**Bias Detection in American Journalism**

Authors: Aaron Carr, Azucena Faus, Dave Friesen

Company Industry: News Media

Company Size: 3 (startup)

GitHub Repository: <https://github.com/fausa/Bias_Detection_in_Journalism>

**Abstract:**

There is a general mistrust of news media where the public follows news that mimic their own political biases. But there is a growing appetite for unbiased news and growing subscriptions for centered news sources.

**Problem Statement:**

The question of mainstream media bias is a significant area of contention in U.S. politics. At the same time, the political 'inclination' of a news outlet often aligns with the personal perspectives of its audience. This overlap provides a unique opportunity and our project's objective: To leverage sophisticated text mining methods and supporting data science techniques to classify the political bias of leading online news sources. This classification will be based on a combination of "politically polarizing" terms as identified through an impartial (academic) source, as well as sentiment analysis context around the use of these terms (Liu et al., 2022).

**Objective:**

The objective of developing this classifier will be to advise our clients who are the executives at a news source that prides itself in being considered politically “centered.” We would use this classifier to analyze their content on a continuous basis and report back the overall/average political “lean” of their articles. The details regarding the metric for overall publication lean will be based on an average of left/right leaning probabilities per article. This feedback will then provide our client with actionable insights so they ensure their overall political lean metric remains centered.

**Goals:**

The success of this project will contribute to the achievement of the following goals:

1. Using article content from Fox News, CNN, Breitbart, and The Washington Post, along with media bias information from AllSides Media (n.d.), develop a classifier with at least a 90% F1 Score on testing data that can differentiate "left" from "right" leaning articles.
2. Train a model to predict political lean on unseen news articles that are sourced from "centered" news outlets and confirm the bias ratings from AllSides Media (n.d.).
3. Make recommendations to our client, The Hill, on whether their online journalism remains centered or has shifted in political lean, as well as provide solutions and next steps.

Ultimately, the goal of this project is to provide clients who wish to stay true to ethical, unbiased journalism with a bias rating for their news content throughout the year, giving them the opportunity to find out which articles are causing that shift and make the necessary steps to mitigate such issues.

**Name of your selected dataset:**

Queried news articles from CNN, Fox News, Breitbert, and The Washington Post covering most of May and part of June, 2023.

**Description of your selected dataset (data source, number of variables, size of dataset, etc.):**

News articles will be sourced via a REST API called NewsAPI, which logs information on “current and historic news articles published by over 80,000 worldwide sources” (NewsAPI, n.d.). Out of the many possible attributes returned for each API query, this project will use six: source name, author, title, url, publishedAt, and content. Prior to data preprocessing, an additional feature (article\_text) will be used to store the scraped data from each specific URL. The size of the final dataset (*N* = 4,026) was limited by both time restrictions, as well as web scraping access to specific sites or articles.

Queries for topics of political interest are used to gather articles from explicitly chosen sources. Independent studies show the political lean for each of these sources (CNN, “left”; Fox News, “right”; The Washington Post, “left”; Breitbert News, “right”) and this will help with training and validation of our classifier (AllSides, n.d.; Ralph & Relman, 2018).

**GitHub Data Locations:**

Master Persisted Train/Test Dataset: data/master.csv

Associated Press Dataset: data/master\_tokenized\_AP.csv

The Hill Dataset: data/master\_business\_TheHill.csv

**Methods Used:**

* Exploratory data analysis (EDA)
* Text data preprocessing (e.g., normalization, tokenization)
* Term frequency-inverse document frequency (TF-IDF) vectorization
* Train/test split
* Classification
* Machine learning
* Hyperparameter tuning
* Model evaluation (e.g., confusion matrix, F1)
* Sentiment Analysis
* Topic Modeling

**References**

AllSides. (n.d.). *AllSides Media Bias Chart*. Retrieved June 6, 2023 from <https://www.allsides.com/media-bias/media-bias-chart>

Liu, R., Jia, C., Wei, J., Xu, G., & Vosoughi, S. (2022). Quantifying and alleviating political bias in language models. *Artificial Intelligence*, *304*. <https://doi.org/10.1016/j.artint.2021.103654>

NewsAPI. (n.d.). *Search worldwide news with code*. Retrieved June 5, 2023 from <https://newsapi.org/>

Ralph, P., & Relman, E. (2018, September 2). These are the most and least biased news outlets in the US, according to Americans. *Business Insider*. <https://www.businessinsider.com/most-biased-news-outlets-in-america-cnn-fox-nytimes-2018-8?op=1#and-heres-how-republicans-ranked-them-from-fox-news-to-cnn-20Page>

**PART I:**

**Data Collection**

**PART II:**

**EDA, Text Preprocessing, & Modeling**