Differences In Bias Accross News Sources and Subjects

Brett_Biscoll

2023-05-10

Contents

Background

Bias is omnipresent in the modern press landscape; nearly every source is considered to have a slant in one direction or another. The more contentious question is which specific publications and stories are the most biased, and how do any two given sources compare.

A common thread is that opposing political factions will often accuse publications with opposing viewpoints - or at least the impression of such - of being unfairly biased against them, declaring them to be have a tilted narrative while other political figures get a pass. Due to the difficulty in objectively identifying the level of bias from text, most of this criticism goes past each other without much examination. However, text-as-data methodologies offer the opportunity to form a baseline for answering the question: what are the topics and subjects that news sources with different directions of bias diverge on the most?

Literature Review

The subject of bias has received extensive research by academics. However, much of the recent literature focuses on how misinformation spreads, rather than the degree or character of spin in truthful publications; this is covered in *Political polarization, misinformation, and media literacy* (Gaultney) from Texas State University. Some analysis of content itself has been performed in articles such as *Worth to Share? How Content Characteristics and Article Competitiveness Influence News Sharing on Social Network Sites* (Karnowski) and *Utilizing overtly political texts for fully automatic evaluation of political leaning of online news websites.* (Zhitomirsky-Geffet) which looked into political texts specifically. However, the creation and identification of specific distinct varieties of bias is slightly more obscure, due to the relative difficulty of identifying and classifying this textual lean relative to outright deception; an incorrect statement of fact can be classified as objectively untrue, while many subjective statements can be true on the surface but have a deceptive subtext.

Research Methodology

Goal

The goal of my assignment was to identify specific types of bias, create predictors that could be applied to any body of text, and classify texts as having a distinct bias present or not.

Categories of Bias

For identifying different bias categories, I turned to publicly available resources. While the academic basis was mixed, I found several sources for informing news readers about potential types of biases; I leveraged the *News Literacy Project*, *Columbus State Library*, and *AllSides.com*, the site of a media company specializing in identifying media bias, to develop a groundwork for understanding and classifying bias. Several categories overlapped across the different sites; for this project, I focused on subjective language, sensationalism and lighting (as in positive or negative).

Subjectivity The first of the bias types I selected was subjectivity; this being defined as the presence of language that was not based in empirical evidence, or was heavily influenced by emotional or personal predilections. While not all subjective language should necessarily be considered biased - for example, calling a natural disaster a 'tragedy' is technically a subjective statement, but not one that many would faults a publication for using - a greater presence of subjective headlines in (for example) crime articles may suggest an attempt to spin stories into a greater narrative than is otherwise warranted.

Sensationalism The second bias category I selected was sensationalism. Sensationalism is defined by the Reporter at RIT as "a tactic used in an attempt to gain an audience's attention. Media outlets resort to the use of shocking words, exaggeration and sometimes blatant lies." (Vanacore). Sensationalism can be identified by the presence of attention-grabbing headlines; two headline examples of sensationalism identified from my coded data would be *"Oprah: Dr. Oz Running For Senate Is An Example Of 'One Of The Great Things About Our Democracy' "* (Gentile) and *"Ronda Rousey Announces Life-Changing News After Giving Birth To Daughter"* (Jerkovich). Giveaways

for sensationalism in articles include the presence of superlative language, maximized 'stakes' from a news event, and (for headlines specifically) titles that lead on the reader with the promise of more salacious information in the body, sometimes called 'clickbait'. Sensationalism is a key technique in grabbing attention for building a narrative, and the higher presence of it can be suggestive of bias.

Positive and Negative Lighting The last category I selected was positive or negative lighting, meaning how a news story is portrayed. Like the prior examples, this type of bias can be present even in articles with the same factual information; for example, the hypothetical news story of '10,000 more minimum jobs added' in a local area could be spun as an decrease in unemployment or an expansion of poorly paying work. For my news outlets, I identified the possibility of excessively positive lighting on an article as a potential bias indicator, since the outlets should be covering generally the same events over the same time frame, and major deviations are likely an indication the source is tilting its coverage.

News Sources

With the bias topics identified, the next step was to identify news outlets upon which the models can be run. The United States offers a surplus of options to choose from; an index hosted at projects.iq.harvard.edu lists no less than 176 publications. While this project could be expanded to as many news sources as are interesting, for the sake of focus and bandwidth restrictions I selected a handful of four major outlets; the Associated Press, CNN, Fox News and The Daily Caller. My intention was to cover the spectrum for potential biases; CNN and Fox news representing opposite ends of the political spectrum (left and right), while Associate Press and Daily Caller representing opposite ends of the perceived 'objectivity' spectrum. Additionally, the presence of these news sources at LexisNexis' NexisUni, the source I leveraged, provided the additional benefit of accessibility.

News Topics

In addition to having a diverse set of sources, I wanted to inspect a cross-section of the different subjects news touches on; my belief was that the presence of bias may be especially pronounced in so-called 'hot button' topics. Again, NexisUni had pre-existing topic labels applied to news articles; as a result, I was able to select specific topics and filter by news source (for example, pulling the last 100 CNN articles discussing crime), which offered a fairly easy comparison across sources and topics. As for the topics themselves, I focused on what I considered the most polarizing news topics in the current moment; crime, government news, legal news, and popular trends. Additionally, I selected a catch-all general category as a baseline for the news stories each outlet reports on.

Model Creation

My process for each of the three subjective bias sources began with pulling the test data from NexisUni. For each combination of source & topics, I pulled the maximum number of articles permitted in a single request (100 articles), sorted by the most recent, with a date cutoff at the end of 2022 for the training sets. The intention of this was to ensure that a different time period was pulled for the training data rather than the test data. I pulled the text data in PDF files format, and constructed a corpus from the data within by looping the pdf_text() function across

the articles. I exported 100 articles for each type of bias into a csv file for coding. For subjectivity and sensationalism, I marked two classes; 0 for the absence of the bias in question, and 1 for the presence. For lighting, I used 1 and 0 as well, but with 1 to represent positive lighting and 0 to represent negative (or neutral) lighting. Once the coding was complete, I imported the data once again into R, and reconstituted the corpus with the coded results.

Once the corpus was loaded back in, I worked on preprocessing the coded datasets. For stopwords, I used the stock set of English stopwords, plus an additional list of words and phrases specific to news - this included phrases such as 'All Right Reserved', stand-ins of non-textual data the pdf_text() function could not interpret ("Image", "Graphic" etc) and a handful of additional terms I saw as non-predictive red herrings- In particular, I removed major news source names from the articles to prevent the model itself from becoming biased by a source's origin rather than its text. Additional preprocessing was applied by removing punctuation, separators, converting the tokens to lowercase and applying stemming; the corpus was converted into a DFM by this process.

This DFM was then converted into a matrix and split into testing and training data for the model specifically. For this project, I utilized support vector machines as the predictive model. The SVM's were relatively lightweight (compared to neural networks), which facilitated running the predictive models the dozens of times necessary, and additionally were relatively easy to scale to additional datasets, with some modification. I trained the SVMs on the training matrix built off the coded data, and generated confusion matrices for each to identify accuracy statistics. Additionally, for the accuracy statistic specifically, I used a k folds measure with 5 iterations to determine a consistent average for each model.

Model Results

The average accuracy results for the models were .66 for the subjectivity model, .69 for the sensationalism model, and .72 for the positive and negative lighting model; not perfect, but satisfactory for making an analysis. For the other statistics generated for the single confusion matrix, I yielded precision results of .46 for subjectivity, .73 for sensationalism, and .82 for positive and negative lighting; the latter two were satisfactory, but unfortunately the precision statistic for the subjectivity model was not as precise as I had desired. I proceeded with applying the models into the other datasets as the accuracy was fairly good.

Results

Application on Test Sets

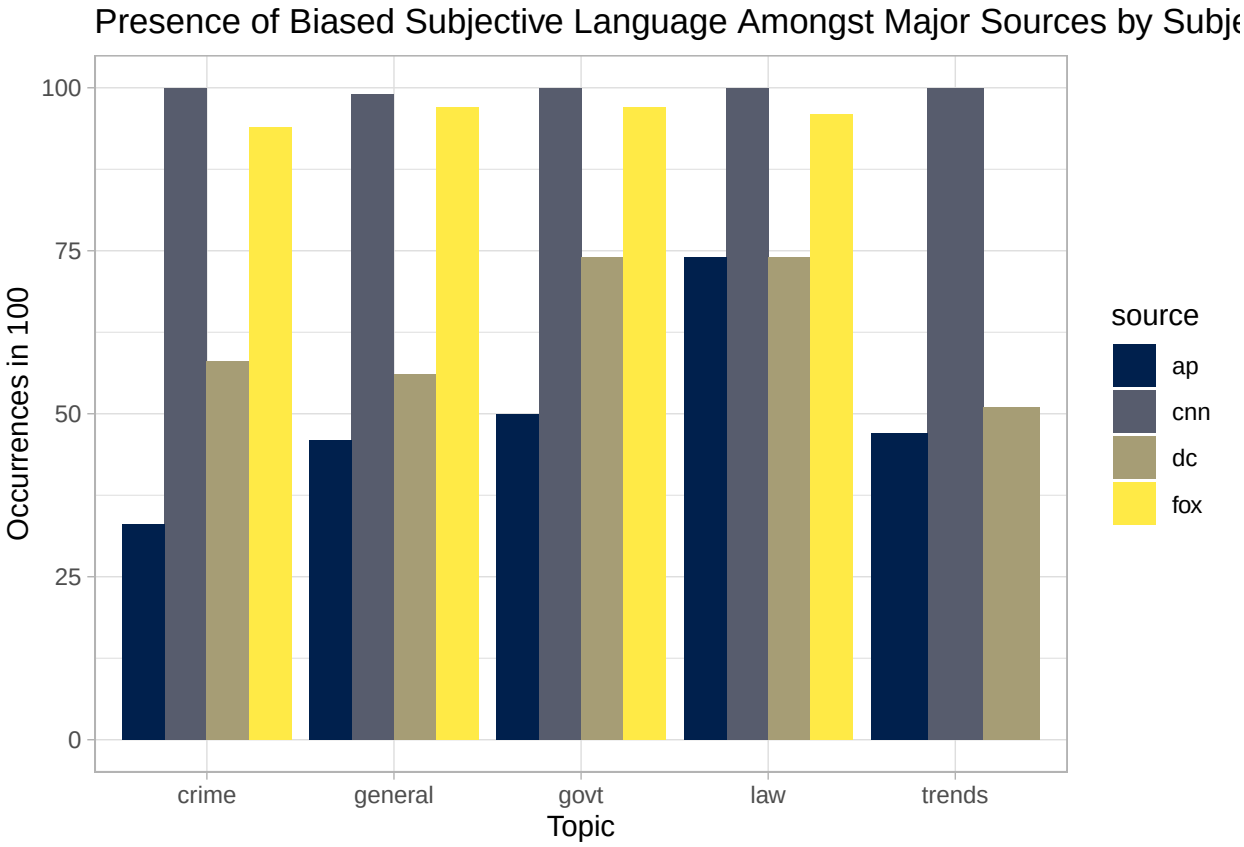
The major issue in applying the SVMs to the new test datasets was that the list of tokens/features was dissimilar to the training matrices the models were constructed on; certain columns present in the training sets were not present in the test sets, and vice versa. As a result, an error 'test data does not match model' was thrown. To resolve this, I adapted the test sets to the models by essentially identifying the column names that overlapped, selecting only those columns, then for the missing columns (i.e. instances where a token appeared in the training set but not in the test set), I added those missing features back with '0's to accurately reflect their absence. Finally, I generated predicted results for each of the 100 articles in the given test set; this generated a vector of 0s and 1s, with 1s indicating presence of bias (or positive lighting) for that given instance. I

bound the results together for each bias topic and source into a 'results dataframe', and exported the file for review.

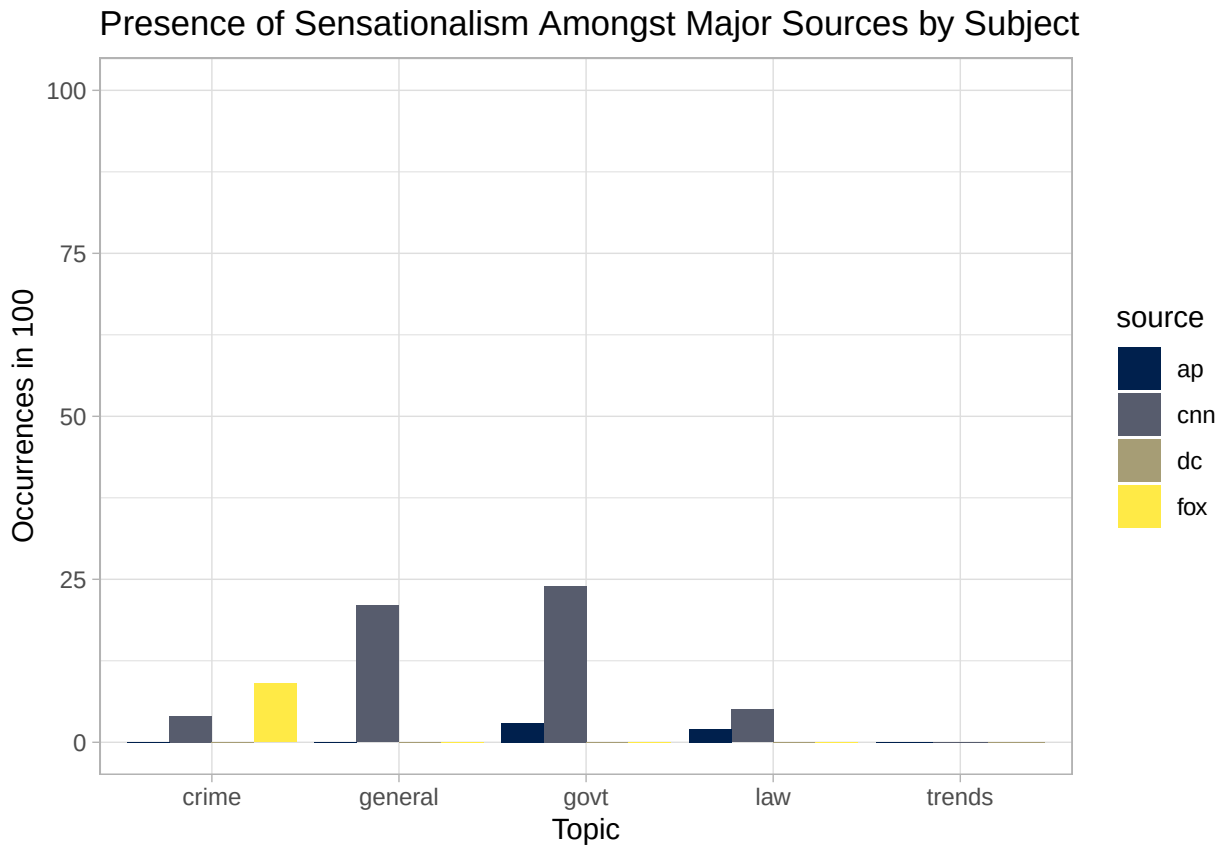
Data Analysis

For analysis, I visualized the data with a set of bar charts. I looked at how prevalent a type of bias was for an entire source-topic combination, instead of focusing on specific examples within the datasets.

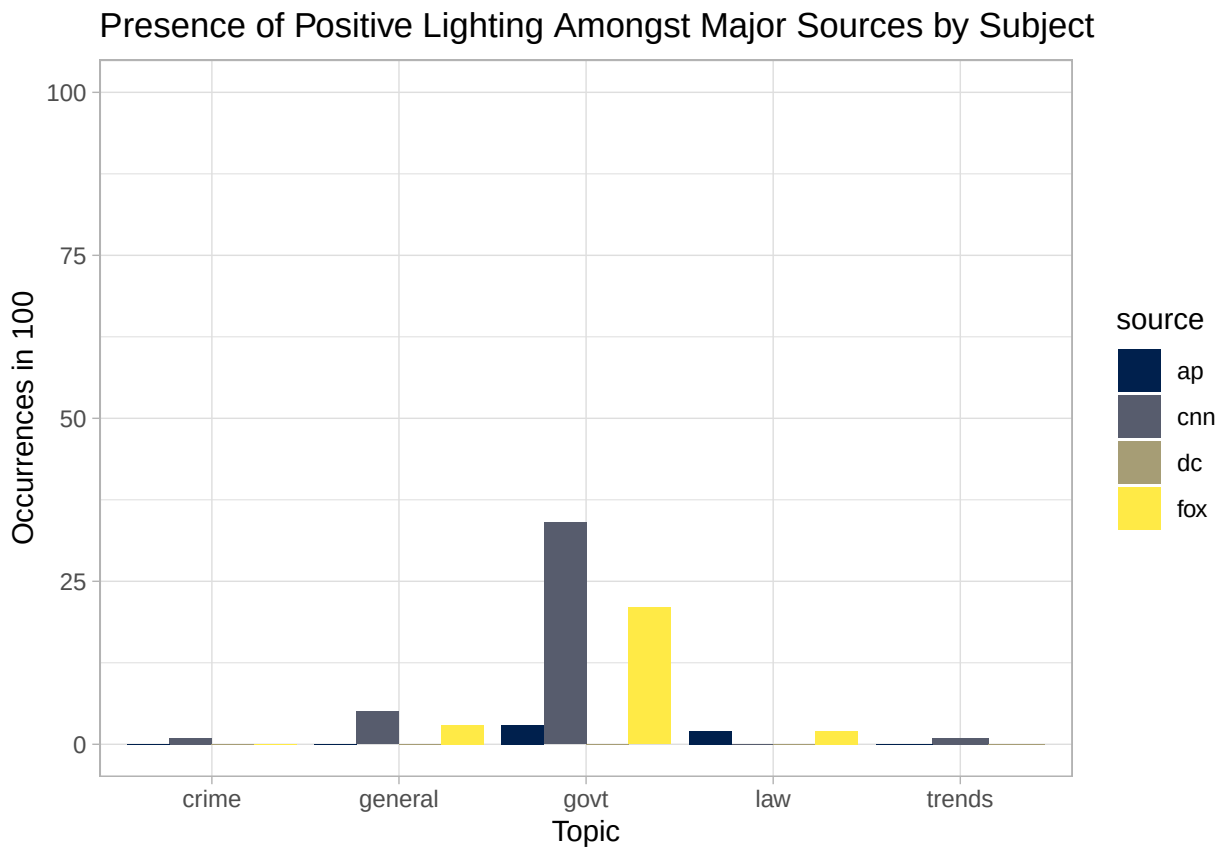
Subjectivity had the greatest prevalence of the bias types across the board; the lowest occurrence of subjectivity was among AP articles relating to crime, and even then there were still 33 occurrences of subjective bias of 100 total. Noticeably, Fox and CNN had the highest rates of subjectivity, with occurrences in over 90% of articles across all subjects (excluding trends for Fox, where there was an issue importing the data). The Daily Caller has the second highest, in the mid 50s to 70s for each subject; while the Associated Press was the most objective, even if subjectivity still occurred in many cases.



Sensationalism was not as frequent an occurrence, happening at most in 24 cases for CNN articles on the government. Notably, CNN has the highest rate of sensationalism on any given topic, with most frequent occurrences in the government and general categories; Fox was the runner up, with a notable level of sensationalism for crime articles.



Lastly, the presence of positive lighting was also comparatively sparse, never exceeding 50%. There were two notable source-topic combinations that made extensive use of positive lighting; those being government stories on CNN and Fox. The Daily Caller never registered positive lighting, while the AP sources rarely did.



The primary takeaway from these findings as that, of the types of bias investigated, subjectivity is the primary tool for conveying a the slant of a given news source. With respect to sensationalism, CNN was the most prone outlet, with the news subjects of government and general having the most sensational news stories; as runner up, Fox had 9 sensationalist news stories, all in the crime subject. Lastly, for lighting, government stories were the most prone to positive lighting by both Fox and CNN; other instances of positive lighting were not unheard of, but negligible in comparison.

To answer the thesis question at the beginning of this report, sources of news diverge the most in their use of subjectivity on articles; on government news, and in favoring sensationalism in different subjects of news- Fox, being considered right wing, had sensationalism mostly in criminal news stories, while CNN had sensationalism in government and general stories.

Further Research Possibilities

For future research in this area, I would expand upon both the sources analyzed and the subjects covered. To have a representative and manageable section of the news 'market', I focused on a small subset of news publications and a limited set of subjects I thought were most likely to show divergence; future research may uncover less known subjects or identify strains of bias in news media that aren't readily apparent among this set.

Furthermore, a time-series analysis of news articles may show change in bias over time, and lend a quantitative backing to the narratives present in news media. Recent narratives have included gun violence and a potential national debt default, and being able to see a divergence in bias present in left and right news publications may begin to offer a timeline of when these stories became a main

focus. Being able to identify when news stories diverge could reveal when new narratives of bias are taking hold, and allow such information to become more readily available to news consumers. This sort of notification of potentially biased news (sometimes known as *pre-bunking*) could help prevent insincere narratives from taking hold and allow for a more honest conversation around public affairs.

As for the models themselves, I was limited again by bandwidth. In developing the training datasets, I would at the very least have an additional user help to code; since bias is rather subjective, my concern is that having one person (myself) identify bias across the datasets would itself show a bias and unfairly attribute bias to one source or anther. we could code the data sets in parallel, and if there was a disagreement on bias on a given piece of news, we could meet or have a third user act as a 'tie-breaker' to hopefully have a more objective rating of bias presence. More coders would help reach objectivity, but there is a certain degree of subjectivity that likely would not be removable; another approach would be having a large set of coders (say, 100) and instead of a 'bias is present, true/false' response, make it a rating resulting from the collective grades of the coders (ex. 66% of coders identified bias). Again, the manpower needed for this far exceeded what was available to for this project, but further investigation would benefit from additional eyes in making the judgment calls.

Works Cited

"Balanced News via Media Bias Ratings for an Unbiased News Perspective." AllSides, 20 Oct. 2022, www.allsides.com/unbiased-balanced-news.

Gaultney, Ira Bruce, et al. "Political Polarization, Misinformation, and Media Literacy." Journal of Media Literacy Education, vol. 14, no. 1, 2022, pp. 59–81., https://doi.org/10.23860/jmle-2022-14-1-5.

Gentile, Luke. "Oprah Says Dr. Oz's Senate Run Is an Example of 'One of the Great Things about Our Democracy'." Washington Examiner, Washington Examiner, 10 Jan. 2022, h ttps://www.washingtonexaminer.com/news/oprah-says-dr-ozs-senate-run-is-an-example-of-one-of-the-great-things-about-our-democracy.

"Index of US Mainstream Media Ownership." Index of US Mainstream Media Ownership, projects.iq.harvard.edu/futureofmedia/index-us-mainstream-media-ownership. Accessed 9 May 2023.

Jerkovich, Katie. "Ronda Rousey Announces Life-Changing News after Giving Birth to Daughter." The Daily Caller, The Daily Caller, 28 Sept. 2021, https://dailycaller.com/2021/09/28/ronda-rousey-announces-giving-birth-daughter/.

Karnowski, Veronika, et al. "Worth to Share? How Content Characteristics and Article Competitiveness Influence News Sharing on Social Network Sites." Journalism & Mass Communication Quarterly, vol. 98, no. 1, 2020, pp. 59–82., https://doi.org/10.1177/1077699020940340.

"Library: Bias in the Media: Types of Media Bias." Types of Media Bias - Bias in the Media - Library at Columbus State Community College, library.cscc.edu/mediabias/typesofmediabias. Accessed 9 May 2023.

"News Literacy Project." News Literacy Project, 9 May 2023, newslit.org/.

Rylan Vanacore | published Nov. 12th, 2021. "Sensationalism in Media." Reporter, https://reporter.rit.edu/news/sensationalism-media.

Zhitomirsky-Geffet, Maayan, et al. “Utilizing Overtly Political Texts for Fully Automatic Evaluation of Political Leaning of Online News Websites.” Online Information Review, vol. 40, no. 3, 2016, pp. 362–379., https://doi.org/10.1108/oir-06-2015-0211.