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

There is evidence of increased political polarization in the United States over the past 16 years. Through voting patterns and self-reported political viewpoints and party affiliation, the political landscape has seen a 'hallowing out' of the middle in resent years. A common categorization of political leanings can be summerized as the spectrum from In simple terms, political beliefs in the United States can be categoried as left, center or right. This political spectrum largely reflects affiliation with one of the two dominate political parties on the federal level, left associated with Democrats and right associated with Republicans. To appeal to difference segments of this political affiliation, publishers of news might want to target content and coverage to issues relevant to only one side of the spectrum or the other. This phenomenon has led to concern over the creation of echo chambers where each side only consumes content made specifically to confirm their own beliefs. Driven by the market demand for confirmation bias, 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 side of the spectrum has been 'moving' more rapidly away from the 'middle'.

2 DATA SOURCES

All data was collected over the course of 2023 using python scripts, the source code for which is available on GitHub [3].

Table 1: News Dataset Sources

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

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

2.1 Memeorandum

The main subject of analysis is a set of news article headlines downloaded from the news aggregation site Memeorandum. The archive spans the years 2005 to 2023 and contains headlines from over 1,700 unique publishers 2. 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].

2.2 AllSides & MediaBiasFactCheck

The media bias ratings are sourced with permission from two media watchdog groups, AllSides and MediaBiasFactCheck. These sources aggregate expert opinion, crowdsourced data and Each source's objective is to assess the bias and factual reporting of media and information sources. Their claim to achieve this using methodologies that combine objective measures like use of primary sources, consistency of reporting accuracy and expert and crowdsourced opinion. Neither claim it realistic to maintain perfect objectivity, and openly admit determining bias is challenging. Both sources provide a categorical value of bias from left to right. It is important to note the bias scale is based on the political landscape of the United States, which differs from that of other countries. For instance, while the Democrats are regarded as centrist or right-center in many nations, they are considered left-center within the US. This bias rating can be convered to a zero centered quantitative measure, with center bias representing zero and left and right representing -2 and 2 respectively. In addition to bias, MediaBiasFactCheck provides a measure of trustworthiness of the publishers' source material. That measure is not used in this analysis, but could be an interesting feature to include in classification analysis in the future.

2.3 Word Embeddings & Sentiment

A common operation in natural language processing is to take text, which is not quantitative, and transform it into a vector representation. There are a couple of common ways to do to. Historically, an algorithm like term frequency–inverse document frequency (TFIDF) was used. Term frequency counts the number of times a word appears within a document relative to the size of the document. Inverse document frequency measures how common a word appears in the corpus. The product of these two values combines the local importance of a term with its global importance across the corpus. In contrast, bidirectional encoder representations from transformers (BERT) use deep learning to capture contextual meaning of words. It learns an embedding by training on a large amount of text data. For the task of sentiment, the text data is labeled by humans with a value from -1, meaning negative, to 1, meaning positive. In this way, the BERT model can be a sophisticated classifier of tokenized text. The labeling of news titles with emotional and sentiment categories was accomplished by using a pre-trained large language model from HuggingFace. The emotional component of this model was trained on a dataset curated and published by Google[1] 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.

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

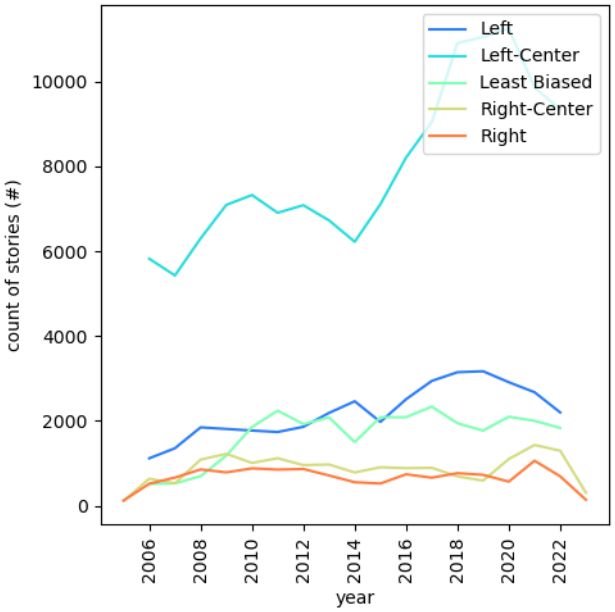

Figure 1: Articles per bias over time.

3 EXPERIMENTS

3.1 Link Similarity Clustering and Classification

The links between breaking news and coverage of that news by other sources can be thought of as a short and wide tree data structure. Or more generally, each headline and linking article are nodes in a directed graph with a single edge pointing from the coverage to the headline. The stories were already in an adjacency list representation, so it was easy to convert it to a adjacency matrix. This matrix then had a row for every publisher with at least one breaking news item, and a column for all publishers in the dataset. The value of the edge from child to parent took on the form of one of three options: onehot encoding, total references and normalized references. The one hot encoding scheme was the simpliest: if a link exists between child and parent, put a one, otherwise, put a zero. The total references scheme was similar, but the references were summed, so each cell contained the sum of all links between parent and child. The normalized scheme extended the total references scheme by dividing each cell by the sum of references across each row, so that each row would sum to one. The result was three matricies, with The creation and reduction of the link graph with principle component analysis will need to be done to visualize the relationship between related publishers.

3.2 Title Sentiment Classification

Of the features used in the analysis, there are enough data points that null or missing values can safely be excluded. The bias ratings do not cover all publisher in the dataset, but there are enough remaining labels to make classification an interesting task.

for every title, tokenize, classify.

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.

4 RESULTS

4.1 Link Similarity Clustering and Classification

Elbow Method. To determine the optimal number of clusters to use for analysis of the link similarity experiments, an plot of squared distances between centroids vs. bin size k was used. Commonly called an 'elbow plot', it helps find a point of diminishing returns, where adding more clusters does not significantly improve the quality of clustering. This analysis heuristically reveals the optimal number of clusters to be in the range of 4 to 7 2.

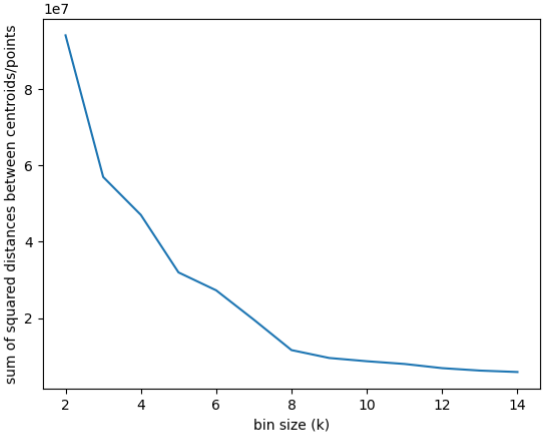

Figure 2: Elbow Criterion for KMeans

The idea behind the elbow plot is to find a balance between having too few or too many clusters. Too few results in under-segmentation, while too many clusters may lead to over-segmentation and less interpretable results.

Clustering. All three encoding schemes (onehot, links quantity, normalized) were ran through a KMeans clustering with $k = 5$. To visualize the results, a principle component analysis (PCA) was performed on the input. PCA is a technique for dimensionality reduction and data visualization. It transforms correlated variables into a smaller set of uncorrelated variables, retaining as much information as possible from the original data. The first two principal components represent the most important patterns or structures in the data. Plotting these two components against the predicted labels of the KMeans clustering algorithm gives us a good gauge on the accuracy of the clustering 5 4 3.

Clustering on link quantity leads to most publishers being labeled with a single class 3.

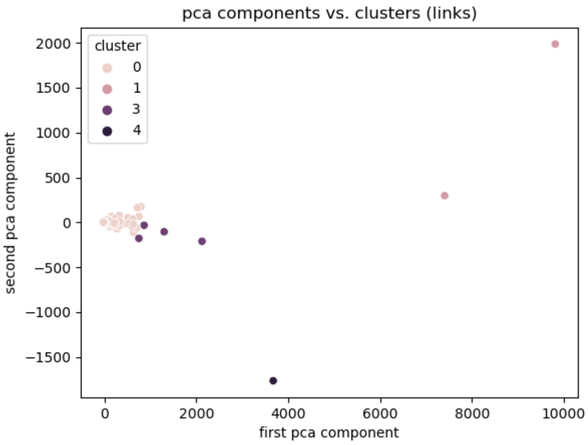

Figure 3: PCA components vs. KMeans Clusters (Links)

The problem of all publishers being labeled a single label is lessed when the link quantity is normalized 4. It was initially thought that the quantity of the links, or the frequency of stories published, would dominate the clustering and the PCA analysis, but less prolific publishers are present in the smaller classes instead.

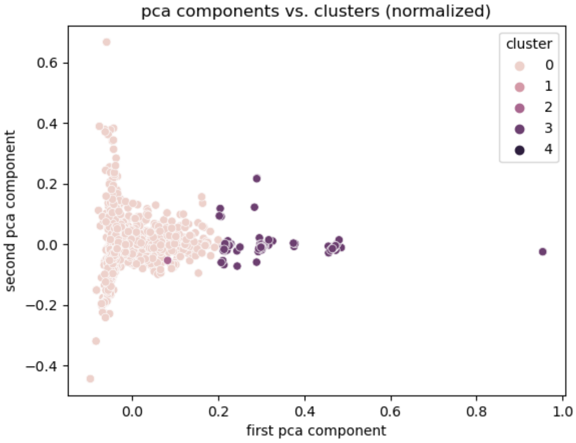

Figure 4: PCA components vs. KMeans Clusters (Normalized)

The classes are more evenly distributed when a one hot encoding is used to generate the KMean clustering 4.

Classification. A k-neighest neighbors (kNN) classification algorithm was trained on the adjacency matrix, the one hot encoding was used for all the following experiements. To test the classification accuracy, the link dataset was separated into a training and a test set with a ratio of 80 : 20 respectively.

To visualize the results, a confusion matrix was generated, plotting the true values of the hold out data against what the kNN model predicts 6.

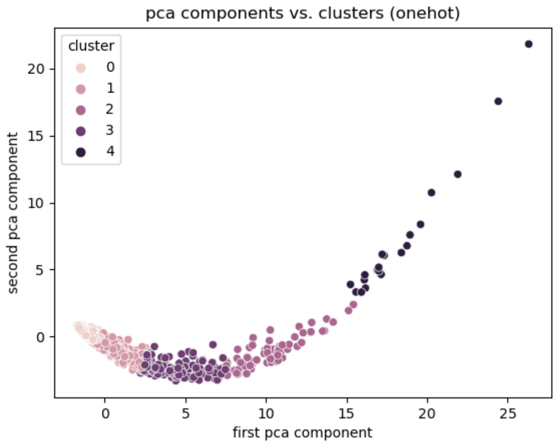

Figure 5: PCA components vs. KMeans Clusters (One Hot)

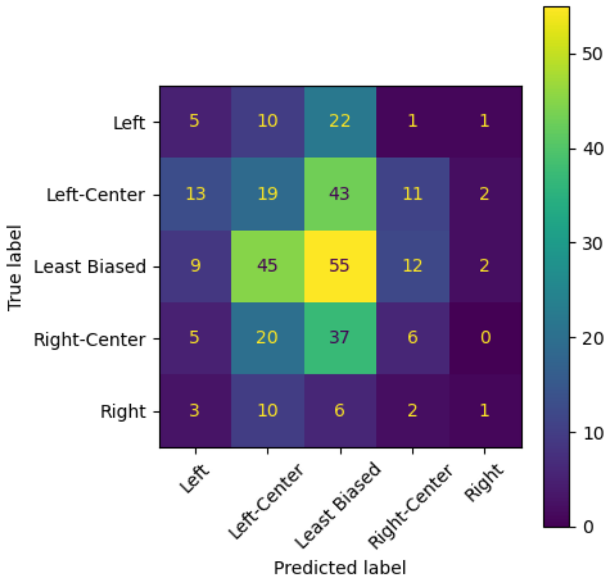

Figure 6: kNN confusion matrix of related links adjacency matrix

Generally, higher numbers along the major diagonal means, where the correct label is the same as the predicted label, is most accurate. Additionally, the distribution of values is relatively balanced, indicating the model is performing consistently across different classes without much skew toward one particular classification.

Overall, the linking between publishers around the same headlines serves as a good feature set for predicting the bias class of a particular publisher. A model such as the kNN, trained on this data, could be used to classify unseen publishers based on where they get their sources from.

4.2 Title Sentiment Classification

Tracking sentiment over time was very straight forward once classes were extracted from the pre-trained model 7. In general, the trend was toward the sentiment of news headlines to decrease from positive to negative over time. The only point over the extend of this dataset where the sentiment was positive (> 0.5), for any publication on the political spectrum, was short periods around 2008 and 2011.

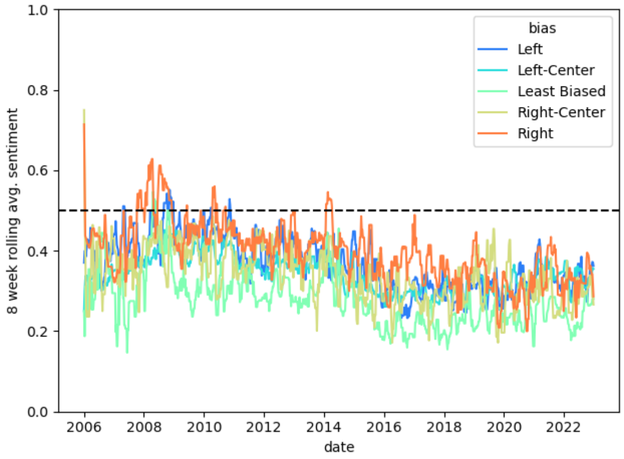

Figure 7: Sentiment vs. bias over time

Emotion. In addition to sentiment, ranging from 0 to 1, another BERT based classifier was used to extract the emotional tone of the headlines. The dominate emotion by far was a neutral one 8

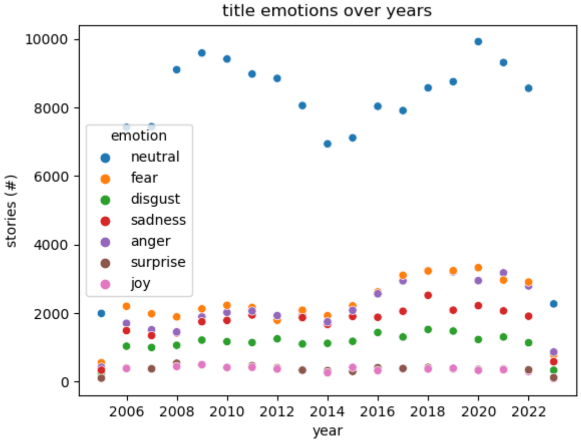

Figure 8: Emotion Tone Frequency Over Time

In general, the emotional tone of headlines have not seen a significant amount of change over time. A regression model was fit to the average emotional content per week. The extreme ends of the political spectrum actually saw the highest increase in neutral titles overtime according to the slope of the regression model 9. The publishers labeled as the least bias saw the highest decrease in neutral emotion expressed.

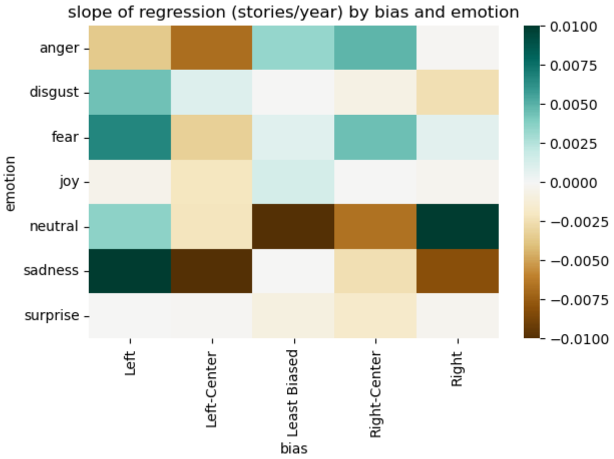

Figure 9: Slope of Regression (stores/year) by Bias and Emotion

5 DISCUSSION

We performed link similarity clustering using three encoding schemes: one-hot encoding, total references, and normalized references. We applied KMeans clustering and visualized the results using PCA. We found one-hot encoding scheme led to more evenly distributed clusters, while the link quantity scheme resulted in most publishers being labeled with a single class. The normalized scheme reduced the issue of all publishers being labeled with a single class but not enough to use in subsequent studies.

For classification, a kNN model was trained on the link adjacency matrix using the one-hot encoding scheme. The accuracy of the classification was evaluated using a confusion matrix, which showed consistent performance across different classes.

We used a pre-trained BERT model to classify the sentiment and emotionaly content of news headlines. We observed a trend towards decreasing sentiment from positive to negative over time. They found that the dominant emotion was neutral throughout the dataset.

Overall, the results of the study indicate a polarization in news headlines over time, with a shift towards more negative sentiment. This study explores the impact of political polarization on the emotional characteristics of news headlines and the link similarity of publication with the same political bias.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. arXiv:2005.00547 [cs]

[2] Matt Jensen. 2023. Data Mining 577: Political Polarization Data.

[3] Matt Jensen. 2023. Data Mining 577: Political Polarization Source Code.

[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

[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

Received 4 April 2023; revised 9 June 2023