News Bias

Alexander Lambert
Computer Science, B.S.
The University of Tennessee
Knoxville, Tennessee
alambe22@vols.utk.edu

Casey Mathews
Computer Science, B.S.
The University of Tennessee
Knoxville, Tennessee
cmathew9@vols.utk.edu

Shivam Patel
Computer Science, B.S.
The University of Tennessee
Knoxville, Tennessee
spatel95@vols.utk.edu

I. Objective

What similarities and differences exist between word usage and patterns in both politically biased and scientifically accurate and inaccurate articles? To find out, we analyzed the following parts of the articles:

- Buzzwords and phrases count
- Emotional word count
- Average word length
- Use of words with negative connotations
- Use of words with positive connotations
- Use of words that indicate opinion (I think, I believe, etc.)
- Use of words that indicate fact (research, studies, etc.)
- First person pronoun usage (Does the author present this as their perspective, or as information)

Using the data we gather, the hope is to find patterns that could be used to analyze new articles for bias or factuality.

II. Data

The data that we used included: articles, article content, and datasets. We obtained the articles using https://mediabiasfactcheck.com/, a site that categorizes articles into different biases, from left-leaning to right-leaning and pseudoscience to pro-science. Using this tool, we grabbed a number of different articles from various biases. Then, using the algorithms discussed in the next section, we retrieved the relevant content from the chosen articles. Relevant content included the body of the article, list tags, and the alt-text of images. The last data we gathered was datasets for the various filters we wanted, which included:

- Emotional words
- Buzzwords
- Positive words
- Negative Words

Other datasets we used were made by ourselves, which included

- Fact phrases
- First person words
- Opinion phrases

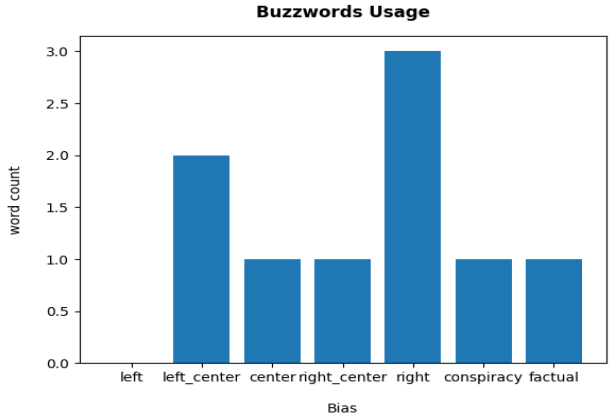
Once we acquired the data, we processed it in a few different ways. Articles were web-scraped and the relevant information was stored as a string inside of our program. These strings were then passed into functions that filtered them against the datasets we had gathered. For each article we also found the average word length. Once all the articles were processed, we analyzed the data by making graphs and comparing the results.

After the article contents were filtered against the datasets, we needed to verify that the filtering had worked properly. In order to this, we manually checked the results. As an article was filtered, we printed out the words it found and how many, which then searched the article's HTML for those words. After checking a sample of the articles we had, we determined our filtering grabs the correct count in almost every case, except for one edge case we were unable to fix. If a list item had a link inside of it that contained the words we were filtering for, the text in the link would be counted.

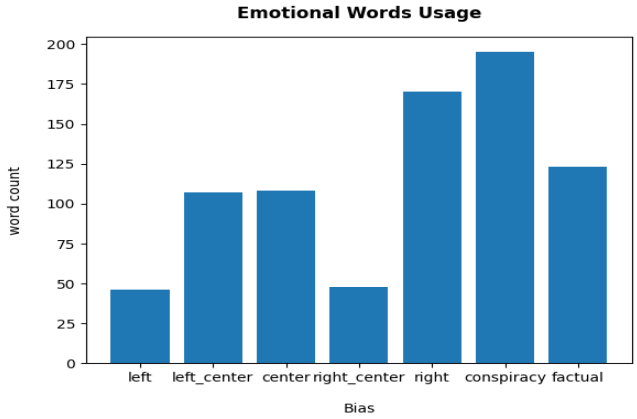
III. Models and Algorithms

We used BeautifulSoup4 to obtain the HTML for online articles and used its tree structure for parsing the articles. First we passed the URL to the news article into BeautifulSoup4. Then, since the majority of the text within the articles was contained in $\langle p \rangle$ and $\langle li \rangle$ HTML elements as well as image element alt text, those elements were targeted for obtaining the content of the articles. We got every word contained within those elements and passed them through some regular expressions to remove punctuation and Unicode characters to help with word comparisons for later. After cleaning up punctuation and special characters, each word was added to a list containing all of the words for that article. Each complete list was used as the input for our code that calculated data about the words used in the articles.

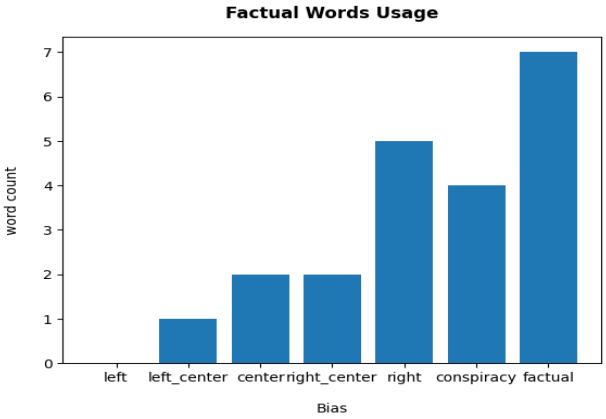
IV. Results



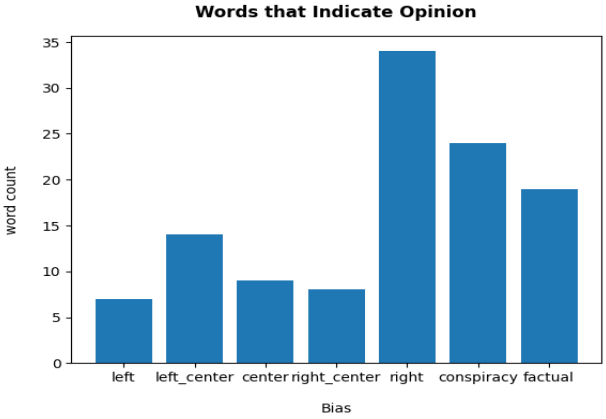
The first filter was for buzzwords. We checked for various politically charged buzzwords, ranging from ”hanging chad” to ”killary” and ”pro-choice”. As can be seen in the graph, our results showed that Left Center and Right biased sources tended to use the most buzzwords. Our initial hypothesis was that both Right and Left biased sites would heavily use buzzwords, so it is interesting to see that this was only partially true. This could be due to our buzzword lists contents, but a pattern is present that Right biased sources tend to use buzzwords.



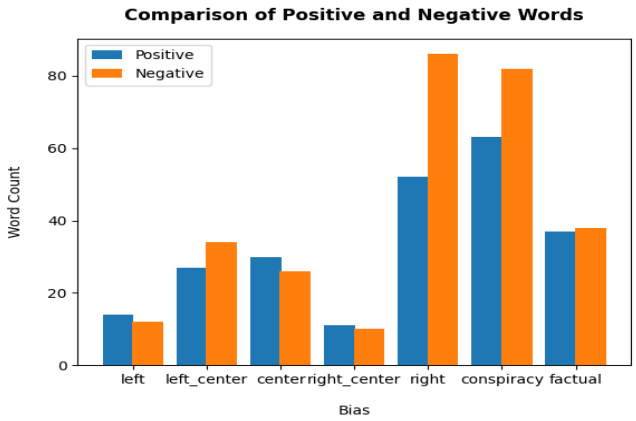
The next dataset considered was for emotional words. These are words that spark emotion in the reader, like ”achieve”, ”alienation”, and ”negative”. Our initial prediction was that heavily biased articles would use lots of emotional language. What we found was that Right biased articles and Conspiracy articles used lots of emotional language, which fits with our prediction. However, Left biased articles used less emotional language than the more Center biased articles. Strangely enough, Factual articles where the third highest for using emotional language. Overall there may still be a pattern here for helping to identify both Right and Left biased sites, along with Conspiracy sites.



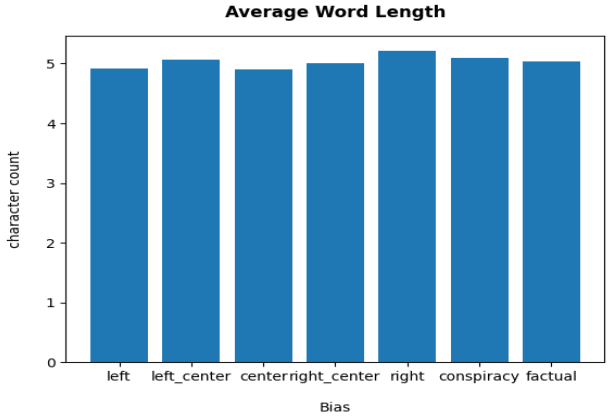
Following emotional words, we checked for ”factual words”. This dataset included words like ”surveys”, ”data”, and ”research”, with the goal of trying to see if an article was trying to show that it had the factual truth. Our prediction was that unbiased, or the three Center categories, to have the highest usage of factual words. We also expected Factual and Conspiracy articles to have a similar usage, as one is presenting research and the other is trying to appear like they are presenting research. In this case, we found that Factual sites used fact words more than any other, which makes sense. Right biased and Conspiracy sites also used ”factual words” a lot, so a pattern may be seen with these three. Left biased articles appeared to use no fact words, which may also be a useful pattern.



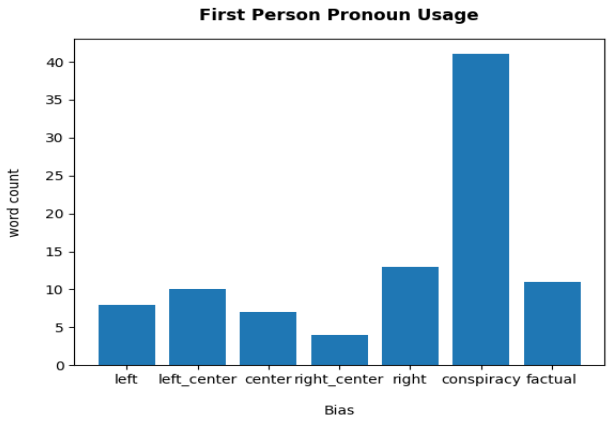
After filtering against emotional words, we checked for words indicating opinion. These include ”believe”, ”expect”, and ”my sense is”. We had initially expected that Center and Center leaning articles would have the lowest opinion word count, as they are trying to just report information. What we found is that this is largely true, with Right biased sources having the highest usage of opinion words. Conspiracy and Factual seem to be rather close, which is a bit surprising. Pattern wise, it seems that Right biased sources tend to use opinion, and more Center and Left biased sources do not use opinion much.



The graph shows the comparison of using positive and negative words for each bias. Positive words can be defined for words like ”achieve”, ”admire”, and ”enjoy”. In contrast, example of negative words are ”chaos”, ”garbage”, and ”nonsense”. The prediction was that conspiracy or biased articles have the highest usage of positive and negative words. The results mostly matches with our prediction because right leaning and conspiracy articles have the highest usage of positive and negative words. A general pattern to note is that negative words are used much more in right leaning articles and conspiracy articles. The other biases were pretty even in using positive and negative words.



This graph represents the average word length in an article for different types of biases. We expected conspiracy based or scientific articles to have a longer word length because they tend to use jargon and longer words. However,we can see a pattern that all different biases are almost the same average word length at around 5 characters.



The bar graph represents the usage of first person pronouns like ”i”, ”we”, ”our”, ”mine”. This is to help indicate if the author is presenting the information from his point of view or giving the information as facts in third person. Clearly, the highest are conspiracy articles with an average around 40 words. The other biases were no more than 15 words. We predicted unbiased articles to have a lower usage of 1st person pronouns and we can see that is partially true with center and right center articles having the lowest word count. No major pattern is recognized besides seeing that conspiracy has a much larger first person pronoun usage compared to the rest of the biased articles.

A. *Conclusion*

Our results show some patterns do exist in the word usage of articles of various biases. In particular, for many categories, Left and Right biased are opposites. Also, we see Conspiracy is often the opposite of Factual. Further research and refinement of this process could allow one to detect the leanings of a new article from its contents alone.

V. Issues

The main issue was that we had to hand pick our articles. This may have led to selection bias. Another issue is that we do not store the contents of the articles and only the URLs since we do not generate a fresh batch of article URLs each time we gather results. That means that each time we run our program, there is no guarantee that each URL is still valid and the program will work properly. Another issue that we ran into while creating this project was having to hand verify our results. This is much more prone to mistakes than an automated process would be. Another issue we had in the design of the project was the creation of many of the data sets to compare the content of the articles against. To determine how many buzz words were contained within an article, we had to create a dictionary of buzz words to compare the text from the article against. This might also lead to some sort of selection bias as well as potentially incomplete results. Lastly, there is an issue with getting words contained within the $\langle li \rangle$ HTML elements. Sometimes when an $\langle li \rangle$ element contains a link to something, it will get the words contained from the URL as well as the words that are used for the hyperlink which can sometimes result in counting the words within the URL

twice. This can potentially slightly skew the results for some articles.

VI. Future Work

The following are ideas for future analysis and ways we can make the process better.

- A web scraper that can autonomously receive data from different types of bias articles (left leaning, right leaning, etc.) so that we wouldn't need to manually pick the articles.
- A validator so we can check for accuracies instantly instead of manually checking the articles for things like accurate count of emotional words.
- Compare the number of pictures in an article
- Analyze if the articles used references
- Store the article's content in a database so it is guaranteed to have the information later for future analysis
- Try and use the results we find to classify a given article's bias.

VII. Org Chart

A. Responsibilities

- Alex Lambert: Designing Filtering, finding articles and datasets.
- Casey Mathews: Web Scraping, finding articles and making datasets.
- Shivam Patel: Graphs and charts, finding articles, and making datasets.

B. Timeline

- By October 20$^{\text{th}}$:
  - Emotional, positive, negative word filtering
  - Datasets for emotional, positive, negative, and buzzwords
  - 15 articles found
  - $\langle p \rangle$, $\langle li \rangle$ and image alt text scrapped.
- By November 5$^{\text{th}}$:
  - Fact and opinion datasets added
  - Filtering code made generic to work with any given dataset
  - All articles found
- By November 16$^{\text{th}}$:
  - Combined work into one notebook
  - Added pronoun check, initial processing of articles
- By December 6$^{\text{th}}$:
  - Report written
  - Results analyzed
  - Accuracy checked