Political Polarization In Media Headlines

CSCI 577 - Data Mining

Matt Jensen
contact@publicmatt.com
Western Washington University
Bellingham, Washington, USA

ABSTRACT

Political polarization in the United States has increased in recent years according to studies [5]. A number of polling methods and data sources have been used to track this phenomenon [4]. A casual link between polarization and partisanship in elections and the community has been hard to establish. One possible cause is the media diet of the average American. In particular, the medium of consumption has shifted online and the range of sources has widened considerably. In an effort to quantify the range of online media, a study of online news article headlines was conducted. It found that titles with emotionally neutral wording have decreased in the share of all articles over time. A model was built to classify titles using BERT-style word embeddings and a simple classifier.

KEYWORDS

data mining, datasets, classification, clustering, neural networks

1 BACKGROUND

Media and new publishers have been accused of polarizing discussion to drive up revenue and engagement. This paper seeks to quantify those claims by classifying the degree to which news headlines have become more emotionally charged of time. A secondary goal is the investigate whether news organization have been uniformly polarized, or if one pole has been 'moving' more rapidly away from the 'middle'. This analysis will probe to what degree has the Overton Window has shifted in the media. Naom Chomsky had a hypothesis about manufactured consent that is beyond the scope of this paper, so we will restrict our analysis to the presence of agenda instead of the cause of it.

There is evidence supporting and increase in political polarization in the United States over the past 16 years. There have been a number of studies conducted in an attempt to measure and explain this phenomenon. [1]

These studies attempt to link increased media options and a decrease in the proportion of less engaged and less partisan voters. This drop in less engaged voters might explain the increased partisanship in elections. However, the evidence regarding a direct causal relationship between partisan media messages and changes in attitudes or behaviors is inconclusive. Directly measuring the casual relationship between media messages and behavior is difficult. There is currently no solid evidence to support the claim that partisan media outlets are causing average Americans to become more partisan.

The number of media publishers has increased and in this particular data set:

These studies rest on the assumption that media outlets are becoming more partisan. We study this assumption in detail.

Table 1: News Dataset Sources

Source	Description
Memeorandum	News aggregation service.
AllSides	Bias evaluator.
MediaBiasFactCheck	Bias evaluator.
HuggingFace	Classification model repository.

Party Sorting: Over the past few decades, there has been a significant increase in party sorting, where Democrats have become more ideologically liberal, and Republicans have become more ideologically conservative. This trend indicates a growing gap between the two major political parties. A study published in the journal American Political Science Review in 2018 found that party sorting increased significantly between 2004 and 2016.

Congressional Polarization: There has been a substantial increase in polarization among members of the U.S. Congress. Studies analyzing voting patterns and ideological positions of legislators have consistently shown a widening gap between Democrats and Republicans. The Pew Research Center reported that the median Democrat and the median Republican in Congress have become further apart ideologically between 2004 and 2017.

Public Opinion: Surveys and polls also provide evidence of increasing political polarization among the American public. According to a study conducted by Pew Research Center in 2017, the gap between Republicans and Democrats on key policy issues, such as immigration, the environment, and social issues, has widened significantly since 1994.

Media Fragmentation: The rise of social media and digital media platforms has contributed to the fragmentation of media consumption, leading to the creation of ideological echo chambers. Individuals are more likely to consume news and information that aligns with their pre-existing beliefs, reinforcing and intensifying polarization.

Increased Negative Attitudes: Studies have shown that Americans' attitudes towards members of the opposing political party have become increasingly negative. The Pew Research Center reported in 2016 that negative feelings towards the opposing party have doubled since the late 1990s, indicating a deepening divide.

- Memeorandum: **stories** - AllSides: **bias** - HuggingFace: **sentiment** - ChatGPT: **election dates**

2 DATA SOURCES

All data was collected over the course of 2023.

Table 2: News Dataset Statistics After Cleaning

stat	value
publishers	1,735
stories	242,343
authors	34,346
children	808,628
date range	2006-2022

3 DATA PREPARATION

3.1 Memeorandum

The subject of analysis is a set of news article headlines scraped from the news aggregation site Memeorandum for news stories from 2006 to 2022. Each news article has a title, author, description, publisher, publish date and url. All of these are non-numeric, except for the publication date which is ordinal. The site also has a concept of references, where a main, popular story may be covered by other sources. Using an archive of the website, each day's headlines were downloaded and parsed using python, then normalized and stored in sqlite database tables [2].

3.2 AllSides MediaBiasFactCheck

What remains after cleaning is approximately 240,000 headlines from 1,700 publishers, 34,000 authors over about 64,000 days 2.

3.3 Missing Data Policy

The only news headlines used in this study were those with an associated bias rating from either AllSides or MediaBiasFactCheck. This elimiated about 5300 publishers and 50,000 headlines, which are outlets publishing only less than 1 story per year. Another consideration was the relationship between the opinion and news sections of organizations. MediaBiasFactCheck makes a distinct between things like the Wall Street Journal's news organization, one it rates as 'Least Bias', and Wall Street Journal's opinion organization, one it rates as 'Right'. Due to the nature of the Memeorandum dataset, and the way that organizations design their url structure, this study was not able to parse the headlines into news, opinion, blogs or other sub-categories recognized by the bias datasets. As such, news and opinion was combined under the same bias rating, and the rating with the most articles published was taken as the default value. This might lead to organizations with large newsrooms to bias toward the center in the dataset.

4 EXPERIMENTS

4.1 Link Similarity Clustering and Classification

4.2 Title Sentiment Classification

for every title, tokenize, classify.

The classification of news titles into emotional categories was accomplished by using a pre-trained large language model from HuggingFace. This model was trained on a dataset curated and published by Google which manually classified a collection of 58,000 comments into 28 emotions. The classes for each article will be derived by tokenizing the title and running the model over the tokens, then grabbing the largest probability class from the output.

The data has been discretized into years. Additionally, the publishers will have been discretized based of either principle component analysis on link similarity or based on the bias ratings of All Sides. Given that the features of the dataset are sparse, it is not expected to have any useless attributes, unless the original hypothesis of a temporal trend proving to be false. Of the features used in the analysis, there are enough data points that null or missing values can safely be excluded.

No computational experiment have been done yet. Generating the tokenized text, the word embedding and the emotional sentiment analysis have made up the bulk of the work thus far. The bias ratings do not cover all publisher in the dataset, so the number of articles without a bias rating from their publisher will have to be calculated. If it is less than 30% of the articles, it might not make sense to use the bias ratings. The creation and reduction of the link graph with principle component analysis will need to be done to visualize the relationship between related publishers.

5 RESULTS

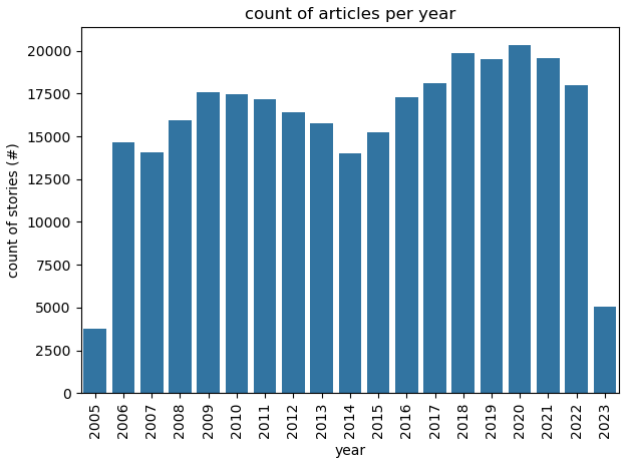

Figure 1: Articles per year.

ACKNOWLEDGMENTS

To Dr. Hearne, for the instruction on clustering and classification techniques, and to Pax Newman for the discussion on word embeddings.

REFERENCES

[1] Seth Flaxman, Sharad Goel, and Justin M. Rao. 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly* 80, S1 (2016), 298–320. https://doi.org/10.1093/poq/nfw006

[2] Matt Jensen. 2023. Data Mining 577: Political Polarization Data. https://data.publicmatt.com/national_news/stories

[3] Matt Jensen. 2023. Data Mining 577: Political Polarization Source Code. https://github.com/publicmatt/data_mining_577

[4] Markus Prior. 2013. Media and Political Polarization. *Annual Review of Political Science* 16, 1 (May 2013), 101–127. https://doi.org/10.1146/annurev-polisci-100711-135242

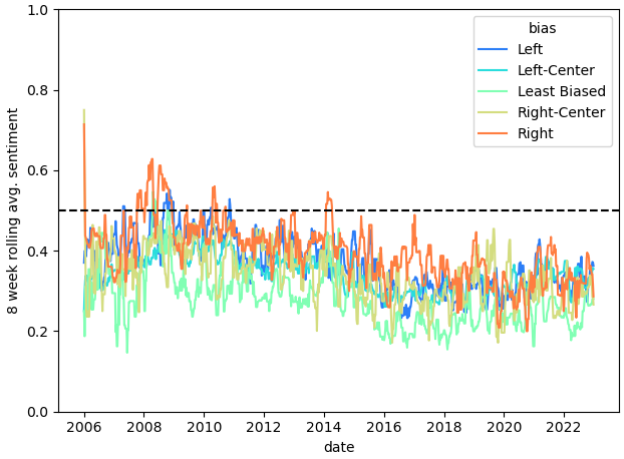

Figure 2: Sentiment vs. bias over time

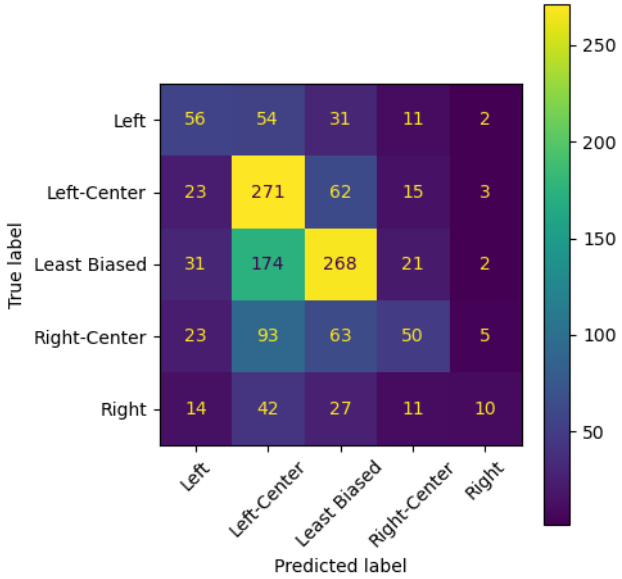

Figure 3: kNN confusion matrix of related links adjacency matrix

[5] Alexander J. Stewart, Nolan McCarty, and Joanna J. Bryson. 2020. Polarization under rising inequality and economic decline. *Science Advances* 6, 50 (Dec. 2020), eabd4201. https://doi.org/10.1126/sciadv.abd4201

A ONLINE RESOURCES

The source code for the study is available on GitHub [3].

Received 4 April 2023; revised 9 June 2023