which were not fully explained by the authors of the dataset. To be more specific, Facebook, GooglePlus, and LinkedIn final popularity variables, some articles had a final popularity rating of -1, which could potentially be seen as an NA value. However, with the lack of background information provided by the authors, we decided to perceive -1 as a value which indicated negative popularity. In order to improve on our research for future purposes, our group believes we should try to have a better understanding of what the -1 values signify to prevent any misrepresentation of data.

5. Which news topic is the most popular?

As stated in the previous question, there are 93,329 articles on a set of different topics: Economy, State of Palestine, President Obama, and Microsoft. Now, to determine which news topic is the most popular, our group decided to follow a similar process to the one used in the previous question. For example, our group decided to create a Facebook-economy variable which would take the sum of all the final popularity ratings of the articles whose topic was economy. Likewise, we created 11 more variables which would follow the same process for each of the other topics for each of the social media platforms. In this case, the sentiment title and sentiment headline variables were not utilized in the calculations.

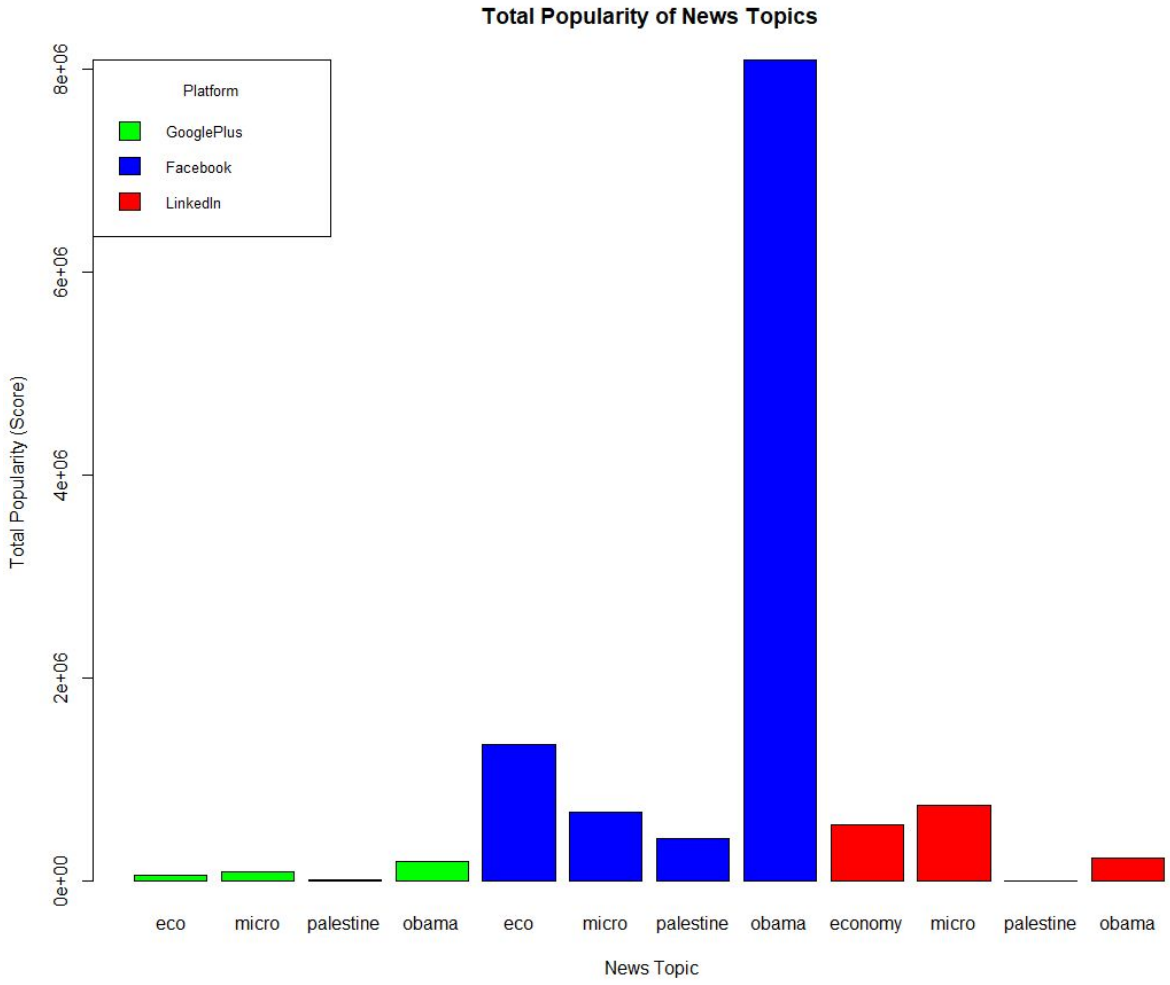


Figure 10. Comparison of total popularity scores of various articles on different topics among different platforms. For each of the social media platforms, with exception to LinkedIn, the Obama articles had a higher total popularity score in comparison to the

articles of other topics. Similar to the previous bar plot, Facebook had particularly high total popularity scores in comparison to the other social media platforms.

Based on the bar plot above, with exception to LinkedIn, articles on President Obama prove to have the highest total popularity score in comparison to other topics among the listed social media platforms. Facebook in particular has a high popularity score for each of the topics listed in comparison to other platforms.

One major finding we found interesting was how President Obama articles were the most popular articles among all the articles listed. Now after some background research, we believe we may be able to identify why President Obama articles are the most popular articles among all the topics listed. To give some background info, this dataset provides data on articles from November 2015 - July 2016. Keeping these dates in mind, the articles were published a year prior to President Obama completing his presidency. Moreover, according to obamawhitehouse.articles.gov, during the timespan of November - July 2016, there were many notable events which took place in relation to President Obama's presidency such as the participation of 2016 NATO summit, the executive order on gun control which would require background checks on those who purchased guns, and most notably the 2016 Presidential Election between Hillary Clinton and Donald Trump. With the election being a time when the majority of U.S. citizens are concentrated on who to elect as a future President-elect, it is possible to say President Obama articles were of great popularity due to the timing of the Presidential election.

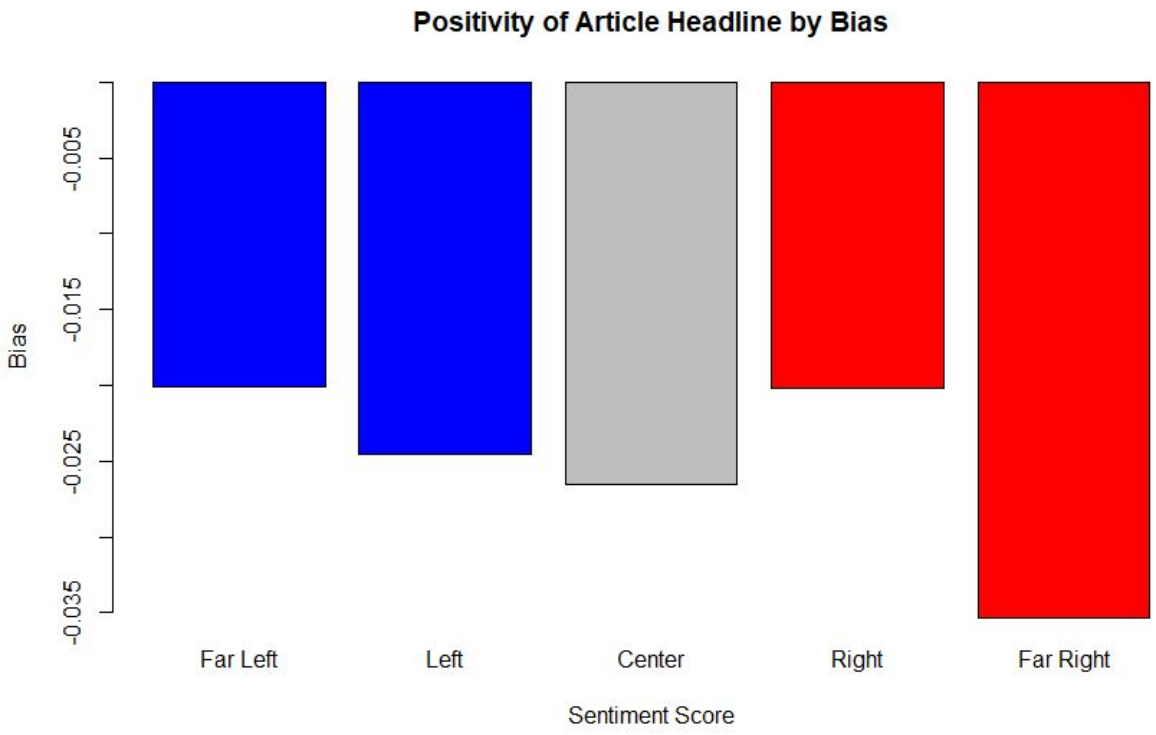
Given the analysis of this question, there is still reason to speculate that the data may be flawed due to the same reason that the previous question discussed could be flawed. In other words, because some of the final popularity ratings for some of the articles were -1, there was still confusion as to what that specific value entailed. However, our group still decided to interpret the -1 as a representation of negative popularity as opposed to positive popularity. Lastly, similar to the previous question, our group believes we should have a better understanding of what the -1 popularity value signifies to improve on research for future purposes.

6. **Is there a connection between political affiliation of news sources and negative/positive news?**

Media outlets often have personal biases associated with them whether that be from the way they report their news or what news they choose to report on. Figures 11 and 12 show the mean value of article headlines and titles grouped by the media outlets bias. All data recorded regarding media outlet bias was taken from a single source as to remain consistent and only the most popular media outlets ranked by articles posted were used.

The figures show a correlation with bias and sentiment scores for articles. News outlets with a more left leaning bias tended to have more positive headlines and titles to their articles than articles from more right leaning outlets. The figures show a nice smooth decline of positivity as you go from the left to the right. This may be due to one of the four news topics reported on within this dataset was President Barack Obama who was a left leaning democratic president. It is also interesting to note that only one of the ten scores recorded here turned out positive meaning most of all of the news reported on was negative news.

This finding is important to note because it gives some insight into what people are most interested in when it comes to consuming media. As we have mentioned earlier media outlets are businesses at their core and successful outlets know exactly what articles will get read and which ones won't. The law of supply and demand has led to this trend we can see in our figures of overwhelmingly negative media across the board.


Positivity of Article Headline by Bias
Bias
-0.005
-0.015
-0.025
-0.035
Far Left
Left
Center
Right
Far Right
Sentiment Score

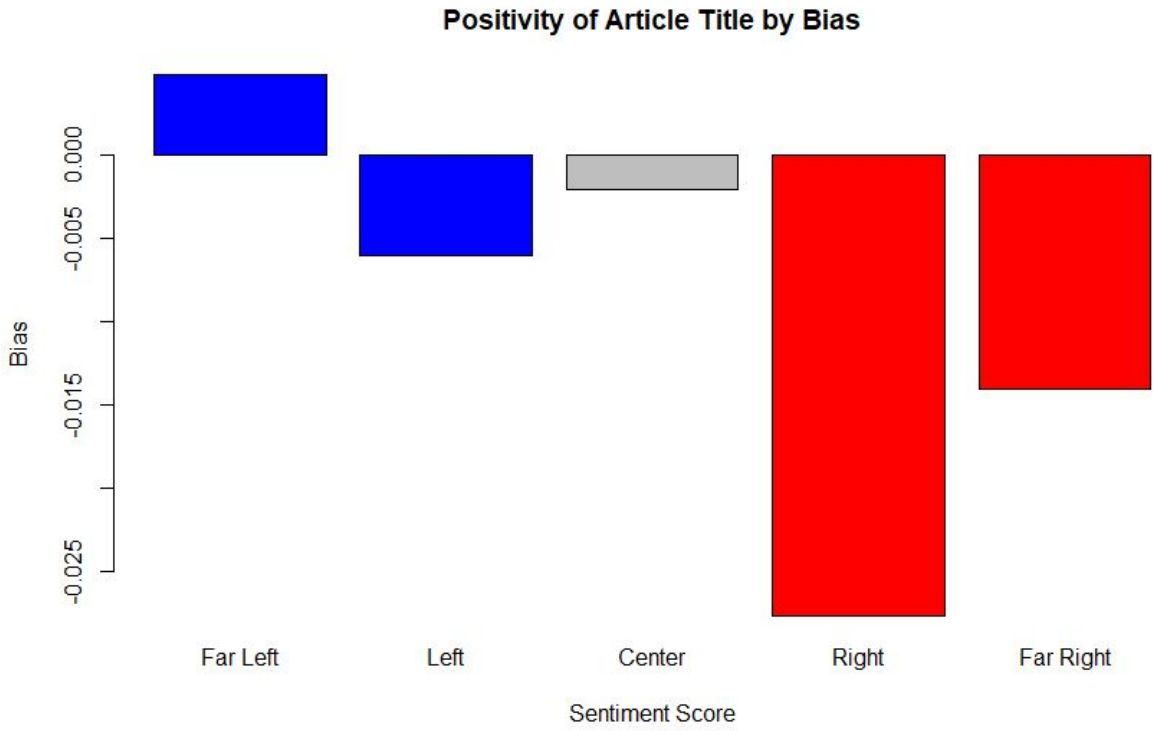


Figure 11 and 12 show comparisons of the positivity of left centered media sources to right centered media sources. The bars have been colored for better viewing and do not represent a variable.

Conclusion

Our group found several interesting results. Overall, articles that were posted in the evening appeared to be more popular than those in the morning. Several seemingly unlikely news sources ranked as the most popular. The most frequently published news articles did not have large differences in sentimentality scoring when the economy or Microsoft topics were involved, but differed slightly on the topic of President Obama and heavily on Palestine. In the case of all three social media platforms, negative articles were much more popular than positive ones. Finally, articles with a more politically-left-leaning bias tended to have higher headline sentiment scores than that of right-leaning.

What complicates all of these results is the perplexing lack of information given by the dataset authors about the popularity and sentiment scoring of the dataset, which, aside from article titles, topics, and publish dates, were the focal variables of our analyses. Based solely on our intuitions, we interpreted these scores to the best of our ability, but without even knowing the units of these scores, let alone the methodology of their acquisition, it is entirely possible we have misinterpreted them, their use, or their purpose. If this is the case, it could damage or destroy all of our findings as we have presented them. For this reason, the chief object of investigation should be these

scores, particularity the popularity scoring between each of the three social media platforms.

Nonetheless, this dataset proved interesting as an exercise for data exploration. The advent of using social media platforms to deliver news to consumers is a vitally important vector for many industries to be on top of (such as news media, media advertising, social networking, etc.), and the insight gleaned through examining such trends is not only academically interesting but can itself be put to use as well.

Code for Question 1 -

```
# How does time of day reported affect popularity?

# Creating a variable for the hour in the day published
News_Final$TimeHr <- substr(News_Final$PublishDate, 12, 13)
News_Final$TimeHr <- as.integer(News_Final$TimeHr)

# Indexing hour of the day into night, morning, and evening
index_night <- which(News_Final$TimeHr >= 20 | News_Final$TimeHr < 4)
index_morning <- which(News_Final$TimeHr >= 4 & News_Final$TimeHr < 12)
index_evening <- which(News_Final$TimeHr >= 12 & News_Final$TimeHr < 20)

# Initializing the time of day variable for use
News_Final$Timeofday <- 1:93239

# Using indexes to fill out the Timeofday variable day
News_Final$Timeofday[index_night] = "Night"
News_Final$Timeofday[index_morning] = "Morning"
News_Final$Timeofday[index_evening] = "Evening"

# Creating new variable combpop containing the combinded score from Facebook,
Googleplus, and LinkedIn for each article
News_Final$combpop = News_Final$Facebook + News_Final$GooglePlus +
News_Final$LinkedIn


#Finding the mean and median of combpop for Morning, Evening, and Night
mean(News_Final$combpop[which(News_Final$Timeofday == "Morning")])
```

```
mean(News_Final$combpop[which(News_Final$Timeofday == "Evening")])
mean(News_Final$combpop[which(News_Final$Timeofday == "Night")])

median(News_Final$combpop[which(News_Final$Timeofday == "Morning")])
median(News_Final$combpop[which(News_Final$Timeofday == "Evening")])
median(News_Final$combpop[which(News_Final$Timeofday == "Night")])

# The two are very different which tells us we have many outliers


# Creating boxplot of combpop for time of the day
boxplot(News_Final$combpop[which(News_Final$Timeofday == "Morning")], ylim = c(0,
50), main = "Popularity of articles posted in the morning", ylab = "popularity")
boxplot(News_Final$combpop[which(News_Final$Timeofday == "Evening")], ylim = c(0,
50), main = "Popularity of articles posted in the evening", ylab = "popularity" )
boxplot(News_Final$combpop[which(News_Final$Timeofday == "Night")], ylim = c(0,
50), main = "Popularity of articles posted at night", ylab = "popularity")
```

Code to get figure 1

```
ggplot(data = subset(News_Final, Timeofday == "Evening" | Timeofday == "Morning")) +
geom_boxplot(mapping = aes(fill = Timeofday, y = combpop)) +
scale_y_continuous(limits = c(0,30)) + ggtitle("Popularity of articles posted in the
morning and evening") + labs(y = "Popularity") + theme_dark()
```

Code for Question 2 -

```
news <- News_Final

# factorizing news source variable
table(news$Source)
news$Source <- as.factor(news$Source)

# factorizing news topic variable
table(news$Topic)
news$Topic <- as.factor(news$Topic)

# creating new variable, Popularity, as average of 3 platform popularity variables
news$Popularity <- (news$Facebook + news$GooglePlus + news$LinkedIn) / 3

source_names <- levels(news$Source)
```

```
#initialize source popularity vector
source_popularity <- c()

#initialize source number
source_number <- 1

# the goal is to populate the vector source_popularity with an average popularity
# value for every article from that news source. to do this, we need to
# first create an index of every article's average popularity from a given news source,
# average these popularity values together, and then add this value to the
# popularity vector. we can use a while loop to do this like so:
while(length(source_popularity) < length(source_names)) {
  pop_index <- news$Popularity[which(news$Source ==
source_names[source_number])]
  pop_avg <- sum(pop_index) / length(pop_index)
  source_popularity <- append(source_popularity, pop_avg, after =
length(source_popularity))
  source_number <- source_number + 1
}

# create new data frame with the source names, each source's popularity, and rank
each source's popularity as a variable
news_pop = data.frame(source = source_names, popularity = source_popularity, rank =
rank(source_popularity))

# subset the data to include only the 30 highest ranked news sources
news_most_pop <- subset(news_pop, rank > 5725)

# draw a barplot of this information! Code for figure 2
ggplot(data = news_most_pop, aes(x = reorder(source, -popularity), y = popularity, fill =
popularity)) + geom_bar(stat = "identity") + labs(x = "News Source", y = "Average
Popularity Score", title = "Most Popular News Sources") + theme(axis.text.x =
element_text(face = "bold", angle = 90, vjust = 0.25), legend.position = "none")
```

Code for Question 3

```
news <- News_Final

# factorizing news source variable
```

```
table(news$Source)
news$Source <- as.factor(news$Source)

# factorizing news topic variable
table(news$Topic)
news$Topic <- as.factor(news$Topic)

news$Source <- factor(news$Source, levels = names(sort(table(news$Source),
decreasing = TRUE)))

# subset news to include only the 10 most published news sources
news2 <- subset(news, Source == "Bloomberg" | Source == "Reuters" | Source ==
"ABC News" | Source == "New York Times" | Source == "The Guardian" | Source ==
"Business Insider" | Source == "Economic Times" | Source == "Forbes" | Source ==
"Washington Post" | Source == "CNN")

# create a barplot of this (Figure 3)
ggplot(data = news2) + geom_bar(mapping = aes(x = Source, fill = Source)) + labs(y =
"Articles", title = "Most Popular News Sources") + theme(axis.text.x =
element_text(angle = -45), legend.position = "none")

# subset news2 to include only articles on the economy topic
news_economy <- subset(news2, Topic == "economy")

# create boxplots of the headline sentiment scores for each source on the economy
topic (Figure 4)
ggplot(data = news_economy) + geom_boxplot(mapping = aes(x = Source, y =
SentimentHeadline), color = "red3") + labs(y = "Headline Sentiment Scores", title =
"Headline Sentiment Scores: Economy") + theme(axis.text.x = element_text(angle =
-45), legend.position = "none")

# subset news2 to include only articles on the microsoft topic
news_microsoft <- subset(news2, Topic == "microsoft")

# create boxplots of the headline sentiment scores for each source on the microsoft
topic (Figure 5)
ggplot(data = news_microsoft) + geom_boxplot(mapping = aes(x = Source, y =
SentimentHeadline), color = "blue3") + labs(y = "Headline Sentiment Scores", title =
```

"Headline Sentiment Scores: Microsoft") + theme(axis.text.x = element_text(angle = -45), legend.position = "none")

# subset news2 to include only articles on the obama topic
news_obama <- subset(news2, Topic == "obama")

# create boxplots of the headline sentiment scores for each source on the economy topic **(Figure 6)**
ggplot(data = news_obama) + geom_boxplot(mapping = aes(x = Source, y = SentimentHeadline), color = "yellow3") + labs(y = "Headline Sentiment Scores", title = "Headline Sentiment Scores: Obama") + theme(axis.text.x = element_text(angle = -45), legend.position = "none")

# subset news2 to include only articles on the palestine topic
news_palestine <- subset(news2, Topic == "palestine")

# create boxplots of the headline sentiment scores for each source on the palestine topic **(Figure 7)**
ggplot(data = news_palestine) + geom_boxplot(mapping = aes(x = Source, y = SentimentHeadline), color = "green3") + labs(y = "Headline Sentiment Scores", title = "Headline Sentiment Scores: Palestine") + theme(axis.text.x = element_text(angle = -45), legend.position = "none")

Code for Question 4 -

creating a News_Final variable to allow for R-Studio to read the comma separated value chart from the UCI Machine learning dataset

News_Final <- read.csv("News_Final.csv")

#create variables which will retrieve the total popularity scores of positive and negative articles for each platform based on sentiment headline score

GooglePlus_Positive_Headline <- sum(News_Final$GooglePlus[which(News_Final$SentimentHeadline > 0)]) #Positive GooglePlus Articles

```
GooglePlus_Negative_Headline <-
sum(News_Final$GooglePlus[which(News_Final$SentimentHeadline < 0)]) #Negative
GooglePlus Articles

LinkedIn_Positive_Headline <-
sum(News_Final$LinkedIn[which(News_Final$SentimentHeadline > 0)]) #Positive
LinkedIn Articles

LinkedIn_Negative_Heaadline <-
sum(News_Final$LinkedIn[which(News_Final$SentimentHeadline < 0)]) #Negative
LinkedIn Articles

Facebook_Positive_Headline <-
sum(News_Final$Facebook[which(News_Final$SentimentHeadline > 0)]) #Positive
Facebook Articles

Facebook_Negative_Headline <-
sum(News_Final$Facebook[which(News_Final$SentimentHeadline < 0)]) #Negative
Facebook Articles

#create variables which will retrieve the total popularity scores of positive and negative
articles for each platform based on sentiment title score

GooglePlus_Positive_Title <-
sum(News_Final$GooglePlus[which(News_Final$SentimentTitle > 0)]) #Positive
GooglePlus Articles

GooglePlus_Negative_Title <-
sum(News_Final$GooglePlus[which(News_Final$SentimentTitle < 0)]) #Negative
GooglePlus Articles

LinkedIn_Positive_Title <- sum(News_Final$LinkedIn[which(News_Final$SentimentTitle
> 0)]) #Positive LinkedIn Articles

LinkedIn_Negative_Title <-
sum(News_Final$LinkedIn[which(News_Final$SentimentTitle < 0)]) #Negative LinkedIn
Articles
```

```
Facebook_Positive_Title <-
sum(News_Final$Facebook[which(News_Final$SentimentTitle > 0)]) #Positive
Facebook Articles
Facebook_Negative_Title <-
sum(News_Final$Facebook[which(News_Final$SentimentTitle < 0)]) #Negative
Facebook Articles
```

Code for Figure 8 -

```
# Create a bar plot to represent the popularity scores for negative and positive articles
among all Social media platforms

Negative_Positive_Popularity <- c("Google", "Google", "LinkedIn", "LinkedIn",
"Facebook", "Facebook") # x-axis/ Social Media Platform Names

Popularity_score <- c(GooglePlus_Positive_Headline, GooglePlus_Negative_Headline,
LinkedIn_Positive_Headline, LinkedIn_Negative_Heaadline,
Facebook_Positive_Headline, Facebook_Negative_Headline) #Total Popularity Score
for each of the positive and negative articles for each platform

legend <- c("Positive", "Negative") #Legend which differentiates between positive and
negative articles for each social media platform

my_colors <- c("green", "red") #Colors used to represent popularity and negativity

barplot(Popularity_score, main = "Popularity of Negative vs Positive
articles(SentimentHeadline)" , xlab = "Social media platform used", ylab = " Total
Popularity Score", names.arg = Negative_Positive_Popularity, col = my_colors, beside
= TRUE)

legend("topleft", legend = c("Positive", "Negative"), title = "Article Sentiment", fill =
my_colors, cex = 0.8) #Legend used to help explain the bar plot
```

Code for Figure 9 -

```
# Create a second bar plot to represent the popularity scores for negative and positive articles among all Social media platforms by Sentiment Title

Second_Popularity_Score <- c(GooglePlus_Positive_Title, GooglePlus_Negative_Title, LinkedIn_Positive_Title, LinkedIn_Negative_Title, Facebook_Positive_Title, Facebook_Negative_Title) #Total Popularity Score for each of the positive and negative articles for each platform by sentiment title

barplot(Second_Popularity_Score, main = "Popularity of Negative vs Positive articles(SentimentTitle)" , xlab = "Social media platform used", ylab = "Total Popularity Score", names.arg = Negative_Positive_Popularity, col = my_colors, beside = TRUE)

legend("topleft", legend = c("Positive", "Negative"), title = "Article Sentiment", fill = my_colors, cex = 0.8) #Legend used to help explain the bar plot
```

Code for Question 5-

```
#Create variables which will receive the total popularity scores of economy, microsoft, palestine, and obama articles for each social media platform

economy_popularity_google <- sum(News_Final$GooglePlus[which(News_Final$Topic == "economy")]) #Popularity of economy articles on GooglePlus

microsoft_popularity_google <- sum(News_Final$GooglePlus[which(News_Final$Topic == "microsoft")]) #Popularity of microsoft articles on GooglePlus

palestine_popularity_google <- sum(News_Final$GooglePlus[which(News_Final$Topic == "palestine")])#Popularity of palestine articles on GooglePlus

obama_popularity_google <- sum(News_Final$GooglePlus[which(News_Final$Topic == "obama")]) #Popularity of obama articles on GooglePlus

economy_popularity_Facebook <- sum(News_Final$Facebook[which(News_Final$Topic == "economy")]) #Popularity of economy articles on Facebook
```

```
microsoft_popularity_Facebook <-
sum(News_Final$Facebook[which(News_Final$Topic == "microsoft")]) #Popularity of
microsoft articles on Facebook

palestine_popularity_Facebook <-
sum(News_Final$Facebook[which(News_Final$Topic == "palestine")]) #Popularity of
palestine articles on Facebook

obama_popularity_Facebook <- sum(News_Final$Facebook[which(News_Final$Topic
== "obama")]) #Popularity of obama articles on Facebook

economy_popularity_LinkedIn <- sum(News_Final$LinkedIn[which(News_Final$Topic
== "economy")]) #Popularity of economy articles on LinkedIn

microsoft_popularity_LinkedIn <- sum(News_Final$LinkedIn[which(News_Final$Topic
== "microsoft")]) #Popularity of microsoft articles on LinkedIn

palestine_popularity_LinkedIn <- sum(News_Final$LinkedIn[which(News_Final$Topic
== "palestine")]) #Popularity of palestine articles on LinkedIn

obama_popularity_LinkedIn <- sum(News_Final$LinkedIn[which(News_Final$Topic ==
"obama")]) #Popularity of obama articles on LinkedIn
```

Code for Figure 10 -

```
# Create a bar plot which resembles the total popularity scores of articles of different
topics among different platforms

Topic_Names <- c("eco", "micro", "palestine", "obama", "eco", "micro", "palestine",
"obama", "economy", "micro", "palestine", "obama") #Topic names listed on x-axis

Topic_Popularity <- c(economy_popularity_google, microsoft_popularity_google,
palestine_popularity_google, obama_popularity_google,
economy_popularity_Facebook, microsoft_popularity_Facebook,
palestine_popularity_Facebook, obama_popularity_Facebook,
economy_popularity_LinkedIn, microsoft_popularity_LinkedIn,
palestine_popularity_LinkedIn, obama_popularity_LinkedIn)
#Total Popularity Scores of each News Topic for each of the social media platforms
```

plot_colors <- c("green", "green", "green", "green", "blue", "blue", "blue", "blue", "red", "red", "red", "red") #colors used to represent the social media platforms

legend_colors <- c("green", "blue", "red") #Legend colors which correspond to plot_colors to help explain the colors in the bar plot

barplot(Topic_Popularity, main = "Total Popularity of News Topics" , xlab = " News Topic", ylab = " Total Popularity (Score)", names.arg = Topic_Names, col = plot_colors, beside = TRUE)

legend("topleft", legend = c("GooglePlus", "Facebook", "LinkedIn"), title = "Platform", fill = legend_colors, cex = 0.8) #legend which is used to explain the bar plot

Code for Question 6 -

```
# Taking the top 20 news sources from the data

sort(table(News_Final$Source), decreasing = T)[1:20]

# Assigning political bias to each news outlet from outside source https://www.allsides.com/media-bias/media-bias-ratings?field_featured_bias_rating_value=All&field_news_source_type_tid[1]=1&field_news_source_type_tid[2]=2&field_news_source_type_tid[3]=3&field_news_source_type_tid[4]=4&field_news_bias_nid_1[1]=1&field_news_bias_nid_1[2]=2&field_news_bias_nid_1[3]=3&title=ZDNet

# Initializing the bias variable for use

News_Final$bias <- 1:93239

# Assigning news sources their bias

News_Final$bias[which((News_Final$Source == "Bloomberg") | (News_Final$Source == "ABC News") | (News_Final$Source == "New York Times") | (News_Final$Source == "The Guardian") | (News_Final$Source == "Economic Times") | (News_Final$Source == "Washington Post") | (News_Final$Source == "CNN"))] <- "Left"

News_Final$bias[which((News_Final$Source == "Reuters") | (News_Final$Source == "Business Insider") | (News_Final$Source == "Forbes") | (News_Final$Source == "Wall Street Journal") | (News_Final$Source == "Wall Street Journal") | (News_Final$Source == "WinBeta") | (News_Final$Source == "CNBC") | (News_Final$Source == "Reuters via Yahoo! Finance") | (News_Final$Source == "The Hill") | (News_Final$Source == "Financial Times") | (News_Final$Source == "USA TODAY") | (News_Final$Source == "ZDNet"))] <- "Center"
```

```
News_Final$bias[which((News_Final$Source == "Breitbart News") | (News_Final$Source == "Daily Mail")
| (News_Final$Source == "Fox News") | (News_Final$Source == "Daily Caller"))] <- "Far Right"

News_Final$bias[which(News_Final$Source == "Huffington Post")] <- "Far Left"

News_Final$bias[which((News_Final$Source == "Washington Times") | (News_Final$Source ==
"MarketWatch"))] <- "Right"

# Calculating Means

Far_Left_H <- mean(News_Final$SentimentHeadline[which(News_Final$bias == "Far Left")])
[1] -0.02008842
Left_H <- mean(News_Final$SentimentHeadline[which(News_Final$bias == "Left")])
[1] -0.02457964
Center_H <- mean(News_Final$SentimentHeadline[which(News_Final$bias == "Center")])
[1] -0.02650299
Right_H <- mean(News_Final$SentimentHeadline[which(News_Final$bias == "Right")])
[1] -0.02020408
Far_Right_H <- mean(News_Final$SentimentHeadline[which(News_Final$bias == "Far Right")])
[1] -0.0353037
Far_Left_T <- mean(News_Final$SentimentTitle[which(News_Final$bias == "Far Left")])
[1] 0.004821679
Left_T <- mean(News_Final$SentimentTitle[which(News_Final$bias == "Left")])
[1] -0.006061708
Center_T <- mean(News_Final$SentimentTitle[which(News_Final$bias == "Center")])
[1] -0.002105816
Right_T <- mean(News_Final$SentimentTitle[which(News_Final$bias == "Right")])
[1] -0.02766623
Far_Right_T <- mean(News_Final$SentimentTitle[which(News_Final$bias == "Far Right")])
[1] -0.0140379

x_H <- c(Far_Left_H, Left_H, Center_H, Right_H, Far_Right_H)
names(x_H) <- c("Far Left", "Left", "Center", "Right", "Far Right")
```

Code for figure 11

```
barplot(x_H, main = "Positivity of Article Headline by Bias", ylab = "Bias", xlab = "Sentiment Score", col =
c("Blue", "Blue", "Gray", "Red", "Red"))


x_T <- c(Far_Left_T, Left_T, Center_T, Right_T, Far_Right_T)
names(x_T) <- c("Far Left", "Left", "Center", "Right", "Far Right")
```

Code for figure 12

```
barplot(x_T, , main = "Positivity of Article Title by Bias", ylab = "Bias", xlab = "Sentiment Score", col =
c("Blue", "Blue", "Gray", "Red", "Red") )
```