Proposal: Analyzing Differences and Identifying Biases in News Sources

DATA 450 Capstone

Kamila Palys

February 8, 2024

1 Introduction

All people have their own biases, whether they are conscious or subconscious. Larger entities and organizations are also known to show bias across many industries, and bias coming from widely known or used sources can be even more dangerous because of how influential ideas can be. In particular, this project will take a close look at biases present in various news sites. These sometimes have to do with the political party that some sources are affiliated with. Sources may also report some news with a positive or negative connotation, when they really have a responsibility to inform of news in a neutral way, not give their opinion on them. This project hopes to provide insight into the word choice of selected news sources and which ones may have a tendency to be more biased so that readers may make informed decisions on where to hear about news.

2 Dataset

[The dataset used for this project will be presented in various formats to allow for the usage of different techniques. One format will have each feature be a token, or word, each row be an article, and the values be a 0 or 1 to represent if the token appears in the article. Another format will have the same columns and rows, but the values being the frequency of the token appearance in the article. In each format, there will also be one column that will tell which news site the article comes from. The articles will come from the following sources: CNN, The New York Times, The Washington Post, Fox News, New York Post, MSNBC, NBC, MSN, The Wall Street Journal, and BBC.

In this section, desribe the dataset(s). This includes things like where you obtained the dataset. Include a full citation, as specified here. Describe how the data was obtained by

the data owner/curator, as best as you can. List the variables that you plan to use in your analysis, for example:

- weight: The patient's weight (kg)
- sex: The patient's sex, male or female
- age: The patient's age (months)

]

3 Data Acquisition and Processing

The data used in this project is collected from a variety of sources. Ten total news sites are studied, consisting of CNN, The New York Times, The Washington Post, Fox News, New York Post, MSNBC, NBC, MSN, The Wall Street Journal, and BBC. For the training set, five major political events and topics from 2023 are covered from each of sources, for a total of fifty articles. The events covered are the indictment of Trump, the Supreme Court's ruling against affirmative action, Biden's low approval rates in polls, the Chinese surveillance balloon, and the deadliest attack by Hamas to date. For the testing set, an additional 2 events and topics are covered from each source for a total of twenty articles. These two events include the Pentagon documents leak and the Russia-Ukraine War. This will make for an approximately 70/30 training/testing split.

The text from each of the articles will be copied and pasted into a text file. Some preprocessing to be done on this text will include, but not be limited to, the removal of punctuation and stopwords, making all words lowercase, and stemming. All of the words, or tokens, from all articles will be brought together into a single pandas dataframe. Each token will be made into a feature, with an additional feature being made to represent the source of the article.

[In this section, if applicable, describe how you will obtain the data (if it's anything more complicated than a simple download). Discuss what data processing steps will be needed, such as recoding variables, data cleaning, data tidying, imputing missing values, etc. See sections 1c, 1d, 1e in the "Good Enough Practices" paper.]

4 Research Questions and Methodology

[In this section, list each of the questions you will explore. Following each question, provide a detailed and specific plan for how you plan to answer the question. Include the specific steps you will take, what form the answer will take (a number? table? visualization? model? Give all the specifics), and estimate how many hours each question will take to complete.]

1. Are the news sources that are similar to each other in wording associated to the same side of the political spectrum? To answer this, I will perform hierarchical clustering on the news sources. To cluster news sources as a whole, I will have to represent each row as a combination of all the articles from a single news source. I will display the results in a dendrogram and observe which news sources were clustered together first, meaning they were most similar. Based on commonly held public opinions shared on the internet, I will then determine if the sources clustered together lean the same way politically.

2. How does wording differ between the news sites? This question will be answered by comparing the frequencies and TF-IDF's of each word in each news site. Similarly to the last question, for this question, each instance will have to represent the whole news source and combine all of its articles. Then, this will be visualized in word clouds, with one word cloud being created per source. Insights will be given by comparing most frequent or significant terms across sources that are supposed to be reporting on the same events.

3. Can some news sources be said to be more positive or negative than others? This question will be answered with the help of various Python libraries built for sentiment analysis, like Natural Language Toolkit (nltk) and TextBlob. These libraries will assign a polarity score to each article to represent how much of a positive or negative connotation it has. The articles will have to be saved as a string for this question. I may also compute the average score per source. Scores may be visualized in a bar chart, with the y-axis representing the score, each bar representing an article, and the bars being grouped by the news source.

4. Is the wording of various news sources different enough to be able to correctly predict which source an article comes from? This question will be answered through machine learning. Models will be trained and tested on the respective sets of data to be able to predict which news site an article came from, given its raw text. Functions will be created along the way that will be able to take any article text, such as from the testing set of articles, and perform the necessary preprocessing to a dataframe. Some classifiers to be used for this modeling are Naive Bayes, Random Forest, and Stochastic Gradient Descent (SGD), given that these handle multiple classes. Accuracy of the models will be looked at as the performance metric and a high accuracy may imply that there is a strong distinction in the wording between these sources. A confusion matrix will also be generated for each model with a heatmap on top to analyze model performance.

5. Can bias be detected in any of the studied news sources? To answer this question, bias will be determined through sentiment analysis. Results from question three may be used to compare average polarity scores across news sites and see if any one of the news sites' average scores is an outlier, determined by the interquartile range formula. Having an average score that is an outlier will mean that the source has an unusually positive or negative connotation to it compared to the other sources, indicative of bias, since news articles should be neutral. Results from question three using TextBlob will also give a subjectivity score, which measures how much personal opinion rather than factual information is in the text.

5 Work plan

Week 4 (2/12 - 2/18):

- Save all article text into text files (2 hours)
- Data cleaning (remove stopwords, punctuation, etc.) (3 hours)
- Create dataframe with binary values (2 hours)

Week 5 (2/19 - 2/25):

- Create dataframe with term frequency values (1 hour)
- Create dataframe with term TF-IDF values (2 hours)
- Create dataframe where each row represents all articles from one source (1 hour)
- Q1: Clustering and dendrogram (3 hours)

Week 6 (2/26 - 3/3):

- Q2: Create word clouds (2 hours)
- Q3: Sentiment Analysis and its bar charts (3 hours)
- Research past work on detecting biases or subjectivity in text data (2 hours)

Week 7 (3/4 - 3/10):

- Presentation prep and practice (4 hours)
- Q5: Calculate average polarity scores, detect outliers, calculate subjectivity scores, and interpret (3 hours)

Week 8 (3/11 - 3/17): *Presentations given on Wed-Thu 3/13-3/14. Poster Draft due Friday 3/15 (optional extension till 3/17).*

- Poster prep (4 hours)
- Q4: Naive Bayes model (1.5 hours)
- Presentation peer review (1.5 hours)

Week 9 (3/25 - 3/31): *Final Poster due Sunday 3/31.*

- Peer feedback (3.5 hours)
- Poster revisions (3.5 hours)
- [Do not schedule any other tasks for this week.]

Week 10 (4/1 - 4/7): * Q4: Random Forest model (1.5 hours) * Read about Stochastic Gradient Descent (SGD) model and its algorithm (2 hours) * Q4: Create (SGD) model (2 hours) * Calculate performance metrics of models and create bar chart comparing accuracy scores (1.5 hours)

Week 11 (4/8 - 4/14): * Confusion matrices with heatmap on top, interpret (2 hours) * Start writing up blog (2 hours) * Search for and collect articles from sources not studied (3 hours)

Week 12 (4/15 - 4/21): * Use models on those articles from other sources to see how they would be classified (2 hours) * Continue blog post and interpreting results (5 hours)

Week 13 (4/22 - 4/28): *Blog post draft 1 due Sunday night 4/28.* [All project work should be done by the end of this week. The remaining time will be used for writing up and presenting your results.]

- Draft blog post (4 hours).

Week 14 (4/29 - 5/5):

- Peer feedback (3 hours)
- Blog post revisions (4 hours)
- [Do not schedule any other tasks for this week.]

Week 15 (5/6 - 5/12): *Final blog post due Weds 5/8. Blog post read-throughs during final exam slot, Thursday May 9th, 8:00-11:20am.*

- Blog post revisions (2 hours)
- Peer feedback (2 hours)
- [Do not schedule any other tasks for this week.]

Here's an example of citing a source (see Phillips 1999, 33–35). Be sure the source information is entered in "BibTeX" form in the `references.bib` file.

6 References

[The bibliography will automatically get generated. Any sources you cite in the document will be included. Other entries in the `.bib` file will not be included.]

Phillips, T. P. 1999. "Possible Influence of the Magnetosphere on American History." *J. Oddball Res.* 98: 1000–1003.