Data Mining CS 571

Matt Jensen

2023-04-25

Abstract

News organizations have been repeatedly accused of being partisan. Additionally, they have been accused of polarizing dicussion 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.

1 Data Preparation

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, url and related discussions. The site also has a concept of references, where a main, popular story may be covered by other sources. This link association might be used to support one or more of the hypothesis of the main analysis. After scraping the site, the data will need to be deduplicated and normalized to minimize storage costs and processing errors. What remains after these cleaning steps is approximitely 6,400 days of material, 300,000 distinct headlines from 21,000 publishers and 34,000 authors used in the study.

2 Missing Data Policy

The largest data policy that will have to be dealt with is news organizations that share the same parent company, but might have slightly different names. Wall Street Journal news is drastically different than their opinion section. Other organizations have slightly different names for the same thing and a product of the aggregation service and not due to any real difference. Luckily, most of the anaylsis is operating on the content of the news headlines, which do not suffer from this data impurity.

3 Classification Task

The classification of news titles into emotional categories was accomplished

by using a pretrained large langauge 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 probabilty 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.

4 Experiments

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

TODO.