Assessing Party Bias in CNN News

2023-05-10

Introduction

Party polarization and ideological divides are becoming more prominent in recent years (Pew Research Center, 2014). One of the indicators of party polarization is polarized trust in media along the partisan lines. 70% of Democrats trust the media, whereas less than 30% of Republicans share the same sentiment (Brenan, 2022). Similarly, the readers of mainstream national news sources are divided along partisan lines as well. Study conducted by Pew Research center in 2020 shows that among all people who resort to CNN as their primary political news source, nearly 80% identify themselves as Democrats (Grieco, 2020). In other words, disproportionately more Democrats choose and trust CNN than Republicans. The question, therefore, is what about CNN that attracts a Democrats' audience pool. Does CNN prime information to align with the Democratic priorities? This project sets out to examine if CNN is creating a Democratic-specific information environment. Does CNN exhibit party bias by reporting the Democratic Party more favorably than the Republican party?

Method

Sentiment analysis

As part of natural language processing (NLP), sentiment analysis allows users to identify the emotion in text. It is widely used in market and healthcare research to evaluate clients' experience (Wankahade et al., 2022). By using text-mining tools and different lexicons, this study compares the sentiment in Democrat, Republican, and Obama-related news texts. These texts are first tokenized to words and then assigned a categorical class. As long as there is one occurrence of the word "democrat", "republican", or "obama", the text is labelled as such-related. Sentiment analysis is applied to assess the emotion of each Democrat-, Republican-, and Obama-related text. The sentiment could be positive, negative, or neutral based on different sentiment analysis method (the Afinn sentiment score, Bing vocabulary, and NRC lexicon).

Inverse document frequency

Inverse Document Frequency (TF-IDF) is used to uncover the unique and important terms in Democrat-concerning and Republican-concerning texts respectively. It has been supported that TF-IDF is more accurate in producing the important terms (Wankahade et al., 2022). The equation of IDF is shown below in Equation 1. *Equation* 1. $IDF(term) = ln\ (\frac{Total\ number\ of\ documents\ t\ present\ in\ a\ document}{Number\ of\ documents\ containing\ term})$

Topic modeling

Latent Dirichlet Allocation (LDA) topic modeling can estimate the topics in the corpus. The number of topics estimated in this study is two, and the log value can be used to identify the words falling under each topics. Examining if the words differ along partisan lines would allow us to determine if the two topics are politically-related or unrelated. *Equation* 2. $log\ ratio = \frac{beta\ 1}{beta\ 2}$

List of packages used

The following packages used to perform data wrangling, the pre-processing of the text data, text mining, data visualization, and topic modeling.

```
library(dplyr) library(stringr) library(tidytext) library(tidyverse) library(tidyr) library(tm)
library(ggplot2) library(wordcloud) library(reshape2) library(ggwordcloud)
library (topicmodels)
```

Data Overview

The data for this study is credited to Qian & Zhai. They scraped seven categories of CNN news, including crime, entertainment, health, living, politics, technology, and travel from Jan. 1st - Apr. 4th, 2014. Since this project is on party bias, only the politics category is used. In total, 409 pieces of CNN politics news are included in the data set. The original variables in data files include URL, title, abstract, text. For the sake of this study, only text and document number are kept.

Data was pre-processed first by transforming the text column into a tidytext format so that each observation is a single token/word. Then, stop words are also removed. After pre-processing, the total of 194,696 observations were generated. Finally, replacements are made. The plural form of the word "republicans" and "democrats" with the singular form "republican" and "democrat". The merge of "democrat" and "democrats" should not implicate "democratic", because democratic is a multiple-meaning word. While "democratic" is relevant with the Democratic Party, it can also be used to describe democracy, irrelevant to this study. The word "obama's" and "obamacare" are replaced with "obama" because the question is how often Obama is mentioned and when a text talks about Obama, what the sentiment is in that text. Both "obama's" and "obamacare" are about President Obama, and these replacement maximize the number of news text under the label "obama" for analysis. The project is interested in the sentiment difference between texts under the three different labels "republican", "democrat", and "obama".

```
CNN_politics<-CNN %>%
  mutate(docno = 1:nrow(CNN)) %>%
  unnest_tokens(output = 'word', input = 'text') %>%
  mutate(word = str_replace_all(word, "republicans", "republican")) %>%
  mutate(word = str_replace_all(word, "democrats", "democrat")) %>%
  mutate(word = str_replace_all(word, "obama's", "obama")) %>%
  mutate(word = str_replace_all(word, "obamacare", "obama")) %>%
  anti_join(stop_words)
```

Word Frequency

Fig. 1 CNN Politics Wordcloud

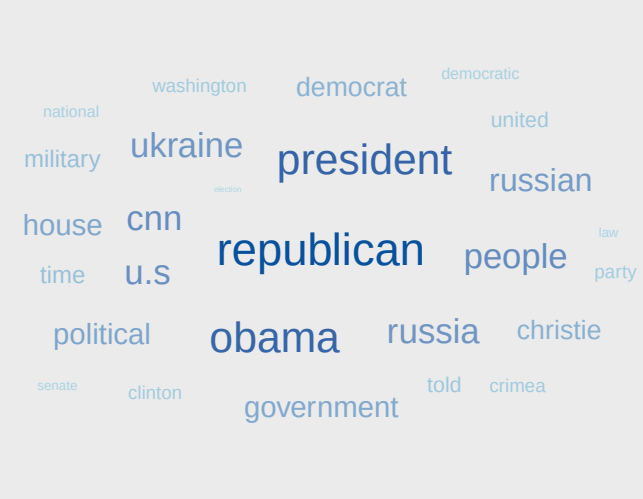
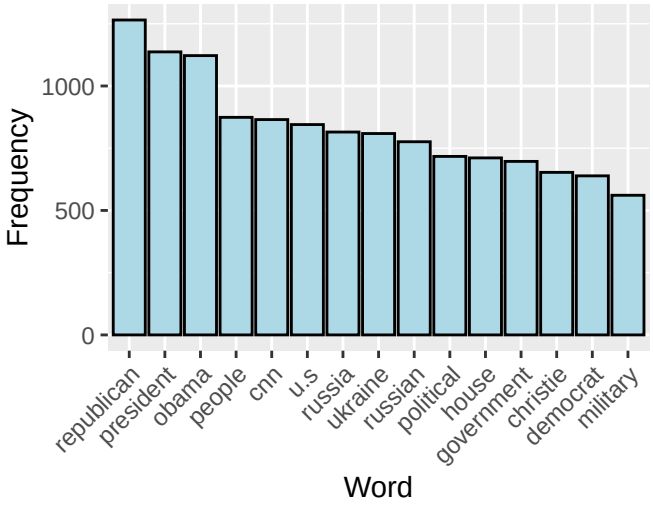


Fig. 2 Top 15 Common Words



The word cloud presents the words that on average appear at least once in each of the 409 documents, as the word frequency cutoff value for the word cloud visualization is set to 409. As seen in both Fig.1 and Fig.2, the most frequent words are CNN Politics are "obama", "republican", and "president", each with more than 1,000 occurrences.

It is worth noting President Obama was in office in 2014. One could make an argument about CNN being Democratic leaning given that "obama" is the most common word. However, determining political bias requires further sentiment analysis to see if CNN casts the Democrats, the Republicans, and President Obama in a more positive or negative light.

On top of that, the coupling occurrence of "russia", "russian", and "ukraine" speaks to the Russian-Ukrainian War that began in February 2014 (Wood et al., 2016).

Sentiment Analysis

Word cloud of sentiment-bearing words

This is a representation of the common sentiment-bearing words across all CNN news data, separated by positive and negative sentiment. There is no statistics presented in this visualization, but the words can be contextualized in the period between January and April 2014. For instance, many Affordable Care Act-related consumer protection are effective starting on Jan 1st, 2014, including the small businesses' access to health benefit plans in a new transparent insurance market and the affordable care through through tax credits (Forum on Medical and Public Health Preparedness for Catastrophic Events, 2014). In fact, words such as "support", "benefits", "reforms", and "affordable" can be associated with the Affordable Care Act, and their presence could mean that the positive sentiments over these three months are mainly contributed by the Affordable Care Act, which President Obama takes credit for.

Filtering

As explained earlier, the research interest lies in the emotion of the documents that are Republican-related, Democrat-related, and Obama-related. For instance, to select the documents that are *about* Republicans, first a list of document number that contains the word "republican" is generated, and then that list is used as a filter to get the words in all the Republican-related document. The process is repeated every time to get a label-related data set.

```
rep_list<-CNN_politics %>%
  filter(word %in% c("republican"))

obm_docno_vec <- as.vector(unique(obm_list$docno))

obm_sentiment<-CNN_politics %>%
  filter(docno %in% obm_docno_vec) %>%
```

Afinn sentiment score

Afinn assigns a sentiment score to every word in the range of -5 to 5 (the most negative to the most positive). The sentiment score for that piece of news hence is the sum of sentiment scores of each word in the piece. If they add up to 0, the sentiment of the piece is assigned neutral; if the sum value is greater than 0, the sentiment is positive; if the sum is smaller than 0, the sentiment is negative. Finally, the distribution of positive, neutral, and negative sentiments in Republican-related news can be plotted.

```
...
  inner_join(get_sentiments("afinn")) %>%
  summarise(sentiment_score = sum(value))%>%
  mutate(sentiment = ifelse(sentiment_score > 0, "positive",
                            ifelse(sentiment_score < 0, "negative", "neutral"))) %>%
    ...
```

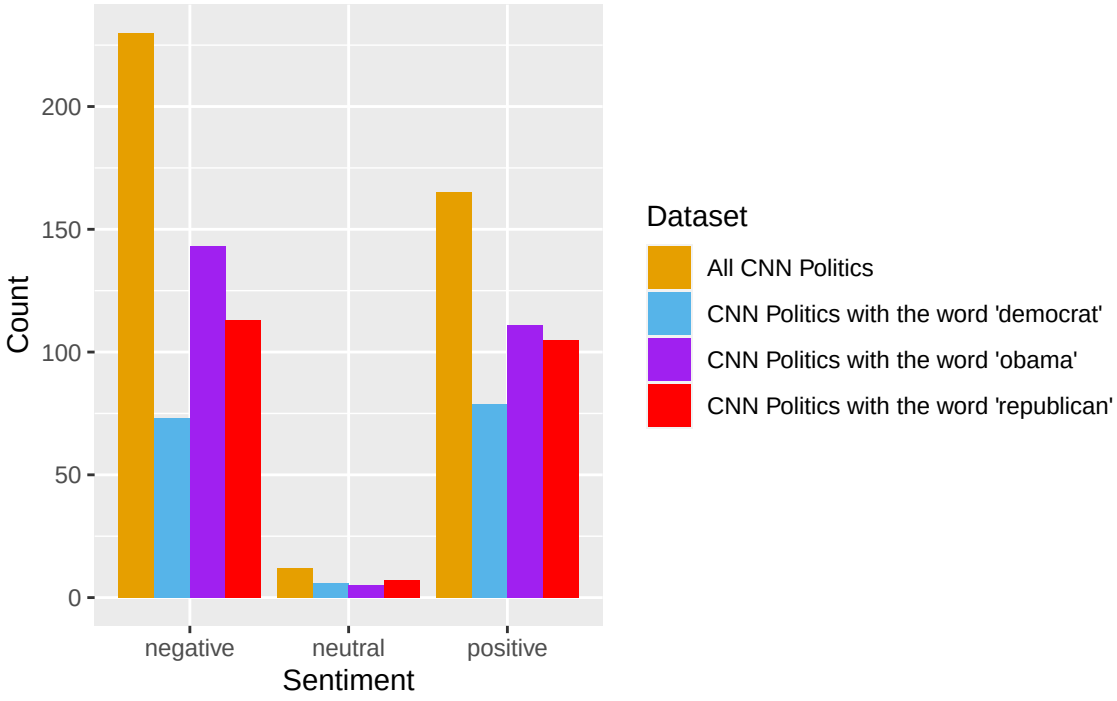
Using Afinn sentiment score for sentiment analysis, CNN overall casts contents negatively. The average sentiment score for all 409 documents is -5.5. Although there is no study linking CNN to negative news, Bellovary et al. (2021) have found the link between the politically-affiliated news organizations and their delivery of negative emotional content. They also show that the level of negativity delivered is positively correlated with readers' engagement. In other words, CNN may have intentionally made their contents disproportionately negative to encourage more user engagement. Certainly, the dominance of negative news is not unique to CNN but is a worldwide phenomenon. Scholars have contributed the constant bad news to people's negative bias and the evolutionary outcome of a more active response to negativity. As audience pay more attention to negative news, they are also shaping the news (Khan, 2019). One could therefore conclude that CNN's negative news flow is not partisan-driven but is simply a marketing strategy. Alternatively, one could say that with an already Democratic audience makeup, CNN is using negative news to carve a

homogeneous Democratic user group, because the viewers' fixation on CNN as the news consumption venue is going to grow with more negative news flow.

CNN's negative portray on Republicans, however, is less likely to be innocent. In Fig. 3, the number of negative Republican-related contents is slightly more that of positive Republican-related contents (113>105), while the number of positive Democrat-related contents slightly exceeds that of negative Democrat-related contents (79>73). CNN thus pictures the Democrat more favorably. CNN has precedents of showing party bias against the Republicans. In the 2008 presidential campaign reports, CNN tends to picture the Republican candidates in a negative light (Pew Research Center's Journalism Project, 2007). Therefore, CNN's historically left political leaning may explain the negative sentiment when the text is about Republicans.

However, this reasoning about political orientation and sentiment does not apply to the sentiment distribution for the Democrat President Obama, because counter-intuitively, Obama is cast in a negative light along with the Republicans, which should not be true if CNN were Democrat-leaning.

Fig. 3 Comparison of Sentiments Using Afinn Sentiment Score




```
## # A tibble: 3 x 5
##   sentiment CNN_all CNN_dem CNN_obama CNN_rep
##   <chr>       <int>   <int>     <int>   <int>
## 1 negative      230      73       143     113
## 2 neutral        12       6         5       7
## 3 positive      165      79       111     105
```

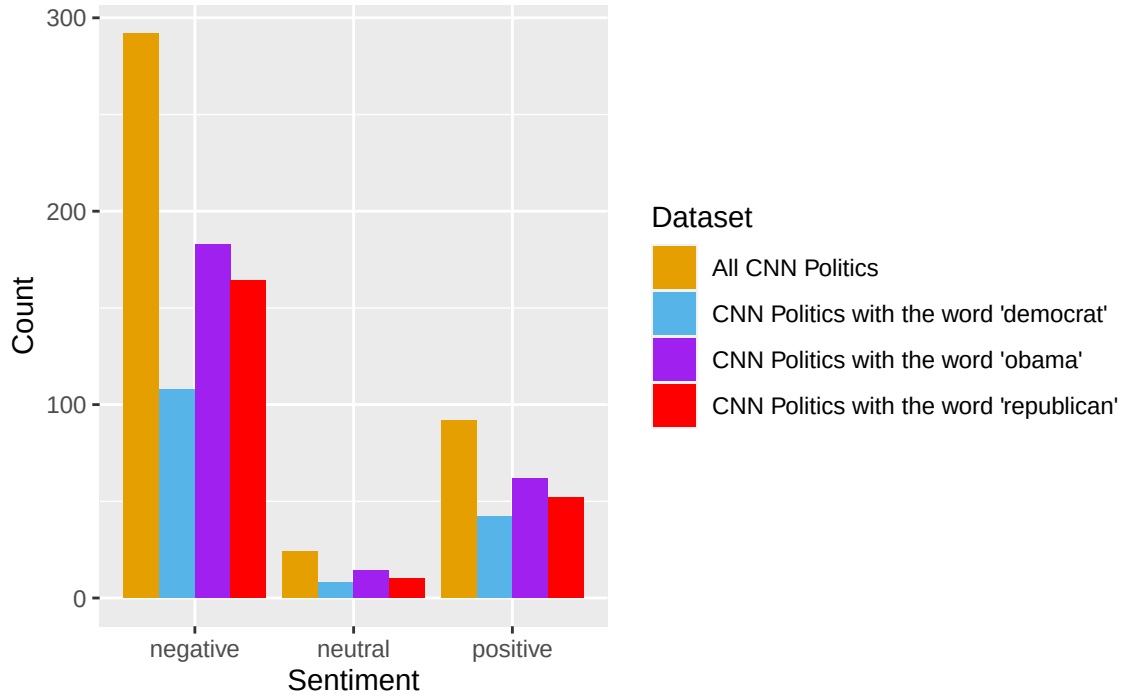
Bing lexicon-based sentiment analysis

Instead of giving a sentiment score, an alternative way is to use the bing vocabulary to assign every word either "positive" or "negative". Then, based on the relative number of positive words to the number of negative words in each document, the document is assign "positive" if the number of positive words is greater than the number of negative words, and vice versa. If the number of negative words and positive words in the document are the same, the document is assigned a "neutral" sentiment.

```
...
inner_join(bing_words) %>%
...
mutate(sentiment = ifelse(positive > negative, "positive",
                          ifelse(positive < negative, "negative", "neutral"))) %>%
  ...
```

Using the alternative method, the same conclusion of the negative overall sentiment can be drawn. Just as in the Afinn sentiment score analysis, the Bing lexicon analysis shows that CNN produces more negative news than positive news and neutral news combined. The only deviation from the Affin lexicon analysis is that Democrat, Republican, and Obama altogether are now cast in a negative light (Fig. 4). There is fewer evidence of pro-Democrat bias in the bing lexicon-based sentiment analysis.

Fig. 4 Comparison of Sentiments Using Bing Sentiment Lexicon




```
## # A tibble: 3 x 5
##   sentiment CNN_all CNN_dem CNN_obama CNN_rep
##   <chr>       <int>   <int>     <int>   <int>
## 1 negative      292     108       183     164
## 2 neutral        24       8        14      10
## 3 positive       92      42        62      52
```

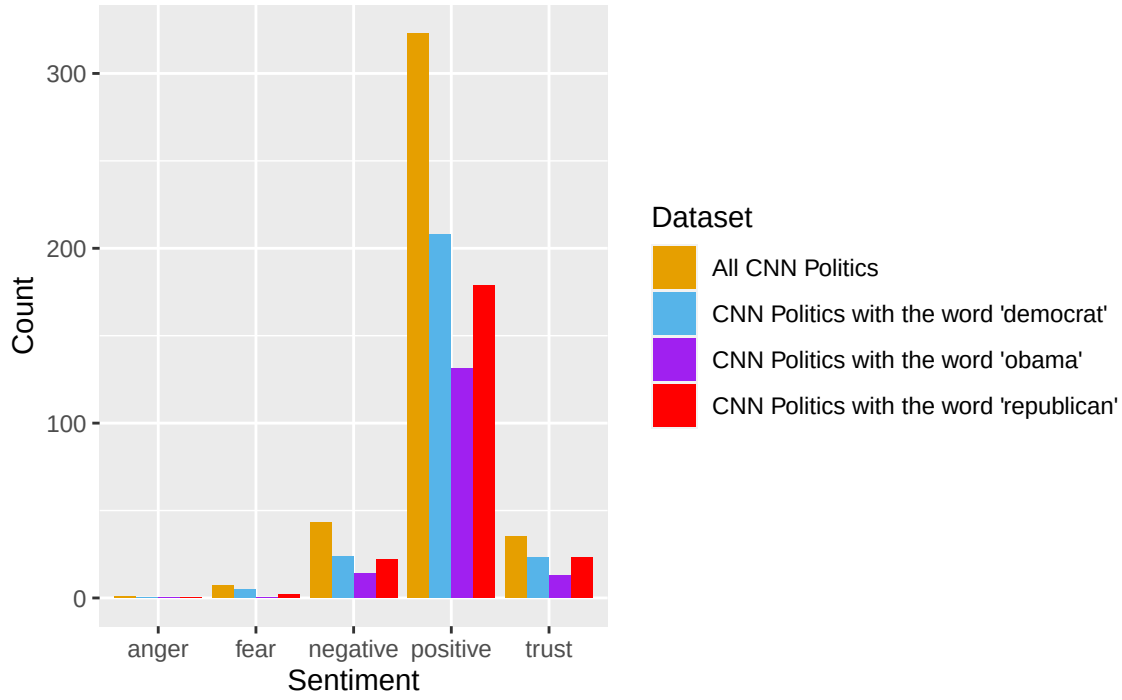
NRC lexicon-baed sentiment analysis

The final sentiment analysis employs the NRC vocabulary that divides sentiment into ten categories, including trust, fear, negative, anger, surprise, positive, disgust, joy, and anticipation. After assigning each ward and counting the number of words in each sentiment, the text's sentiment takes the sentiment with the highest counts of all ten emotions.

```r
obm_sen_nrc<-CNN_politics %>%
  filter(docno %in% obm_docno_vec) %>%
  inner_join(nrc_words, multiple = "all") %>%
  count(docno, sentiment, sort = T) %>%
  pivot_wider(names_from = sentiment, values_from = n, values_fill = 0)
obm_sen_sum_nrc <- obm_sen_nrc %>%
  mutate(sentiment = names(select(obm_sen_nrc, -1))[apply(select(obm_sen_nrc, -1), 1, which.max)]) %>%
```

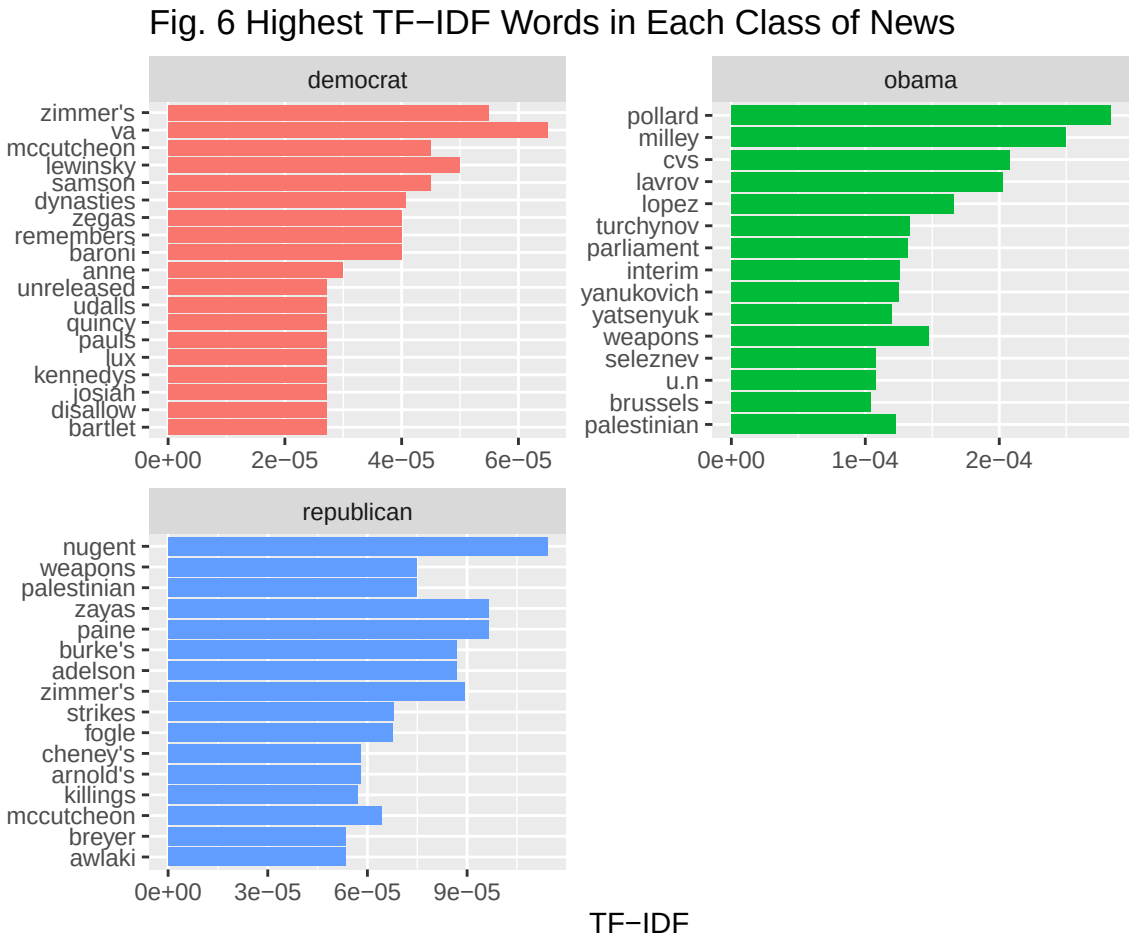
The result from NRC dictionary-based sentiment analysis is a significant departure from the previous two analyses. First, using NRC lexicons, every class of text, is cast positively rather than negatively. This presents contradictory finding from the previous two analyses, concluding that CNN delivers more positive than negative news Between the different categories of data set, there are 208 pieces of positive news on Democrats, 179 positive news on Republicans, and 131 on Obama. This analysis thus shows potential pro-Democrat bias in CNN. Moreover, NRC lexicon-based sentiment reveals fear as the distinct sentiment in Democrat-related contents – there are only 7 pieces of news marked as fear, and 5 of them are Democrat-relevant (Fig. 5).

Fig. 5 Comparison of Sentiments Using NRC Sentiment Lexicon




```
## # A tibble: 5 x 5
##   sentiment CNN_all CNN_dem CNN_obama CNN_rep
##   <chr>       <int>   <int>     <int>   <int>
## 1 anger           1       0         0       0
## 2 fear            7       5         0       2
## 3 negative       43      24        14      22
## 4 positive      323     208       131     179
## 5 trust          35      23        13      23
```

Inverse Document Frequency



Separating the corpus into three classes (Republican, Democratic, and Obama) allows one to identify the frequent and important words that are rare in the corpus but are unique to this class. IDF eliminates the overlapping words that frequently show up across the texts and extracts the most rare and important terms to each class. By performing IDF, the story of Democrat-related, Republican-related, and Obama-related coverage are distinguished from one another.

The differences can be seen in in Fig. 6. For instance, the highest TF-IDF term for Democrat-relevant class is "va", which may refer to Veterans Health Administration (VHA) controversy of 2014. The scandal is about the negligent treatment of veterans at VHA facilities. CNN followed up on the investigation extensively because the whistle blower came to CNN with allegations (Devine & Bronstein, 2014). The highest TF-IDF value term for the Republican-relevant class is "nugent". Obviously, Republican-related news coverage attends to the guitarist's Ted Nugent's racist insults on President Obama (Whitaker, 2014) more than the Democrat-related and Obama-related news. The highest TF-IDF term for the Obama-relevant class is "pollard", which means the Obama-related coverage has more accounts of the Israeli spy Jonathan Pollard's release decision (Myre, 2014). These differences in TF-IDF values reflect the distinct vocabulary and focus used in Republican, Democrat, and Obama-related news.

The different classes of text do share some important terms in Fig 6. For instance, both the Republican and the Obama class have the term "weapons", but "weapons" takes a higher TF-IDF value in the Republican class, which means that the term "weapons" comes up more often in Republican-related news. In the same vein, "Palestinian" is more significant in Republican-related news than in Obama-related news, again speaking to the difference in focus for the three category of news.

The last observation in Fig 6. is term "killings" in Republican-relevant news. The term is also the only term with explicitly negative sentiment. Other than "killings", the rest of the terms are mostly neutral nouns. "killing" as an important term in Republican-relevant news may indicate a party bias in CNN news that paints republican-related contents less favorably.

LDA Topic Model

The estimated two topics are associated with the words in Fig. 7. In Fig. 8, the positive and negative log ratio represents affiliation with topic 1 and topic 2 respectively. Crimea, russian, russia, ukraine, united, u.s, military, government, obama, president belong to topic 1; people, political, republican, democrat, christie belong to topic 2. It is reasonable to infer that topic 1 is about the 2014 Ukraine crisis, and topic 2 is about election and U.S politics. If party bias were present, one would expect to see topics divide along partisan lines. However, if there are more topics, word divergence along partisan lines is possible.

Fig. 7 Topics of CNN Politics News

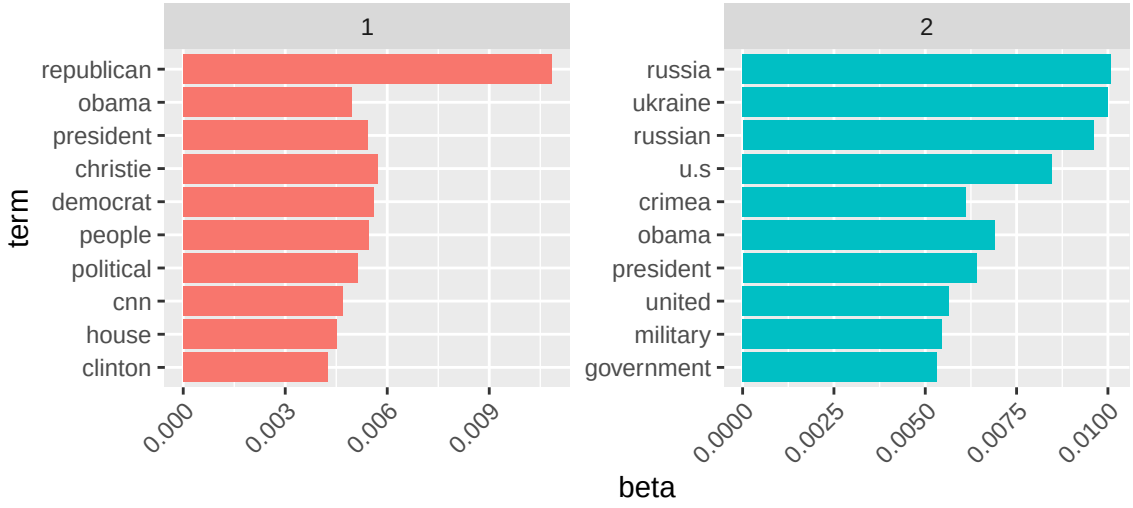
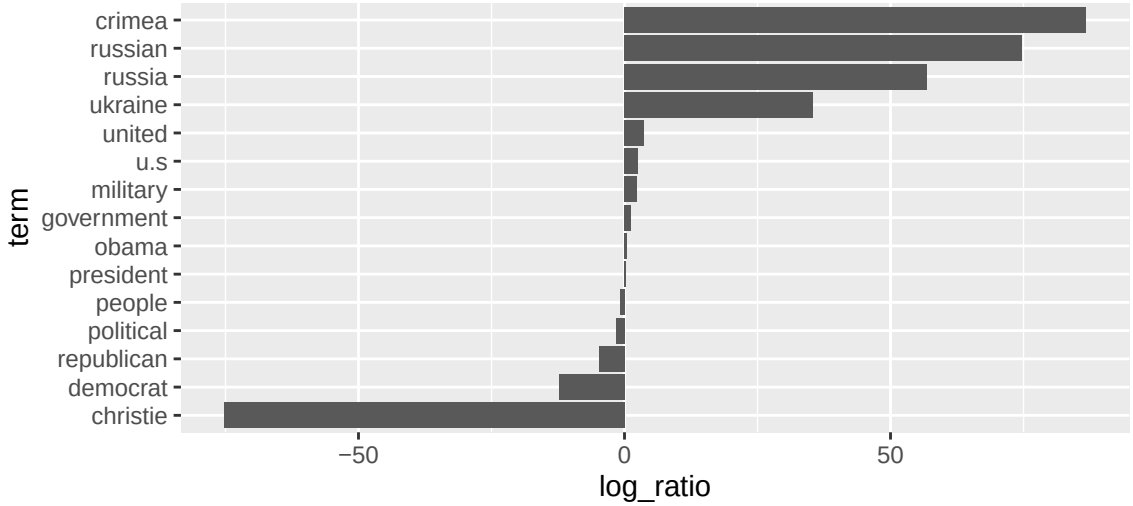


Fig 8. Words with the Greatest Difference Between Topic 1 and



Limitations

The major limitation is that Democrat-, Republican-, and Obama-related news are not exclusive of one another. One text could have the word "democrat", "republican", and "obama" at the same time and thus be labelled as such. The overlap does not affect the determination of sentiments, but it does introduce a lot of repetition in the counts. For example, document number 1 with both "democrat" a "republican" in the text could contribute to counts of positive democrat-related news and positive republican-related news. If the sample can divided into mutually exclusive categorical groups, the results may have greater statistical significance. In this study, the observed difference between positive and negative counts is usually small.

On top of that, the text-labeling method used in this study does not generate the equal size of democrat-, republican-, and obama-related pieces of news. The sample size of democrat-related news is small (n=158) compared to republican- (n=226) and obama-related (n=260), not to mention a large part of the data intersect with the other two classes.

Finally, I only use the word "democrat" to filter the Democrat-related news pieces, because "democratic" could either be about about the Democrat Party or an adjective of Democracy. In the cases of former scenario, this study undercounts the number of Democrat-related news.

Conclusion

The sentiment analysis does not give consistent results on CNN's Democratic leaning. Only with the NRC sentiment lexicon is CNN displaying pro-democrat tendencies. However, it does not negate the hypothesis that CNN's left-wing political orientation is channeled thorough information presentation, given that the sample size of Democrat-related news is smaller than the Republican- and Obama-related news. LDA model reveals the potential of topic modeling to identify finer and smaller topics that may be about partisan leaning. IDF analysis shows that these three classes of news do tell different stories, but that is not equivalent to party bias. The reasons that different subjects are featured across these classes of news are multi-fold, and further studies are needed to find more support on politically-mediated information priming in CNN.

Bibliography

Brenan, M. (2022, October 18). *Americans' Trust In Media Remains Near Record Low.* Gallup.com. https://news.gallup.com/poll/403166/americans-trust-media-remains-near-record-low.aspx

Bellovary, A.K., Young, N.A. & Goldenberg, A. Left- and Right-Leaning News Organizations Use Negative Emotional Content and Elicit User Engagement Similarly. *Affec Sci* **2**, 391–396 (2021). https://doi.org/10.1007/s42761-021-00046-w

Devine, C., & Bronstein, S. (2014, September 18). *VA Inspector General admits wait times contributed to vets' deaths*, CNN politics. CNN. https://www.cnn.com/2014/09/17/politics/va-whistleblowers-congressional-hearing/index.html

Forum on Medical and Public Health Preparedness for Catastrophic Events. (2014). *The Impacts of the Affordable Care Act on Preparedness Resources and Programs Workshop Summary.* The National Academies Press.

Grieco, E. (2020, August 1). *Americans' main sources for political news vary by party and age.* Pew Research Center. https://www.pewresearch.org/short-reads/2020/04/01/americans-main-sources-for-political-news-vary-by-party-and-age/

Khan, A. (2019, September 5). *Why does so much news seem negative? human attention may be to blame.* Los Angeles Times. https://www.latimes.com/science/story/2019-09-05/why-people-respond-to-negative-news

Myre, G. (2014, April 1). *The arguments for and against releasing Jonathan Pollard.* NPR. https://www.npr.org/sections/parallels/2014/04/01/297675675/the-arguments-for-and-against-releasing-jonathan-pollard

Pew Research Center (2014, June 12). *Political Polarization in the American Public.* Pew Research Center. https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/

Pew Research Center: Journalism & Media staff. (2007, October 29). *The invisible primary - invisible no longer.* Pew Research Center's Journalism Project. https://www.pewresearch.org/journalism/2007/10/29/the-invisible-primaryinvisible-no-longer/

Robinson, J. S. and D. (n.d.). *3 Analyzing Word and Document Frequency: Tf-IDF: Text mining with R.* | Text Mining with R. https://www.tidytextmining.com/tfidf.html

Whitaker, M. (2014, January 22). *Ted Nugent calls obama "Subhuman Mongrel."* MSNBC. https://www.msnbc.com/politicsnation/ted-nugent-calls-obama-subhuman-mongrel-msna252356

WOOD, E. A., POMERANZ, W. E., MERRY, E. W., & TRUDOLYUBOV, M. (2016). *Roots of Russia's War in Ukraine.* Columbia University Press. https://doi.org/10.7312/wood70453

Wankhade, M., Rao, A.C.S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif Intell Rev* 55, 5731–5780 (2022). https://doi.org/10.1007/s10462-022-10144-1