News ETA: Report

Becky Desrosiers | rn7ena@virginia.edu | DS5001 F24 | 13 December, 2024

All throughout this class, I have learned about tools and methods that can take a long-form text or corpus of texts and turn it into a data model that tells a story. I played with toy examples for 12 weeks, and then finally had the chance to put everything together into this investigation. I found results that excited me because they matched up with what I know, results that surprised me because they didn't match what I thought, and some results that disappointed me because they were inconclusive or not very useful. In all, I think I got the full experience of a project in Exploratory Text Analytics (ETA)! Let me share it with you.

I used a dataset provided by the course materials, which was a collection of news article snippets from 16 different sources, ranging from CNN to Fox. The publish dates range from November 05, 2013 to February 27, 2020, although there were significantly less documents from before 2017 and data from before November 2016 were very sparse (Appendix A). There were over 1 million documents, but for feasibility reasons I took a random sample of 10,000 documents with random seed 5001. The sources were biased towards US News, which accounted for about 50% of the documents. When accounting for the political leaning of each source, they were predominantly liberal, and conservative sources were underrepresented (Appendix B). The results of my analysis may be skewed because of the imbalance of the dataset, and estimates on conservative sources may be less accurate because of the underrepresentation. Even so, there were almost 800 conservative documents in the corpus: enough to do analysis on, in my opinion.

I started by processing the raw text into a Standard Text Analytic Data Model with annotations, vector space models, and analytical models. The most exciting result, in my opinion, was the topic modeling. I used Latent Dirichlet Allocation (LDA) to categorize the tokens into topics, then summed the tokens for each document to find the overall topics addressed by each document. The topics, described by the top 3 words, turned out to be generally intelligible and seperable. Overall, the news sources focused a lot on what I deemed police-related stories (Appendix C). The top three words in this topic were man/police/authorities.

To explore further, I compared topic coverage based on political party and news source (Appendix D). I found that my model indicated conservatives focus more on politics and crime, whereas liberal sources focus more on wellbeing-related matters like healthcare and the environment, which seems on par with my experience. I was excited to see that my topic model matched up with my own perception of news sources! Another interesting trend noticeable in the heatmap in Appendix D is the horizontal lighter-colored bars corresponding to national politics and state politics. This indicates that many news sources cover more politics than other topics. Interestingly, the Drudge Report covered a large majority of the topic I dubbed "general" – the top 3 words were story/link/stories, and it was difficult to determine what that topic was actually about.

I investigated the 5 topics with the highest change in coverage frequency over time (Appendix E), because it was too hard to read if all 15 topics were plotted. The most interesting observations included that coverage of police-related news increased at the beginning of the period and stayed high throughout, and that coverage of the environment saw an increase in 2018. Both of these correspond to increased awareness of police brutality and climate change that I have witnessed in the past decade.

I was more surprised by the results from my sentiment modeling (Appendix F). I used word sentiment scores from the NRC Word-Emotion Association Lexicon (aka EmoLex) to assign sentiment values to terms, summed the values for every token in each document, then segmented the documents by source and political leaning of the source. Sentiment scores across sources/political leanings were normalized so that bar charts would not be dominated by liberal sentiment scores. I expected more negative and fearful language from conservatives, and more hopeful discourse from liberals, possibly represented by higher trust scores and positive polarity. It turned out, however, that while conservative, liberal, and center sources all demonstrated the same amount of positive word usage, but documents from liberal and center sources also tended to include more negatively-scored words. Overall, the polarity score of conservatives was more than double that of liberals, and over 6 times that of the center sources. Documents from conservative sources demonstrated notably more trust, and less sadness, anger and fear than liberal and center sources. Even so, overall trust was the most prominent emotion for all three political leanings and most sources in general. Polarity and

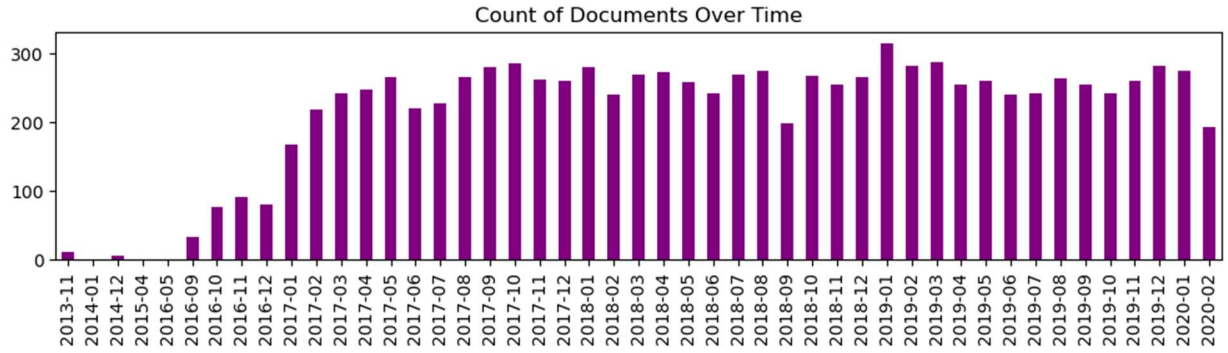
emotions over time did not show very interesting trends, so I will now move on to my most disappointing results.

My results from principle component analysis (PCA) were not very conclusive or interesting, in my opinion. I used the top 50% of words (proper nouns excluded) by TFIDF sum to create my covariance matrix, which was factored via eigendecomposition to produce my PC loadings and explained variance. The first principle component (PC0) explained less than 3.5% of the variation in the dataset, and the first two combined explained only just over 5% of the variance, so it was difficult to map the data onto lower dimensions using principle components (Appendix G).

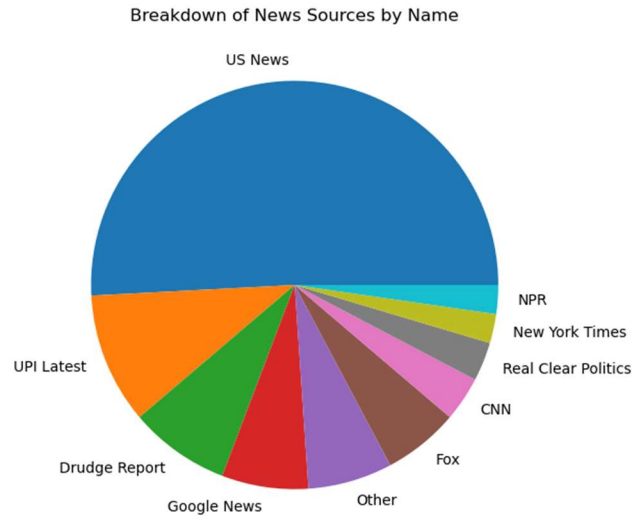
The table in Appendix G shows the 10 most extreme words on PC0 and PC1 on both the positive and negative sides. There does not seem to be an obvious latent feature that either of these axes deal with. I might venture to say that PC0 could be a scale from official to personal, where the highest loadings correspond to words like vote, trial, and third, whereas the negative loadings correspond to more personal words like family, beating, and crashed. Even so, this interpretation seems a little arbitrary and is not fully supported by all the words in the top/bottom 10. It is even more difficult to find any significant trend in the PC1 loadings. One thing to note is that all of the words hug the axes of PC0 and PC1 in the scatterplot in Appendix G. There are no points which have high magnitudes on both axes. I thought this was interesting, though I do not know exactly what it signifies.

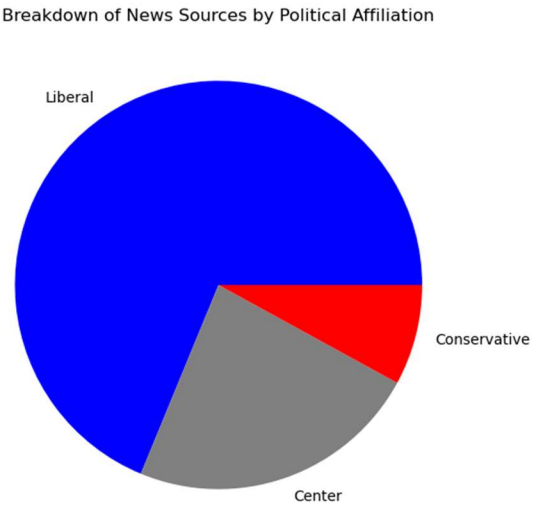
In conclusion, I had mixed results from my analysis, which I believe is par for the course. ETA, along with much of machine learning, is as much an art as it is a science, and requires some massaging to get useful results. Moving forward, I would try investigating different numbers of topics to possibly break up the 60% “general” topic that Drudge Report covers and testing principle components with an updated vocabulary – perhaps a larger or smaller proportion of top-TFIDF-sum words. Another future step would be to repeat my analysis with more current text, as the most recent document in the corpus I used is from February 2020. From this project, I gained a deeper understanding of the mechanics of ETA methods and a deeper appreciation for the results they can produce.

Appendix A

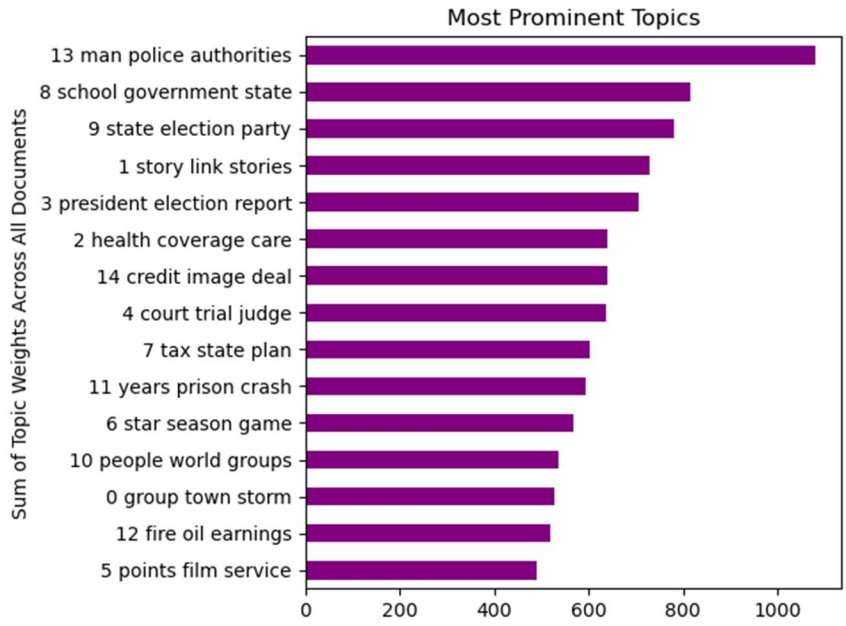


Appendix B: Breakdown of News Sources

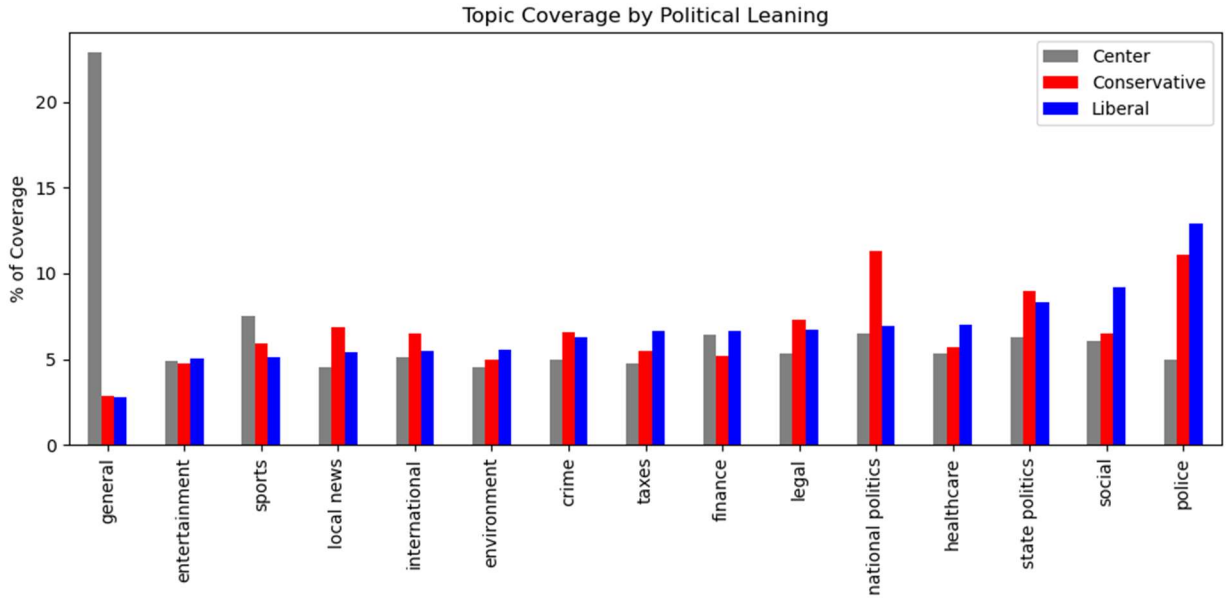


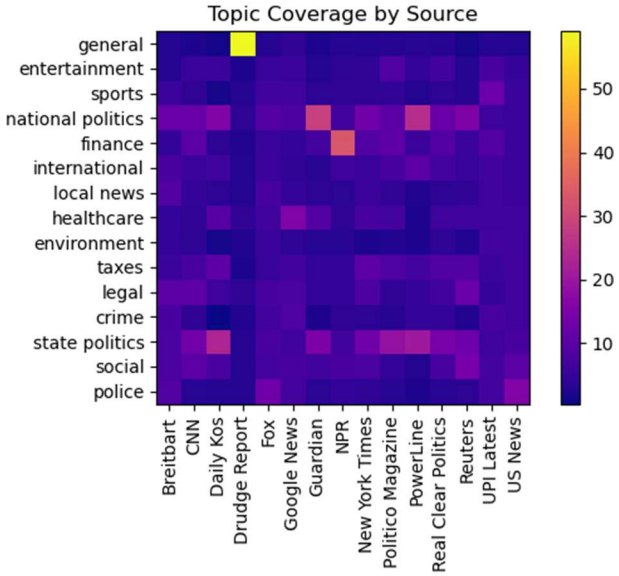


Appendix C

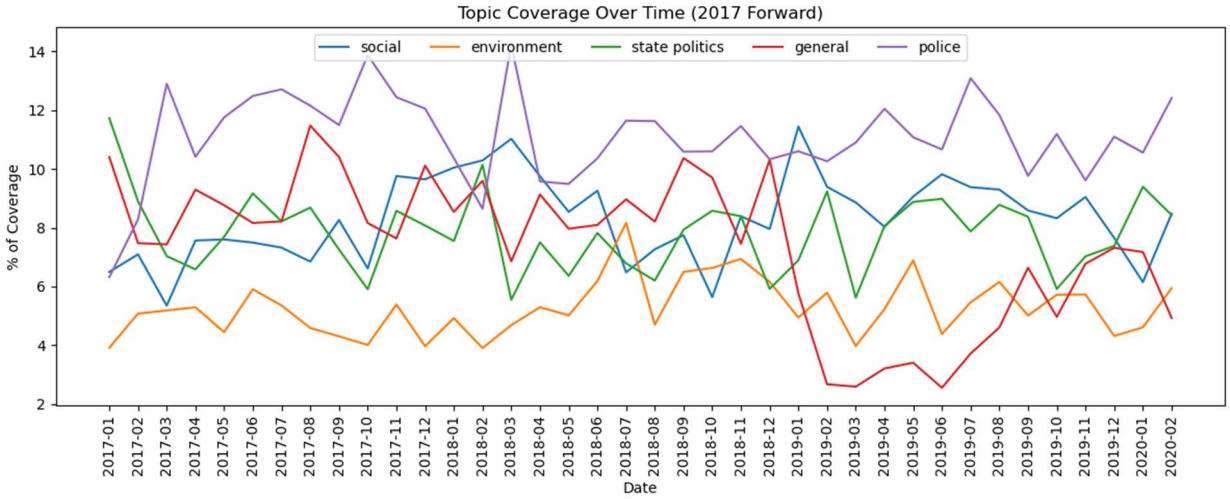

Most Prominent Topics
Sum of Topic Weights Across All Documents
13 man police authorities
8 school government state
9 state election party
1 story link stories
3 president election report
2 health coverage care
14 credit image deal
4 court trial judge
7 tax state plan
11 years prison crash
6 star season game
10 people world groups
0 group town storm
12 fire oil earnings
5 points film service
0
200
400
600
800
1000

Appendix D: Breakdown of Topic Coverage

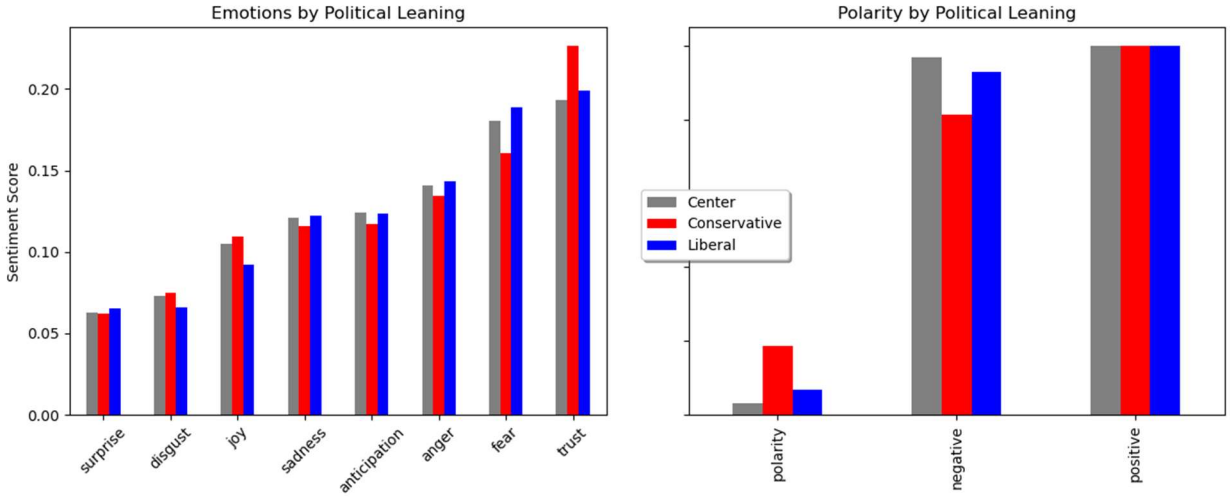




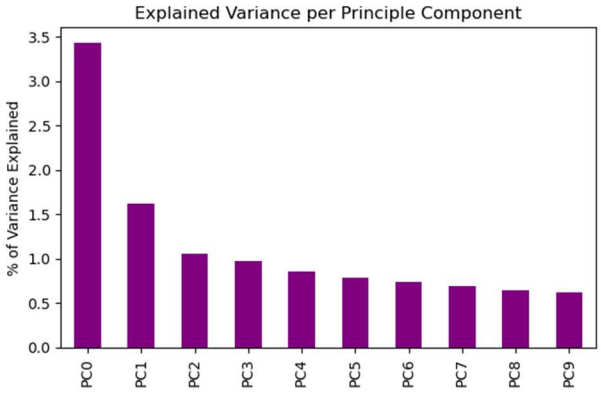
Appendix E

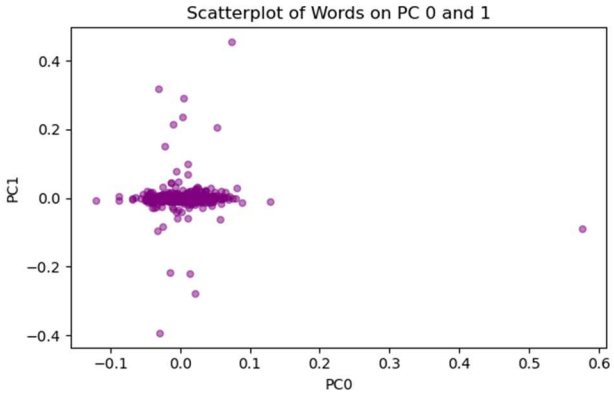


Appendix F: Sentiment by Political Leaning



Appendix G





top 10 PC0		bottom 10 PC0		top 10 PC1		bottom 10 PC1	
PC0_loading		PC0_loading		PC1_loading		PC1_loading	
0.068153	ago	-0.121335	official	0.069276	woman	-0.392926	one
0.069128	vote	-0.088456	family	0.079242	new	-0.279345	state
0.071204	taking	-0.087647	work	0.100374	year	-0.220157	killed
0.073617	says	-0.068767	beating	0.149858	years	-0.217544	police
0.075332	making	-0.068489	told	0.206051	earnings	-0.094152	--
0.080174	trial	-0.064697	despite	0.213628	say	-0.088425	-
0.080959	home	-0.055895	crashed	0.234944	people	-0.083090	column
0.089237	third	-0.053326	back	0.290048	stories	-0.062242	week
0.129641	month	-0.050024	sentenced	0.318191	said	-0.060318	man
0.575931	-	-0.049113	weeks	0.454605	says	-0.058712	full