Data Mining - CSCI 577

Project Status Report I

2023-04-04

Participants

Matt Jensen

Overarching Purpose

I hope to use a dataset of new articles to track the polarization of news over time. I have a hypothesis that news has become more polarized superficially, but has actually converged into only two dominate views points. I think there is a connection to be made to other statistics, like voting polarity in congress, or income inequality, or consolidation of media into the hands of the few.

Data Source

To test this thesis, I will crawl the archives of memeorandum.com for news stories from 2006 onward. I will grab the title, author, publisher, published date, url and related discussions and store it in a .csv. The site also has a concept of references, where a main, popular story may be covered by other sources. So there is a concept of link similarity that could be explored in this analysis too.

Techniques

I am unsure of which technique specifically will work best, but I believe an unsupervised clustering algorithm will serve me well. I think there is a way to test the ideal number of clusters should exist to minimize the error. This could be a good proxy for how many 'viewpoints' are allowed in 'mainstream' news media.

Project Status Report II

2023-04-11

Participants

Matt Jensen

Dataset Description

The dataset I will be using for my analysis has the following attributes:

- title
  - a text description of the news item.
  - discrete, nominal.
  - ~800k distinct titles.
- url
  - a text description and unique identifier for the news item.
  - discrete, nominal.
  - ~700k distinct urls.
- author
  - a text name.
  - discrete, nominal.
  - ~42k distinct authors.
- publisher
  - a text name.
  - discrete, nominal.
  - ~13k distinct outlets.
- related links
  - an adjacency matrix with the number of common links between two publishers.
  - continuous, ratio.
  - counts are less than total number of stories, obviously.
- published date
  - the date the article was published.
  - continuous, interval.
  - ~5.5k distinct dates.

In addition, I will augment the data with the following attributes:

- title word embedding
  - a vectorized form of the title from the output of a LLM or BERT model which embeds semantic meaning into the sentence.
  - continuous, nominal.
  - 800k vectors, of 768 values.
- political bias of the publisher
  - a measure of how voters feel the political leanings of the publisher map to the political parties (Democrat/Republican).

  - continuous, ordinal.
  - ~30% of the publishers are labelled in allsides.com ratings.
- estimated viewership of the publisher
  - an estimate of the size of the audience that consumes the publisher's media.
  - continous, ratio.
  - I still need to parse The Future of Media Project data to get a good idea of this number.
- number of broken links
  - I will navigate all the links and count the number of 200, 301 and 404 status codes return.
  - discrete, nominal
  - size of this dataset is still unknown.

Purpose

I want to analyze data from the news aggregation site memeorandum.com and combine it with media bias measurements from allsides.com. My goal for the project is to cluster the data based on the word embeddings of the titles. I will tokenize each title, and use a BERT style model to generate word embeddings from the token.

Word embedding output from language models encode semantic meaning of sentences. Specifically, BERT models output embeddings of 768 dimensional space. Clustering these vectors will map from this semantic space to a lower dimensional cluster space.

My understanding of cluster leads me to believe that this lower dimensional space encodes meaning like similarity. In this way, I hope to find outlets that tend to publish similar stories and group them together. I would guess that this lower dimensional space will reflect story quantity and political leanings. I would expect new outlets with similar quantity of stories and political leanings to be grouped together. Another goal is to look at the political alignment over time. I will train a classifier to predict political bias based on the word embeddings as well. There is a concept of the Overton Window and I would be curious to know if title of new articles could be a proxy for the location of the overton window over time.