Measuring Bias in COVID-19 News Articles

Andrew Pagtakhan, Kwan Bo Shim, Cinthia Jazmin Trejo Medina

May 10th, 2021

Contents

1 Preface

The full code and dataset can be found here: https://github.com/ajpag/US-News-NLP

All analysis was completed in R.

2 Motivation

2.1 Research Question

Are there underlying patterns in news articles related to COVID-19 across major news sources that suggest bias, and are these patterns predictive of which news source is it likely from?

2.2 Background

The means in which informaton is communicated with regards to the COVID-19 pandemic has had major influence on how we read and learn about the virus through a multitude of media outlets. Some of these major sources include television, YouTube, social media forums, and major news companies. Major news companies in particular carry large influence based on the audiences it can reach. For example[1], Fox News Channel averaged 2.5 million primetime viewers (8pm - 11pm) in February 2021, and CNN averaged 1.7 million during the same time period.

According to King G.[2] & et. al., *"...the exposure to news media causes Americans to take public stands on specific issues, join national policy conversation, and express themselves publicly"*. Furthermore, Holman E.[3] & et. al, suggest a correlation between raising level of stress and prolonged media exposure to "community-based traumas (e.g., mass shootings, natural disasters)". Recently, Holman[4] has suggested that COVID-19 is a particular case to study since multiple stressors have arose at the same time. To mention a few stressors: financial crisis, elections, and health crisis, among others. It can be said that news can influence the decisions, general views and mental health of Americans.

The large influence that major news companies have on how information is commmunicated to its audiences, and the impact news have on its readers makes it vital to quantify how different these sources are in relation to COVID-19 news. By measuring potential bias in relation to each source, this analysis examines how different major news sources are when reporting on COVID-19. It also explores underlying patterns such as subtopics within COVID-19 that are published more in news sources over others. And if these patterns are predictive of which news source the article came from.

2.3 Applications

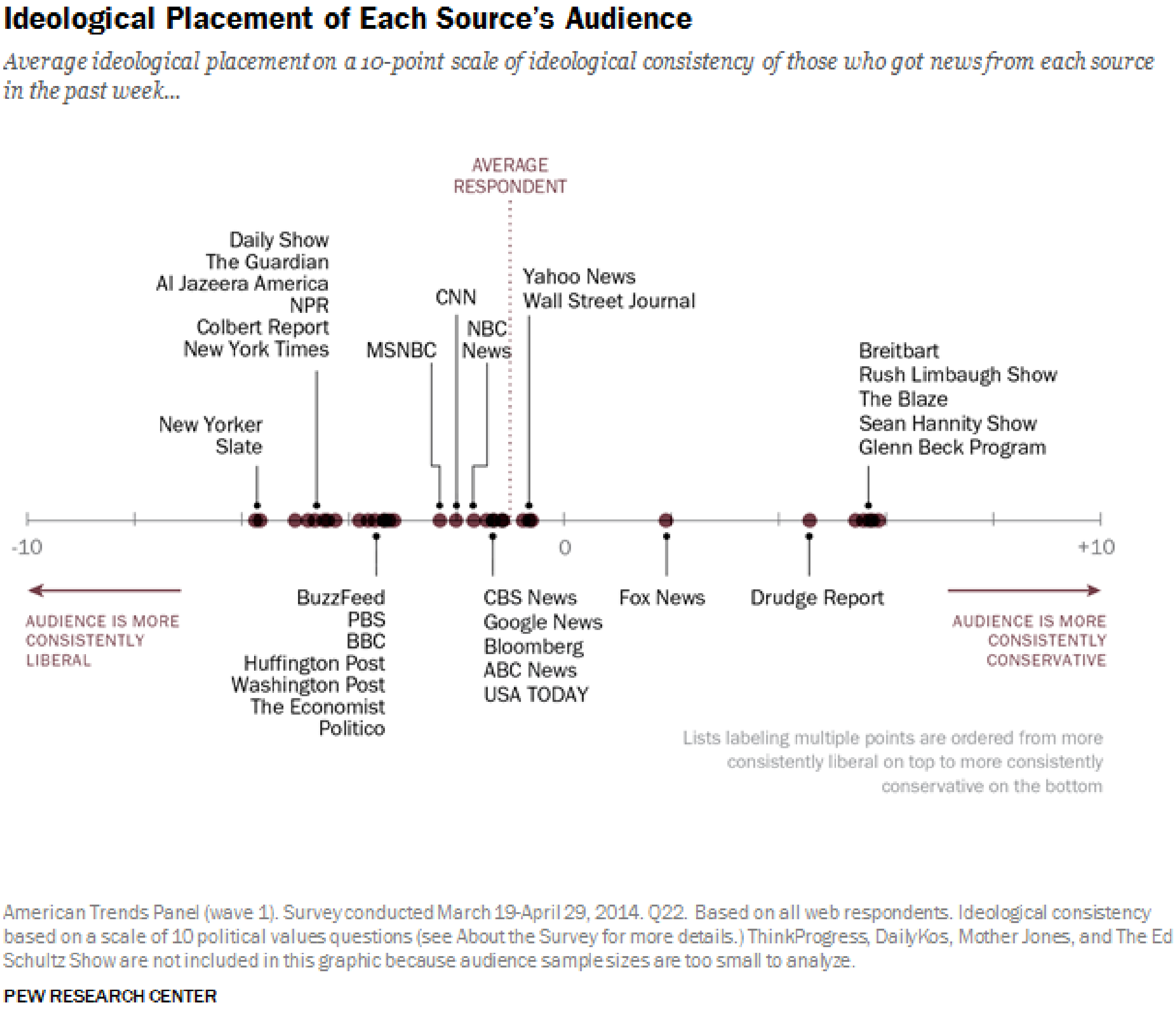
By quantifying underlying differences on COVID-19 reporting and examining the predictive power of these patterns to identify the news source, this study can be useful for a number of cases. For example, understanding biases in article text can help the reader understand inherent ideological leanings towards certain news sources. This can help to equip them with greater understanding and critical examination of news consumption. In the same way, it can help readers to be more selective when choosing news sources and reduce their stress impact.

From a policy perspective, greater impact and studies could be done to influence greater transparency[6] in reporting across the news companies. This can assist in making informed decisions on which news sources to read or be aware of the bias different sources might have.

3 Data Description

3.1 Data Sourcing

In order to choose major news sources to analyze, looking at diversity of news sources is important. Therefore, we used the figure below created by the Pew Research Center. This figure shows where on the US political spectrum various news companies fall. This criteria was used to get a mix of sources across the conservative and liberal spectrum, while meeting limitations of computing resources. The following news sources were used to procure a dataset.



Ideological Placement of Each Source's Audience

Average ideological placement on a 10-point scale of ideological consistency of those who got news from each source in the past week...

Lists labeling multiple points are ordered from more consistently liberal on top to more consistently conservative on the bottom

American Trends Panel (wave 1). Survey conducted March 19-April 29, 2014. Q22. Based on all web respondents. Ideological consistency based on a scale of 10 political values questions (see About the Survey for more details.) ThinkProgress, DailyKos, Mother Jones, and The Ed Schultz Show are not included in this graphic because audience sample sizes are too small to analyze.

PEW RESEARCH CENTER

Figure 1: Source: Pew Research Center

Source: https://www.journalism.org/2014/10/21/political-polarization-media-habits/pj_14-10-21_mediapolarization-08/

- BBC
- CNN
- The Wall Street Journal
- Reuters

Leveraging the **GNews API**, news article data related to COVID-19 were pulled for the selected major news sources (filtered to US articles). The time frame used for this analysis was between January 1st 2020 - April 9th 2021.

A total of 5,360 articles were called from the API (20 articles per week and news source, for 67 weeks). The API call returned 3,454 records for article-related data (due to REST api query limitations). Below are some of the key fields from the API call, with an example:

- `article url`: https://www.cnn.com/2021/02/08/health/covid-19-antigen-tests-states-cnn-analysis/index.html
- `article description`: "Covid-19 antigen tests not counted among cases in some states, CNN analysis shows - CNN"
- `date and time of article publication`: 2021-02-08T08:00:00Z
- `article source name`: CNN

Note: Due to legal restrictions, Fox News data was not pulled

3.2 Data Cleaning: Web scraping

After gathering the article URLs, the full news text for each article was pulled. Each data source carried unique idiosyncrasies in its html structure. The next sections highlight the unique aspects of web scraping each data source.

3.2.1 BBC

BBC news articles used different HTML structures based on each section. Hence, there was a need to use the keyword aguments in the "GNews API" to extract articles only in BBC's "News" tab. This allowed all body text to be scraped. However, it also scraped advertisements and extraneous information (such as disclaimers). As a result, there was a need to remove irrelevant lines and words to prevent these features from affecting the analysis.

3.2.2 CNN

There were two distinct HTML structures. One for the first sentence, and another for the remainder of the article. Additional cleaning needed to be done to remove the "(CNN)" and "(CNN Business)" text at the beginning of each article, as well as additional escape characters scattered throughout the article.

3.2.3 Reuters

This was relatively streamlined compared to the other news sources. One unique aspect of Reuters was that a number of their articles were not text articles in the traditional sense, but slideshows, such as : https://www.reuters.com/news/picture/coronavirus-outbreak-spreads-in-china-idUSRTS2ZART

As such, these articles were not scraped due to significantly different html structures (n = 18).

3.2.4 The Wall Street Journal

The original plan was to also extract articles from a conservative newspaper, such as Fox News. However, Fox news updated their Terms of Use which strictly forbids scraping: *"... any automated means, including"robots," "spiders," or "offline readers"... ".* Thus, The Wall Street Journal was scraped instead. One downside of using The Wall Street Journal was that it is a subscriber-based website and only allows readers without a subscription to extract a fraction of each article. This could potentially explain the small amount of average word count per article. Similarly to BBC, scraped body text contained irrelevant information and it was removed with further cleaning processes.

4 Exploratory Data Analysis

To better understand the data, sentiment analysis utilizing various lexicons were explored. Of the lexicons available in `tidytext::get_sentiments()` the most informative one for this analysis was Afinn lexicon.

This provided a range of sentiments from negative to positive on a scale from -5 (most negative) to 5 (most positive). It also allows for more streamlined sentiment analysis, such as finding average sentiment. Key visualizations of the data are illustrated below.

Excluding The Wall Street Journal, BBC had the lowest average words per article, whereas CNN had the most words per article.

There were a couple of CNN articles that asked for reader feedback, hence low counts. Example: https://www.cnn.com/2020/05/29/us/coronavirus-infected-in-hospital-callout-invs/index.html (words = 33)

Reuters has articles with short reports, such as: https://www.reuters.com/article/us-china-health-azerbaijan/azerbaijan-reports-first-case-of-coronavirus-ifax-idUKKCN20M1IR

The two longest articles were written by CNN and The Wall Street Journal:

- https://www.cnn.com/2020/11/30/asia/wuhan-china-covid-intl/index.html
- https://www.wsj.com/articles/what-we-know-about-the-coronavirus-11579716128

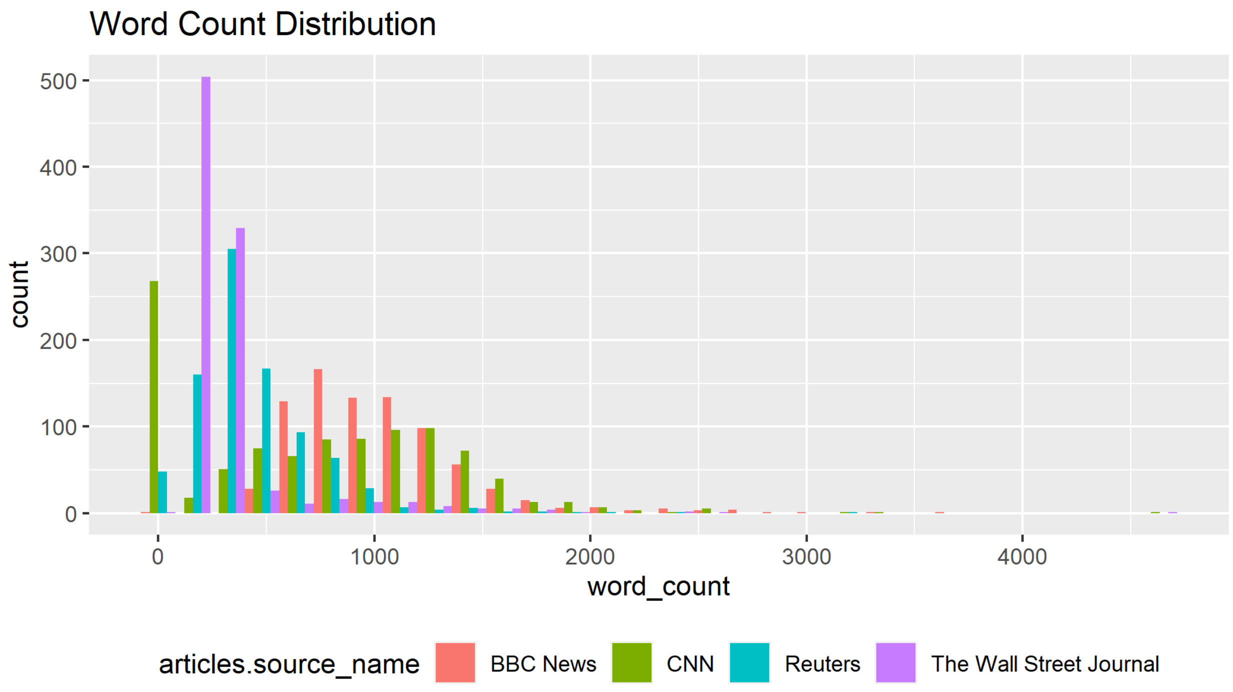


Figure 2: Words Per Article

BBC News had a larger variety of top words compared to the other news sources.

All news sources had average weekly sentiment mainly hovering between -1 to 0, using the Afinn lexicon. This means that average sentiment was slightly negative.

The Wall Street Journal had the largest range in average weekly sentiment. It is plausible there is some correlation between this sentiment and the stock market, and would require further analysis outside of this study.

Afinn lexicon: https://www.tidytextmining.com/sentiment.html

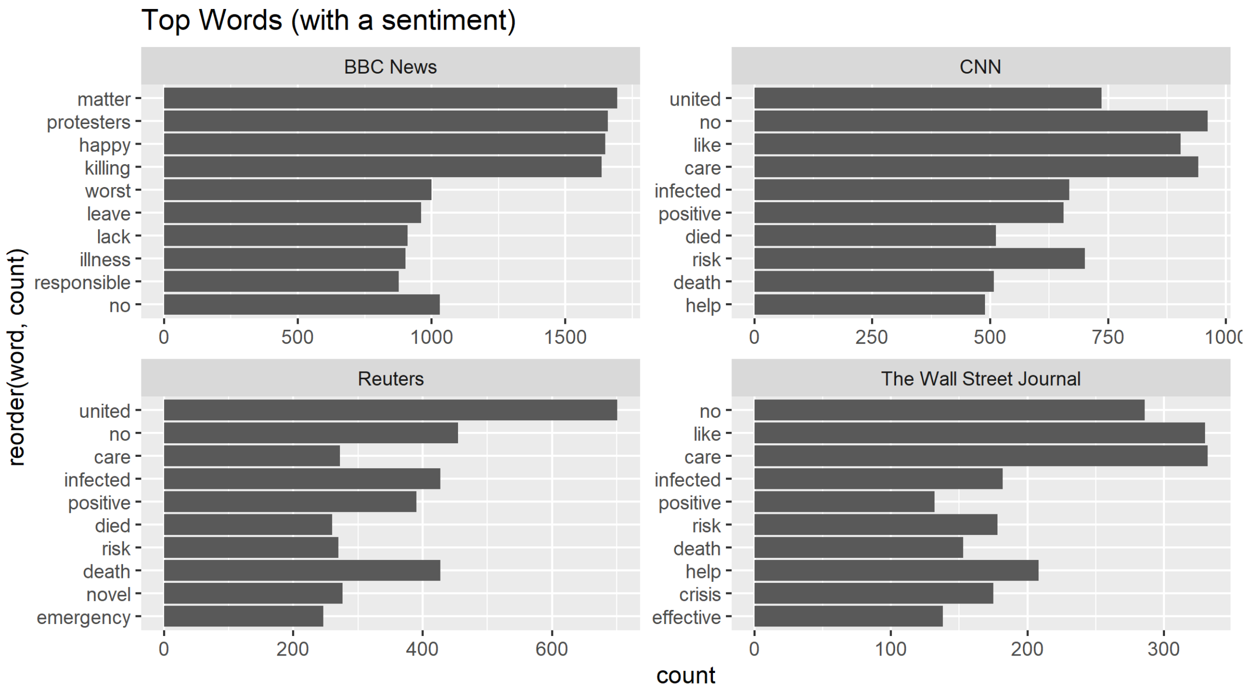


Figure 3: Top Words with a sentiment

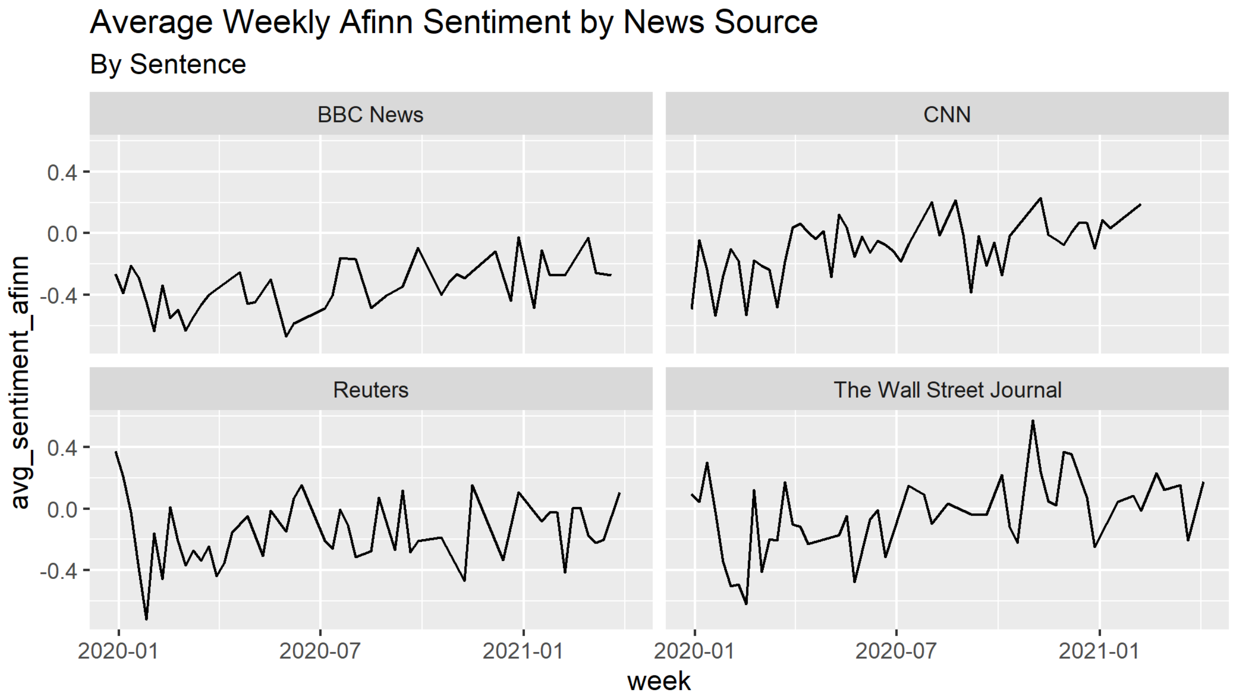


Figure 4: Average Sentiment by Week

5 Methodology

In order to predict which news source the article was written by, classification modeling was performed. The analysis was done with three sets of features in order to evaluate the bias and different patterns accross all the news sources. The purpose of each set of features is to explain a different dimension in which the news sources might be differentiated. The three dimensions modeled were: 1.) sentiment based on key words,2.) vocabulary previously studied in other papers, and 3.) topics determined through topic modeling.

The analysis was split in two parts. First, the bias across sources was measured with a chi-squared test in each of the explained dimensions. Second, the patterns were analyzed through various classification algorithms.

5.1 Feature Engineering

In order to create features to build classification models for predicting a news source based on text and measure bias across sources, the below features were generated.

5.1.1 Average Sentiment and Word counts

One feature was generated for each article:

- average sentiment by word
- average sentiment by sentence
- word count
- word count with a sentiment

5.1.2 Keyword features

As a baseline, keyword features based on the researchers' assumptions on the following topics was used to generate average Afinn sentiment by word for each article. One feature was generated for each term below (noted in quotes).

- Politics: "Trump", "Biden"
- Business: "stock market", "financial"
- Pandemic: "death", "pandemic", "disease", "illness"

5.1.3 Topic features from prior research

In order to evaluate existing literature on politicization and polarization, features based on previous published papers were generated.

The first paper used to generate features (Hart, Chinn and Soroka[8]) introduced a dictionary, which assembled major words with the most relevant vocabularies such as Covid-19, Scientist, Republican and Democrat, to explore politicization and polarization in pandemic news in the U.S. newspapers from March 2020 to May 2020.

- COVID 19: "corona", "coronavirus","covid"
- Scientist: 'scientist', 'research', 'professor', 'health official', 'doctor','dr', 'health commission','expert', 'health leader', 'health service','health authorit', 'world health organization', 'centers for disease control and prevention', 'cdc', 'national institutes of health', 'health and human services', 'mayo clinic', 'johns hopkins' , 'fauci', 'birx', 'tedros'
- Republican: "republican", "gdp", 'conservative', "trump", "pence", "mcconnell", "white house", "administration"
- Democrat: "democrat", "liberal", "progressive", "pelosi" , "schumer", "biden", "obama", "newsom" , "whitmer", "cuomo" , "biden," , "sanders"

The second paper by Green et al[9] demonstrated an absolute difference in the proportion of words used by political party using Tweet data from January to March 2020.

- Republican Words: “coronavirus”, “china”, “businesses”, “realdonaldtrump”, “relief”, “inittogether”, “small”, “together”, “cares”, “great”
- Democrat Words: “health”, “need”, “crisis”, “public”, “workers”, “trump”, “must”, “pandemic”, “care”, “leave”, “paid”, “familiesfirst”, “people”, “sick”, “emergency”

The final research by Schaeffer[10] also shows the difference in views on various topics based on their political stands from January and Feb 2021.

Politically divisive national policies that needs to be addressed this year[10]

- div_words1: “social distance”,“gathering”,“avoid”,“large groups”
- div_words2: “limit”,“carry-out”,“restaurant”,“restaurants”
- div_words3: “closing”,“close”,“k-12”,“school”
- div_words4: “race”,“racial”,“racism”,“blm”
- div_words5: “climate”,“climate change”,“global warming”,“global climate change”

The new features computed the frequency of these list of words used in each article. Since these features are not extracted nor inferred from the collected article data but from a broader analysis, it is plausible that these features can enhance our classification model performance.

5.1.4 Topic Modeling: Latent Dirichlet Allocation

To assess commonalities in topics and article text across articles, Latent Dirichlet Allocation was applied. After tuning, it was decided that 7 topics was the optimal number, based on interpretabilty and minimal overlap of topic words. The VEM method was used: https://www.tidytextmining.com/topicmodeling.html

Topics:

Table 1: Topics from LDA

topic_number	topic
1	Politics
2	Reported deaths and cases
3	China / Wuhan
4	Outbreaks and infections by country
5	Patients and Symptoms
6	Vaccines and Research
7	Business and Economy

5.2 Measuring Bias Across News Sources

To measure bias across news sources, chi-square tests were performed using two of the previous dimensions (topic probability and average sentiment), as well as the word count, to measure the significant differences. The chi-square tests for each of the three metrics suggest that there is indeed bias in article text across the news sources.

Table 2: Chi Square Test Results

chi_sq_test_metric	p_value
topic_probabilities	0.0000000

chi_sq_test_metric	p_value
afinn_sentiment_sentence	0.0024126
word_count	0.0000000

5.2.1 Topic Probabilities

Following intuition, The Wall Street Journal has the highest average topic probability for Business and Economy articles. It is interesting that Reuters has the highest probability of China / Wuhan related articles. CNN has the lowest probability of articles pertaining to vaccines and research.

The chi-square test ($p < 1\%$) indicates there are correlations across each news source in relation to average topic probabilities.

Note: Probabilities were multiplied by a factor of 100 prior to running the test

5.2.2 Average Sentiment

The chi-square test ($p < 1\%$) indicates there are correlations across each news source in relation to average Afinn sentiment.

Note: Sentiment scores were multiplied by a factor of 100 and converted to positive integers prior to running the test

5.2.3 Word Count

The chi-square test ($p < 1\%$) indicates there are correlations across each news source in relation to average word count.

Note: Average word counts were converted to positive integers prior to running the test

5.3 Classification Algorithms

In order to analyze the difference in patterns across news sources, 5 algorithms were used. For rach algorithm, 4 models were run using different sets of features:

- **Topic Modeling features**: per section 5.1.4
- **Keyword features**: per section 5.1.2
- **Topic keyword features**: per section 5.1.3
- All features

6 Conclusion

6.1 Results

Assessing bias using the chi-square test 7 suggests that there is strong evidence that the article text varies significantly across the news sources.

Of all the models that were run, the random forest using all features performed the best (90% accuracy, 98% AUC). The models that performed least favorably were Naive Bayes and Support Vector Machine. The random forest model performed the best because it is a non-parametric model without strong assumptions. Since the news sources have inherent topics (i.e. Business in The Wall Street Journal) they are more likely to write about, but have some degree of overlap, the random forest can model these patterns better vs. other models such as logistic regression.

Naive Bayes and Support Vector Machine did not perform well for a number of reasons:

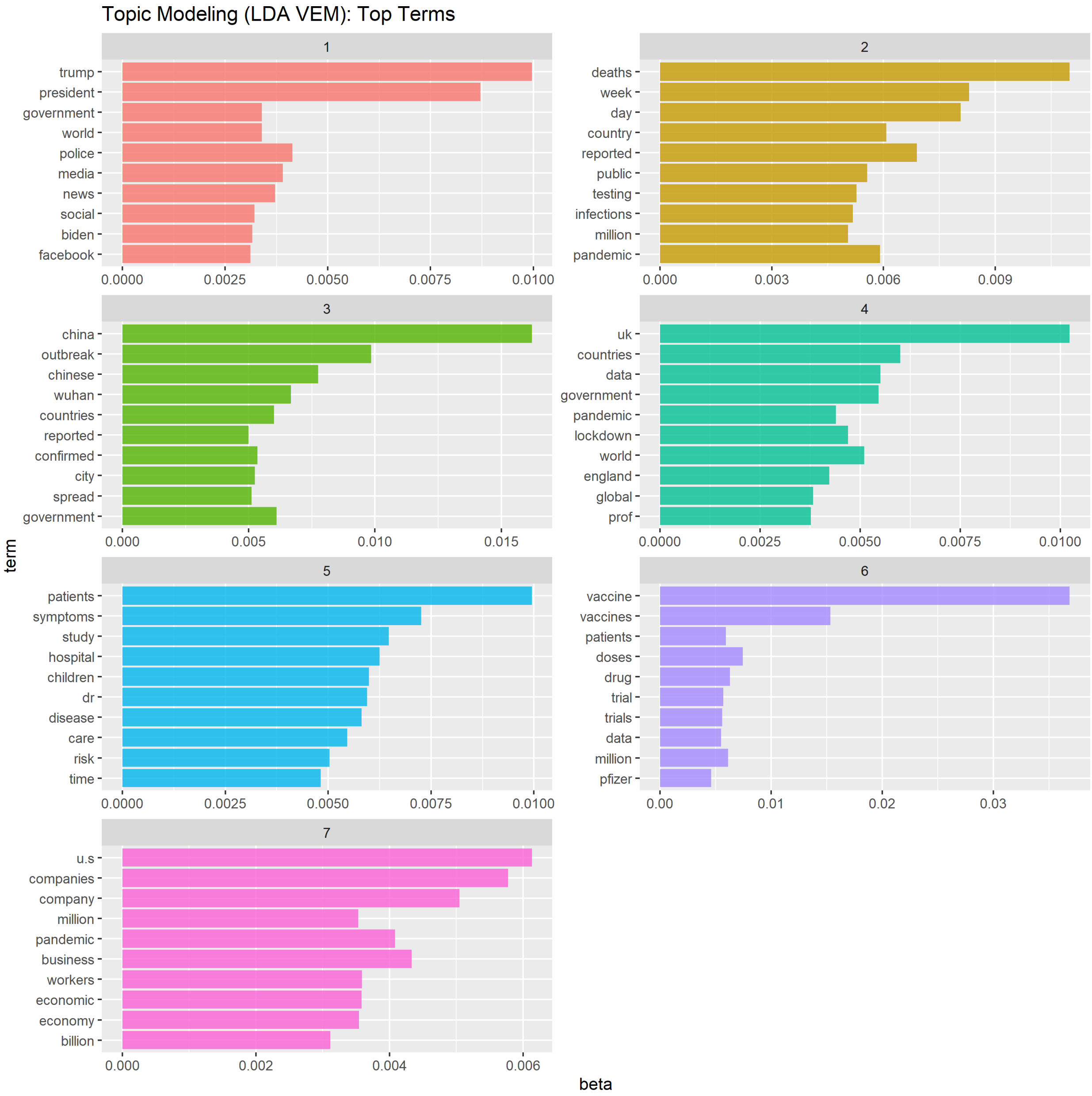


Figure 5: Topic Modeling: Top Words

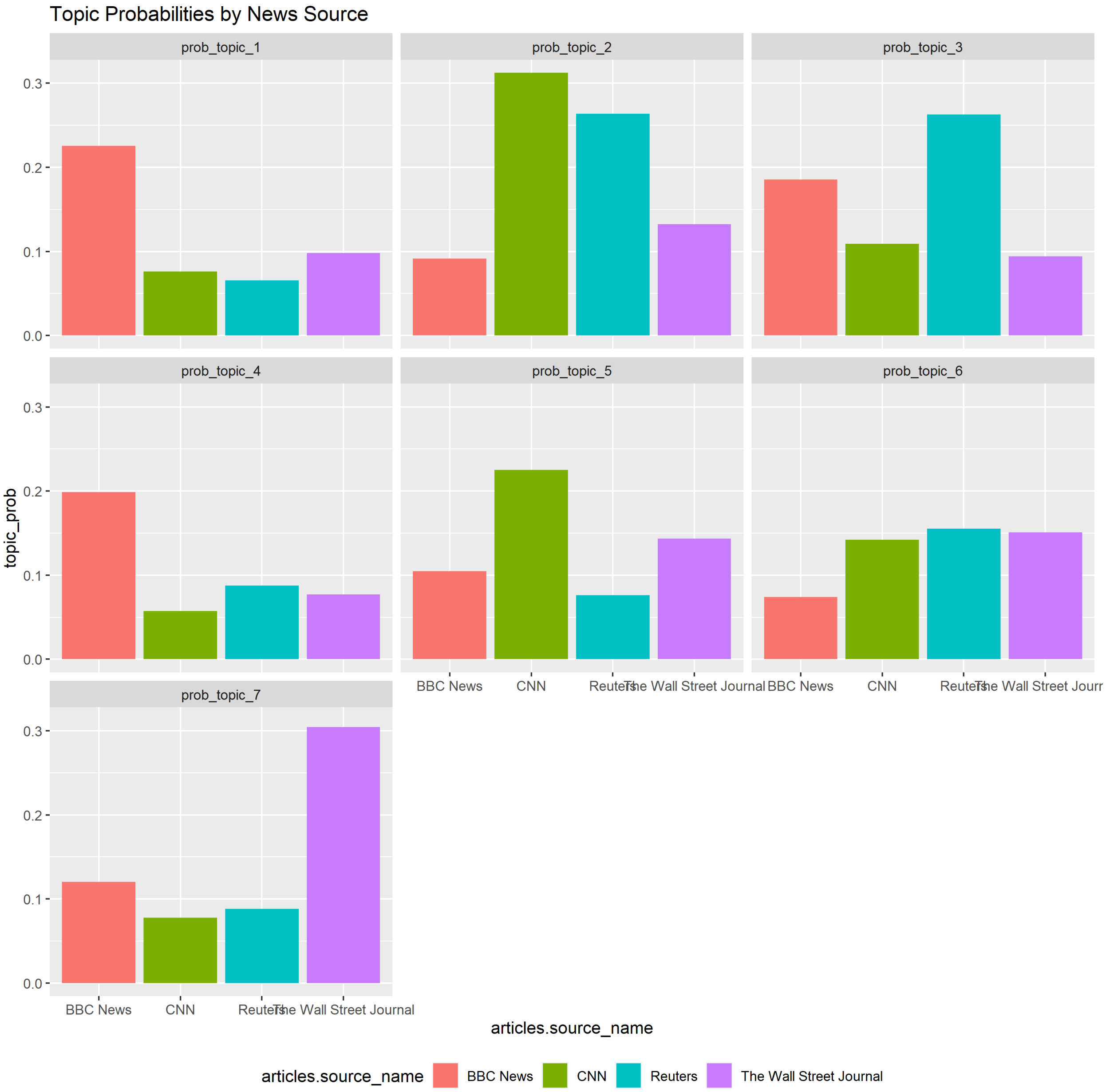


Figure 6: Topic Modeling Probabilities

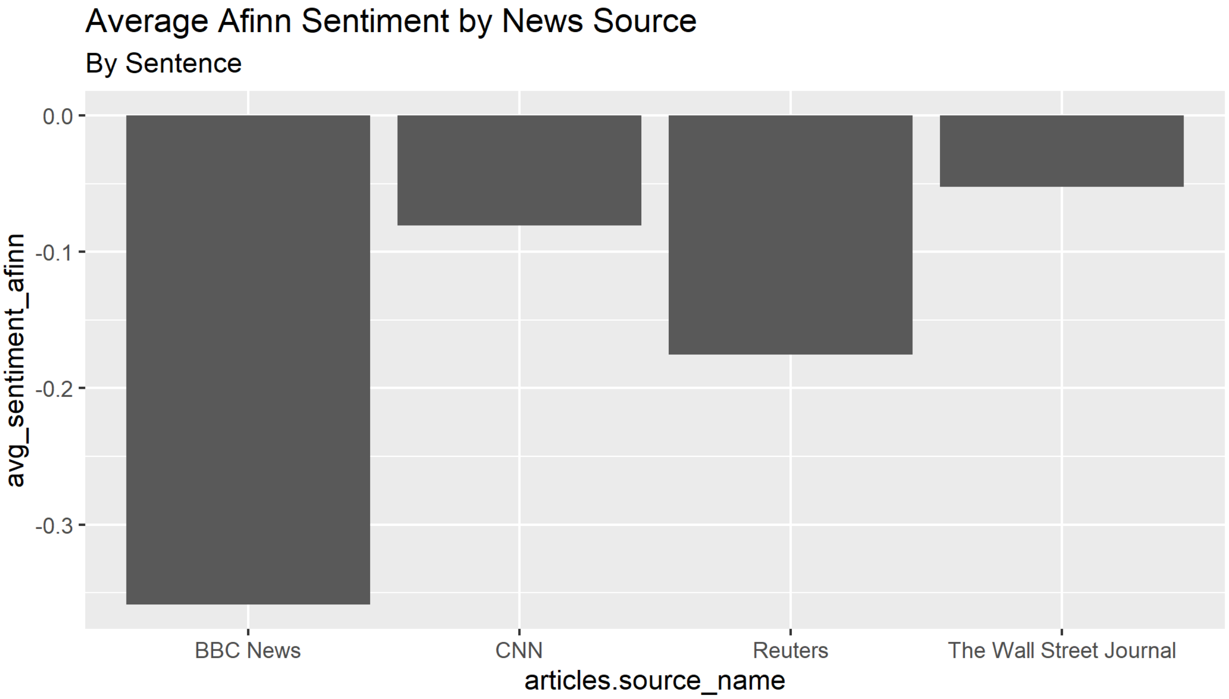


Figure 7: Avg sentiment by sentence across articles

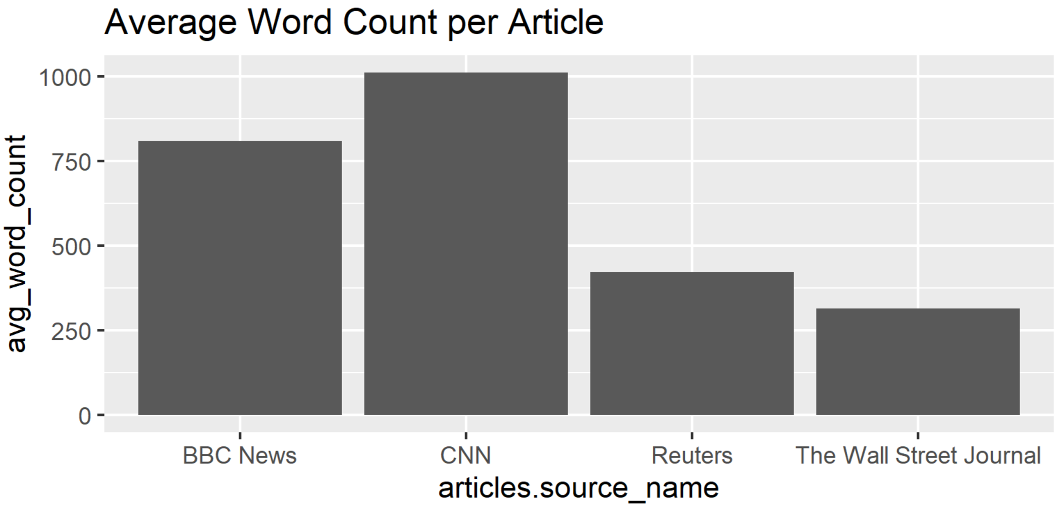


Figure 8: Avg word count across articles

- Strong assumptions: Naive Bayes assumes independence of features, which is not the case in our data.
- Binary vs. Multi-class predictions: The two models are better suited for binary classifications, as opposed to multi-class problems. Even when it is possible to fit the models as binary predictions, it is time consuming and there are other factors that need to be taken into account, such as rebalancing of the training samples.
- Boundary splitting: Naive Bayes and SVM perform better when the data structure naturally splits the data to accurately classify the news source. Since the data does not have clear patterns that can split the data cleanly, it does not perform as well.

The feature importance for the best model, Random Forest (all features), are shown in the below figure.

The most predictive feature was probability topic 7 (Business and Economy), which best split Wall Street Journal from the other news sources.

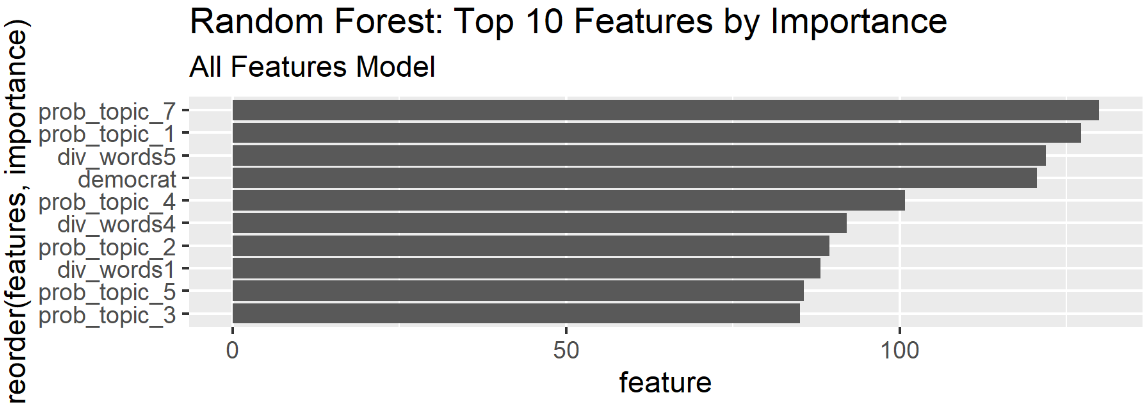


Figure 9: Avg word count across articles

6.2 Implications

It seems that, as other papers have suggested, there is indeed bias across news sources. The bias as shown through the modeling generates patterns that can accurately predict the news sources of which an article is coming from. This information could potentially be used towards making bias more transparent across news sources for the reader. It is also a foundation towards modeling sentiment across different topics to evaluate the impact in healh during the pandemic.

6.3 Limitations

- Selection bias in news articles analyzed: Due to legal restrictions, more conservative news sources, such as Fox News, were not scraped. Also due to legal restrictions, the articles of The Wall Street Journal were not able to be fully scraped.
- Context limitations in sentiment: The sentiment method used, Afinn, only parsed words individually, and not into context of the entire article. So, the sentence sentiment was calculated using the average sentiment of words.
- Parsing limitations: There are many edge cases in which breaking down articles by sentences did not parse successfully. For example, `tidytext::unnest_functions()` incorrectly parsed the following sentence into two since `Ms.` has a period:

Original sentence: `"The manager's decision to send Ms. Coleman home for wearing the headscarf was due to a lack of training," Warren said.`

Parsed sentence(s)

- `"The manager's decision to send Ms."`
- `"Coleman home for wearing the headscarf was due to a lack of training," Warren said."`
- Calculating sentiment scores from news articles to predict the publisher may not be an effective mechanism. The research question hypothesized that the average sentiment of articles from each newspaper on a single phenomenon is not equal. It is true that each newspaper has its own perspectives in general. However, it is difficult to detect and to measure these patterns based on each article's sentiment; news articles are mostly written with an objective and logical tone. It would be more comprehensive if we focus on the headline of the articles[11] or data from social network service such as Facebook or Twitter as they are more likely to include emotional words. In this paper, the researchers used machine learning models that we used to predict three major sentiments based on Twitter data with a higher accuracy.[12]
- Another issue is related to newspaper's several differences in their nature. Firstly, Wall Street Journal is a newspaper specialized in business and economic issues, unlike BBC or CNN newspapers, which cover general topics. Moreover, Wall Street Journal and CNN are based on the United States whereas Reuters and BBC News are based in the United Kingdom. Finally, BBC and CNN are regarded as broadcast news sites which implies that some articles are from press services such as Reuters and AP News[13]. Inequality in different data sources may explain the topic modeling features carry the highest significance in prediction.
- Furthermore, the data size was limited. We have collected 700-800 articles per newspaper with covid-19 topic from 2020 to 2021. Additionally, there was no section selection and as a result there might be a tendency that the newspaper with more articles related to culture, entertainment, and leisure have a more positive sentiment score. Since the number of articles was not large enough, the few positive sentiment article's content with irrelevant topics was enough to impact on the general sentiment analysis. The noise could be reduced by increasing data size or using more specific keyword as well as section types.
- Finally, negativity in newspaper may not be a good indicator to predict the news sources as major U.S. media outlet are more likely to have negative content comparing to any other major media outlet outside the U.S., according to the study done by B. Sacerdote, R Sehgal, and M Cook[14]. As the study used covid-19 related newspaper from 2020 with 9.4 million news stores, it is worth to consider the finding when interpreting the sentiment analysis result.

6.4 Next Steps

A number of follow up studies can be conducted to further quantify bias across news sources. Examples:

- Further studies on how sentiment differs on subtopics across news sources, such as business.
- Conduct longitudinal studies to see how consumers react to different news sources (A/B testing).
- Incorporate more conservative news sources and compare bias.
- Increase number of articles.
- Add additonal sources to capture sentiment from consumers such as social media (YouTube, Twitter).
- Engineer additional features, such as subsections articles belong to, e.g. (Politics, Business, Health, etc.).
- Expand study to international sources to study how sentiment varies across countries).

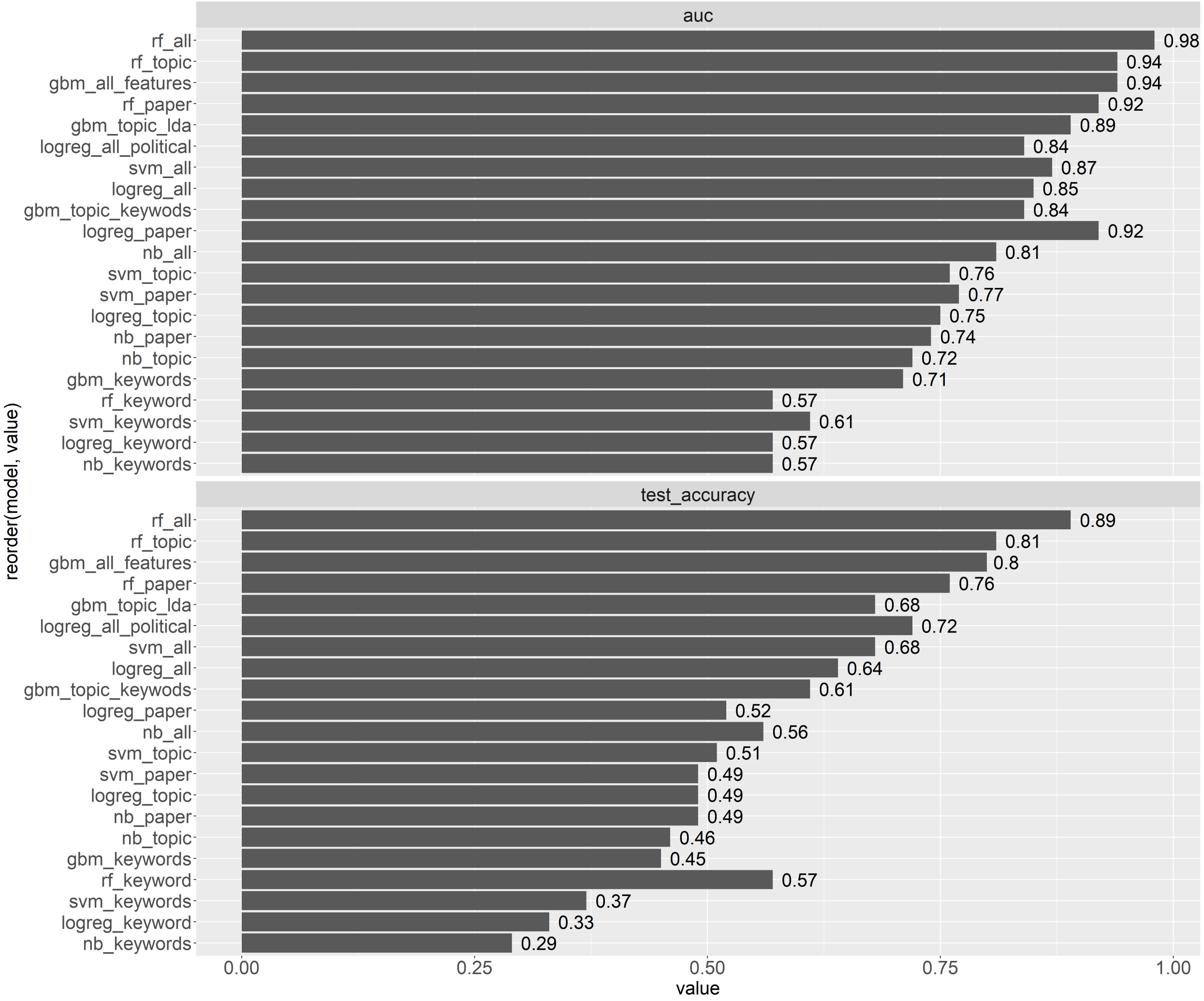


Figure 10: Model Results

7 Reference

[1] Flood, B. (2021, March 2). Fox News finishes February as most-watched primetime network. Fox News Live. https://www.foxnews.com/media/fox-news-finishes-february-most-watched-primetime-network

[2] King, G., Schneer, B., & White, A. (2017). How the news media activate public expression and influence national agendas. American Association for the Advancement of Science. Vol.358 (6364), pp.776-780. https://science.sciencemag.org/content/358/6364/776

[3] Holman, E., Garfin, D., & Silver, R. (2014). Media's role in broadcasting acute stress following the Boston Marathon bombings. Proceedings of the National Academy of Sciences of the United States of America. Vol.111 (1), pp.93- 98. https://doi.org/10.1073/pnas.1316265110

[4] Sue & Bill Gross School of Nursing, UC Irvine (2020, May 7). How (and why) coronavirus is changing our sense of time. University of California News. https://www.universityofcalifornia.edu/news/how-and-why-coronavirus-changing-our-sense-time

[5] Gorvett, Z. (2020, May 12). How the news changes the way we think and behave. BBC news. https://www.bbc.com/future/article/20200512-how-the-news-changes-the-way-we-think-and-behave

[6] (2013). Journalism Essentials. American Press Institute. https://www.americanpressinstitute.org/journalism-essentials/bias-objectivity/understanding-bias/

[7] Mitchell, A., et.al. (2014). Political Polarization & Media Habits. Pew Research Center. https://www.journalism.org/2014/10/21/political-polarization-media-habits/

[8] Hart, P., Chinn, S., & Soroka, S. (2020). Politicization and Polarization in COVID-19 News Coverage. SAGE journals. Vol.42 (5). https://journals.sagepub.com/doi/10.1177/1075547020950735

[9] J. Green, et. al. (2020). Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. American Association for the Advancement of Science. Vol.6 (28). https://advances.sciencemag.org/content/6/28/eabc2717

[10] Schaeffer, K. (2021). Despite wide partisan gaps in views of many aspects of the pandemic, some common ground exists. Pew Research Center. https://www.pewresearch.org/fact-tank/2021/03/24/despite-wide-partisan-gaps-in-views-of-many-aspects-of-the-pandemic-some-common-ground-exists/

[11] Aslam, F., et. al. (2021) Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak. Humanities and Social Science Communications. Vol.7 (23). https://doi.org/10.1057/s41599-020-0523-3

[12] Rustam F, et. al. (2021) A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. PLOS ONE. Vol.16 (2). https://doi.org/10.1371/journal.pone.0245909

[13] News as a Source: Choosing & Using Sources: A Guide to Academic Research. Pressbooks, Ohio State University. https://ohiostate.pressbooks.pub/choosingsources/chapter/news-as-a-source/

[14] Sacerdote, B., Sehgal, R., & Cook, M. (2020). Why Is All COVID-19 News Bad News? National Bureau of Economic Research, Working Paper, Working Paper Series(28110). http://www.nber.org/papers/w28110

[15] Slinge, J. & Robinson, D. Text Mining with R: A Tidy Approach. https://www.tidytextmining.com/topicmodeling.html