Cable Talk Show Language Comparison

By Jennifer Shumway, Rachel Miller, Riley Schenck

Motivation

Cable talk shows provide a platform for the host and their guests to establish issues as important to a society and serve as a forum for discussion of existing issues. Through this work, we explore the language used by cable talk show hosts Rachel Maddow and Tucker Carlson as it relates to issues deemed important by voters through polling carried out by PEW Research Center.

We delve into four NLP techniques with a focus on topic modeling in order to explore the advantages and shortcomings of these approaches. Through this work we expand on previous work completed by other authors[1,2] by making an assessment not only of how how well each talk show mirrors what potential viewers rank amongst the top issues relevant to them but also delving into the technique itself.

Techniques:

Word Frequency
Limited Dictionary Classification
Cosine Similarity with Word Embeddings (GoogleNewsVector)
BERT Topic Modeling

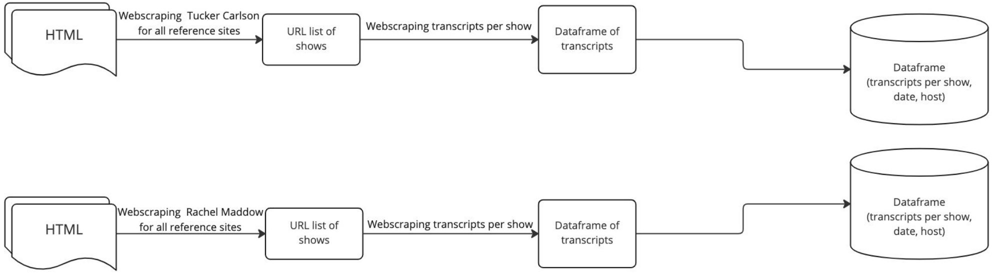
Questions we explore:

What are the topics of interest per host as determined by a given technique?
What are the advantages and disadvantages of each technique?
Based on any given technique, how do the topics tie back (or not) to the PEW polling data?

1. https://tvnews.stanford.edu/data
2. https://www.newswise.com/politics/study-compares-fox-news-and-msnbc-using-52-000-transcripts-283-million-words/?article_id=755937

Methodology

Data Sources



HTML

Webscraping Tucker Carlson for all reference sites

URL list of shows

Webscraping transcripts per show

Dataframe of transcripts

Dataframe (transcripts per show, date, host)

HTML

Webscraping Rachel Maddow for all reference sites

URL list of shows

Webscraping transcripts per show

Dataframe of transcripts

Dataframe (transcripts per show, date, host)

The following data sources were considered for this report:

Pew issue polling data

This dataset is created and curated by Pew Research Center. In their own words "Pew Research Center is a nonpartisan, non advocacy fact tank that informs the public about the issues, attitudes and trends shaping the world. It does not take policy positions. The Center conducts public opinion polling, demographic research, computational social science research and other data-driven research."

We extract 3 data sets which reflect polling taken throughout 2022 (March, August and October). Each data set provides the percentage of respondents (stored as a float) who say a particular issue (captured as a string) is 'VERY' important to their vote broken down by democrat and republican.

https://www.pewresearch.org/politics/2022/08/23/abortion-rises-in-importance-as-a-voting-issue-driven-by-democrats/

https://www.pewresearch.org/politics/2022/10/20/midterm-voting-intentions-are-divided-economic-gloom-persists/

https://www.pewresearch.org/politics/2022/03/24/republicans-more-likely-than-democrats-to-say-partisan-control-of-congress-really-matters/

Cable news show transcripts

Rachel Maddow

https://www.msnbc.com/transcripts/show/rachel-maddow-show

Tucker Carlson

https://www.foxnews.com/category/shows/tucker-carlson-tonight/transcript

This dataset is created from transcripts from both the Rachel Maddow and Tucker Carlson show covering all transcripts between January 2022 and September 2022. Totaling approximately 160 transcripts per host. Each show url provides a full written transcript via a web interface that can be accessed via webscraping and pulled into python and stored as '.tsv'. Through this webscraping approach we are able to capture transcript, date, host and URL all represented as strings, for further analysis.

Notebooks: 1,2 Data Files: 9,10

Methodology

Data Cleaning and Manipulation Methods

Rachel Maddow and Tucker Carlson Transcripts:

Data Cleaning Needs	**Data Cleaning Solution**
Data Types	Converting date string to datetime object including custom function to remove updated date
Duplicate Transcripts	Removing duplicate entries based on timestamp (single show produced per day)
Data Shaping	Use Split-Apply-Combine to transform and reshape data Melt data into long form to aid with creating visualizations
Text processing	Lowercasing - in order to ignore capitalization for techniques such as word frequency Tokenization - breaking text into tokens (words for our use case) Stop word removal - removing common stop words based on NLTK. Removing additional filler words relevant per host Lemmatization - Reducing each word to a single base form ie. running, ran, runs → run

PEW Issue Polling:

This dataset is very clean and organized and requires minimal cleaning/manipulation as it provides consistent labeling, data types and we pull it manually from the PEW research documents . Where there are missing values for a few polling questions, we simply map these missing values to None and choose not to interpolate between existing polling data points.

In order to explore a variety of NLP techniques for topic modeling we also add a dictionary of words per issue. This dictionary evolved through several iterations and has both advantages and disadvantages that we will discuss within our analysis.

Notebook: 3 Data Files: 11,12

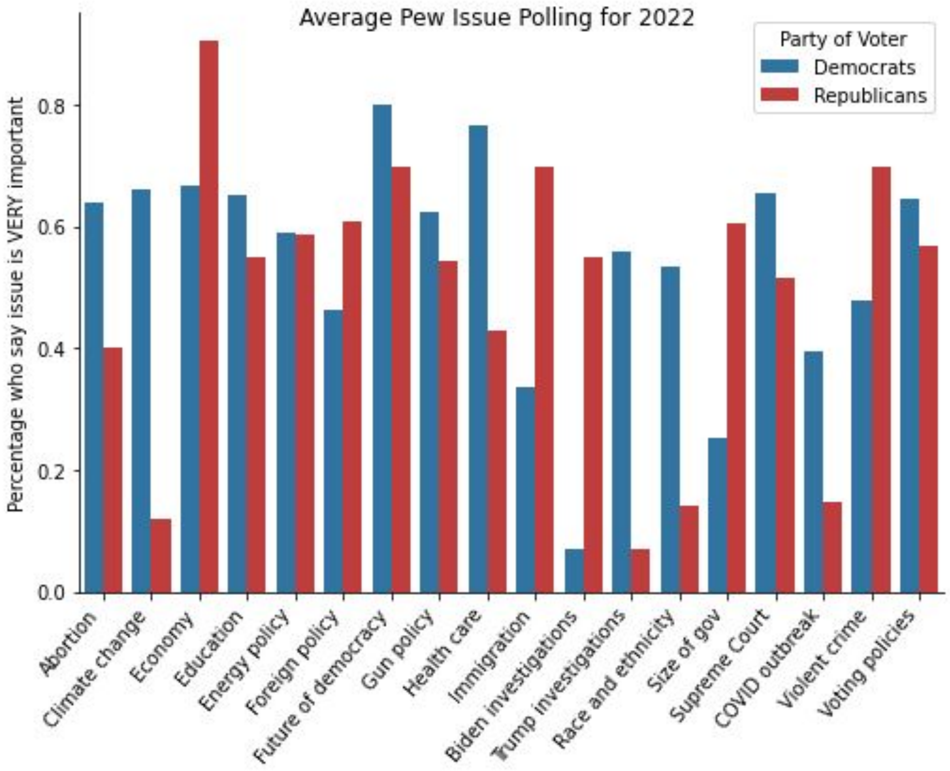
Analysis

PEW Issue Polling

PEW issue polling captures the percentage of respondents who say a particular issue is ‘VERY’ important to their vote. Issues which ranked consistently the highest among democrats throughout each polling period in 2022 was healthcare with 74%, 77% and 79% in March, August and October respectively. Issues related to the economy ranked most important overall amongst Republican voters with 90%, 90% and 92% over the same periods.

Issues where democrats showed little relative interest included size and scope of the federal government and immigration. For Republicans, the topics of limited interest included climate change, issues around race and ethnicity and the coronavirus outbreak.

Throughout the remainder of our analysis we seek to understand how four different approaches to topic modeling perform at capturing these topics and if the topics observed reflect back to the topics potential viewers find most important.



Notebook: 4 Data Files: 14

Analysis

Transcript Word Frequency

Approach: Firstly, we evaluated the language used by both talk show hosts based on word frequency. Here we visualize the top 10 most frequent words used per host over time across their respective transcripts in 2022.

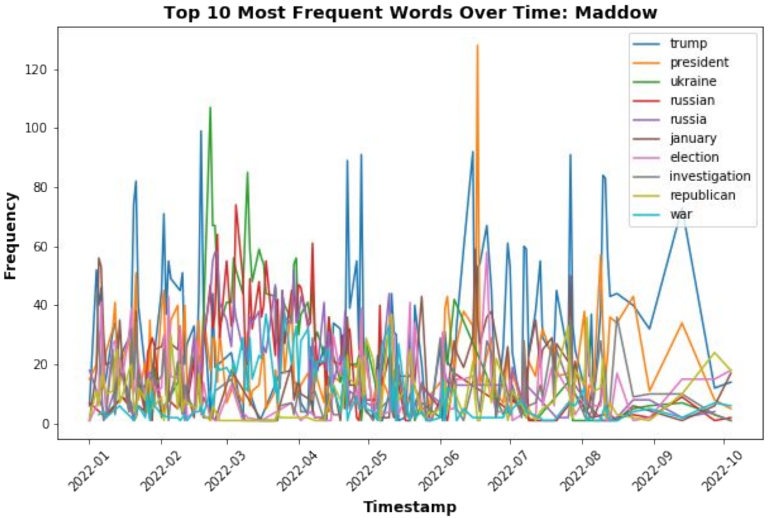
Advantages: Although simplistic, word frequency says a lot about the topics deemed most important by each host. For example, ranking highest for both Maddow and Carlson is references to a particular president ('Trump' in the case of Maddow and 'Biden' in the case of Carlson). Each host holds strong opinions towards the opposing parties leader and comes up in many episodes over time.

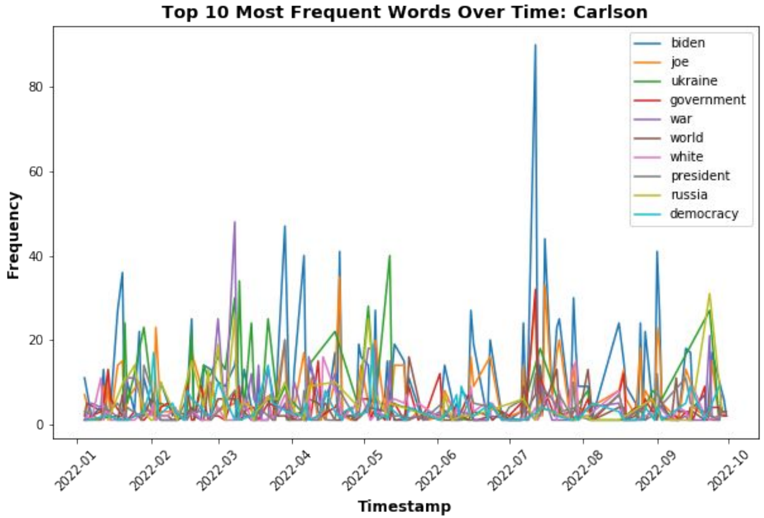
We are also able to extract other topics that both hosts find meaningful. Specifically issues towards Ukraine and Russia appear as a highly frequent word in both cases. This certainly ties back to the prominence of foreign policy and the economy (relating to the impact of russian oil on the global economy) as important issues for both parties as determined by the PEW issue polling.

Disadvantages: Beyond these high level topic extractions word frequency provides a limited view into how all the PEW issue topics are developing over time.

Future Considerations: We would recommend to consider capturing frequency of relevant words per topic over time compared to a baseline of total most frequent words over time. This additional analysis can make word frequency a more robust method of topic modeling.

Notebook: 5 Data Files: 11,12



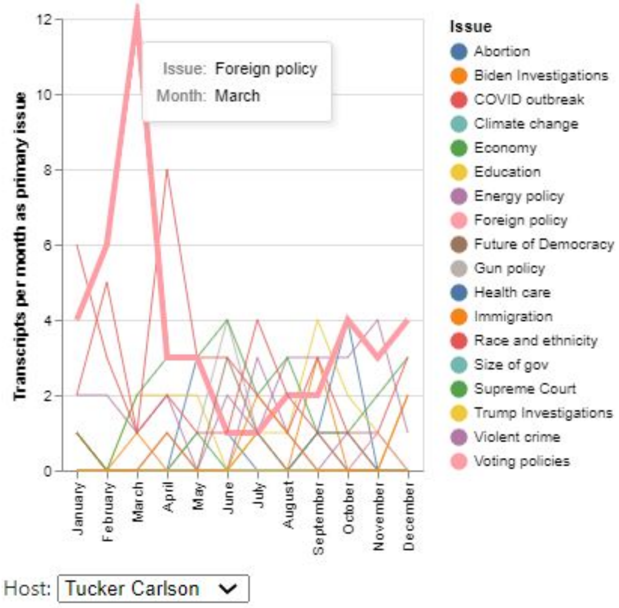


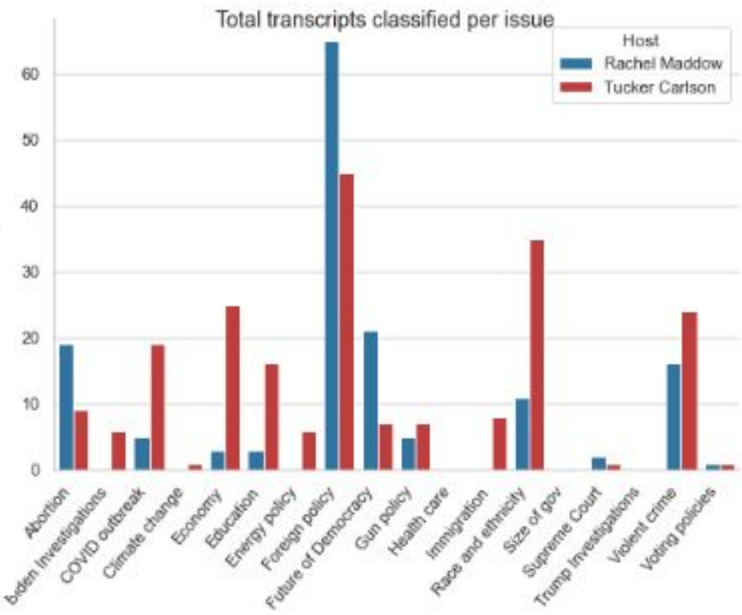
Analysis

Limited Dictionary Classification

Approach:

Here we classify each transcript based on the frequency of words and phrases found within issue-specific limited dictionaries. The dictionaries are based on our knowledge of issue-specific vocabularies as astute political observers. For example, “southern border” counts as a mention towards the “Immigration” issue, whereas “Putin” counts as a “Foreign Policy” mention. Transcripts are then assigned the issue that has the most total mentions from the different limited dictionaries.





Advantages:

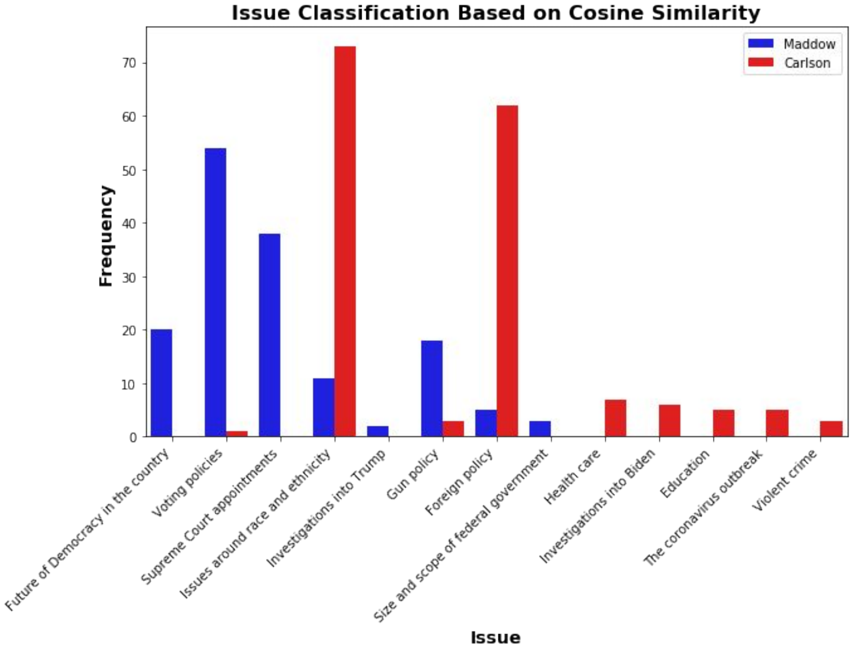
The method’s simplicity makes it easy to interpret. We can understand its strengths and weaknesses without being math majors. Unsurprisingly, this method appears to do a better job of comparing between hosts than between issues since the vocabulary used when speaking about issues varies much more between issues than it does between hosts. For example, Maddow as expected has more transcripts classified about abortion and the future of democracy compared to Carlson, while Carlson has more about issues surrounding race and ethnicity and crime.

Disadvantages: The limited dictionaries are subject to our own biases and understanding about which words and phrases are used by both hosts to talk about each issue, and those selection choices will influence the frequency of the issues mentioned within each transcript.

Notebook: 6 Data Files: 11,12

Analysis

Cosine Similarity with Google News Vectors (1)



Approach: To build on the previous approaches, we explored the pre-trained word embedding model GoogleNews-vectors-negative300.bin. We determined a vector representation per transcript by weighting each word (non-stop word) in a given transcript by it's TF-IDF score and then take the average of these weighted words per transcript. We then established a vector representation for each issue in the PEW data by taking the simple average of each word vector in the primary dictionary that aims to define such an issue. With both the transcript vector representation and the PEW issue vector representation in hand, we take the cosine similarity for each issue/transcript combination and finally take the argmax() to perform a simple classification of each transcript to one issue. Visualized here is the frequency of each issue classification for both Maddow and Carlson. Based on this classification approach, Carlson primarily covers topics related to 'race and ethnicity' and Maddow focuses mostly on issues related to the 'voting policies'.

Advantages: Google News Vectors provide a large vocabulary of over 3 million words and phrases with high quality word representations. By utilizing word embeddings and the cosine similarity approach we are able to capture words that are semantically similar to those in in the primary dictionary and expand on any approach that requires exact matching. Comparing our results here to what potential viewers find most important according to the PEW data we see that Tucker Carlson over weighs on issues towards race and ethnicity compared to the low issue rating it received from Republican poll participants. Issues covered by Rachel Maddow reasonably reflect issues Democrats found important to their vote.

Notebook: 6 Data Files: 11,12

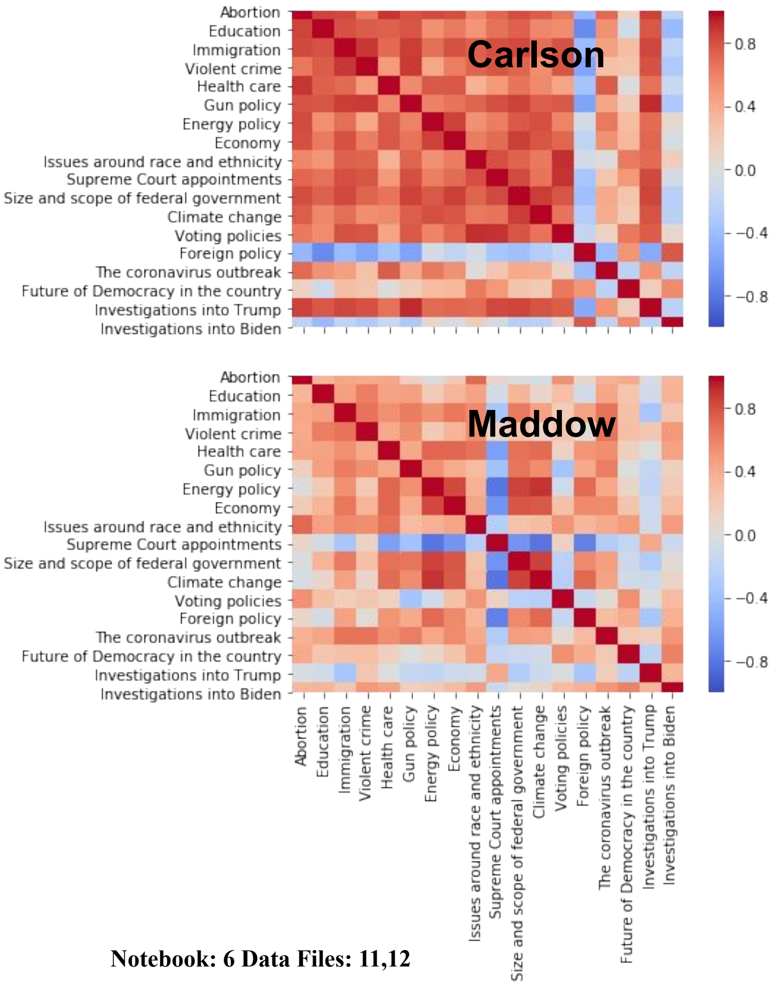
Analysis

Cosine Similarity with Google News Vectors (2)

Disadvantages: While word embedding models are a powerful tool, our classification approach based on the highest cosine similarity to a given pew issue vector proves to be limited. Visualized here is the correlation heatmap between issues for both Carlson and Maddow respectively. We observe that most issues within Carlson transcripts are highly correlated and Maddow transcripts show a larger degree of variation. This data would indicate that Carlson often covers a broad array of topics in each of his broadcasts resulting in a high correlation across many topics while Maddow tends to have more focused shows covering a more limited range of issues. However, we would expect the opposite as Carlson's transcripts are his opening monologues and not his entire show whereas Maddow's transcripts cover her full broadcast.

Like the limited dictionary, a significant downfall of this approach is the dependency on the PEW issue dictionary to properly capture language used to describe an issue with limited overlap between issues. The PEW issue dictionary provides a reasonable first approximation of language per issue but each issue dictionary is certainly not orthogonal.

Future Considerations: To expand on this approach we would recommend to invest further time and effort towards defining an issue dictionary that appropriately captures the issue at hand as deemed by an unbiased 3rd party, while also minimizing overlap within each dictionary.



Notebook: 6 Data Files: 11,12

Analysis

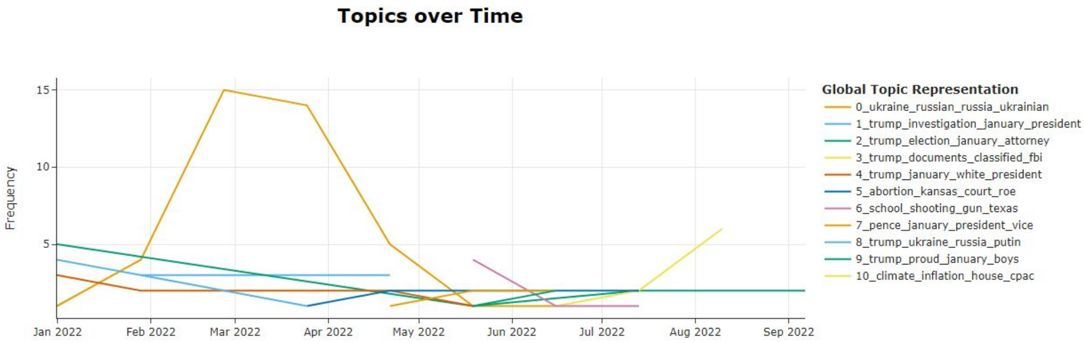
BERTopic

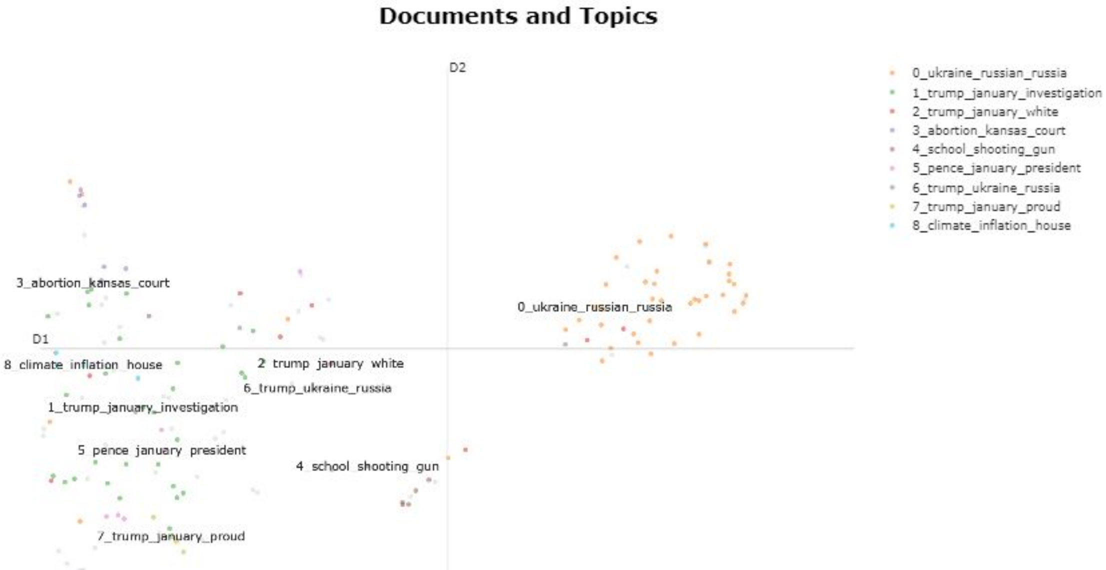
Approach: BERTopic is a pre-trained latent topic extractor that creates document embeddings which are able to capture meaning and be used for grouping documents of similar content.

I "seeded" my topics by introducing our limited dictionaries which increase the weight of the words within the word embeddings, making it more likely that topics form around them. BERTopic is recommended for data sets where you are classifying over 1,000 docs by topic, so using such a small data set (for Rachel Maddow shown here) required lowering the minimum number of documents needed to form a topic. I was then able to manually combine topics 1, 2, and 3 from the first chart into topic 1 for the second chart which shows the distance between the documents color coded and labeled by topic cluster.

Advantages: Leverages word embeddings and cosine similarity like the GoogleNews Vector method, but is not dependent on a limited dictionary to create the topic-specific vectors for comparison with the document vectors since it finds its own primary topic cluster within each documents.

Disadvantages: Can't adjust plot titles like the ones shown here. Can only lightly influence how topics are defined with the topic seeding, and the results vary even with the same data and parameters (the results varied much less however when trained on the full data sets instead of just 2022).





Notebook: 8 Data Files: 11,12

Discussion and Conclusion

We examined four different methods for extracting topic data from Tucker Carlson and Rachel Maddow transcripts in order to analyze how the frequency of topics discussed on both Rachel Maddow and Tucker Carlson broadcasts compare with PEW polling data which asks voters what issues they deem as very important for their vote. The four methods for topic extraction were: Transcript Word Frequency, Limited Dictionary Classification, Cosine Similarity with Google News Vectors, and BERTopic.

One very noticeable delta in results stood out from the Cosine Similarity with Google News Vectors approach specifically on the topic of foreign policy. All four methods identified foreign policy as one of the top topics discussed by the hosts except for the Cosine Similarity method for Maddow. One potential explanation is that Rachel Maddow referenced the war in Ukraine so much throughout 2022 that those words were down weighted when applying TF-IDF weighting per word. Another noticeable difference is how Carlson's topics are comparatively more diverse in the limited dictionary, whereas Maddow's topics are more diverse with Cosine Similarity. Since the average Carlson transcript is much shorter than Maddow's, perhaps the cosine similarity method is better at classifying longer text examples into topics, whereas the limited dictionary method is better suited for classifying shorter text examples. The cosine similarity amplifies two specific issues for Carlson: race and ethnicity, and foreign policy. As someone familiar with Carlson's rhetoric I know that those are two common themes of his that he will oftentimes relate to the issue of the day. For example, one monologue may ostensibly discuss the economy, but Carlson will mention how the "elites" care more about Ukraine than reducing inflation for the common man. Or a monologue about crime rates in Portland inevitably will include ironic mentions of "social justice warriors." Since the format of the Carlson and Maddow transcripts are so different, the cosine method could essentially be performing different types of analysis on Carlson and Maddow.

With googlenewsvectors we are limited by the average vector embedding of the PEW issue dictionary, with BERTopic we are able to extract latent topics that can provide deeper semantic meaning than what an average word vector from a limited dictionary of a few words can provide. The end result is impressively accurate and detailed topic classification that is much more likely to reflect higher level topics like the investigation into the classified material found at Mar-a-Lago, the Uvalde mass shooting, or, the war in Ukraine, especially when expanding the hosts' datasets beyond 2022. Conversely, it's harder to use BERTopic to analyze how often the much more generalized PEW issues are being discussed in the transcripts without significant manual tweaking and merging of topics, which then opens the door to the same researcher biases inherent in the creation of the limited dictionary used in our other methods.

Overall, all four methods were found to have their strengths and weaknesses both for extracting and classifying topics from political talk show host transcripts.

Statement of work and endnotes

Our collaboration across 3 different time zones started off very strong and we were able to deliver web scraping scripts for both hosts based on a combined effort early in the project. We then focused on exploring the data through a variety of methods and worked in parallel while sharing feedback of our different approaches on a semi-regular basis (minimum weekly). We tracked our meetings in a ‘Collaboration Log’. To improve collaboration in the future, we would recommend to clearly define interfaces to each of the building blocks for our analysis and onboard our work to github as early as possible to allow for more straightforward collaboration offline.

Work Items

Jennifer Shumway

- Webscraping
- Data Cleaning
- Pew Issue Polling
- Word frequency
- Google news vectors
- Report

Rachel Miller

- Data cleaning

Riley Schenck

- Webscraping
- Data Cleaning
- Pew Issue Polling
- Limited Dictionary
- BERT topic modeling
- Report

References

Jupyter Notebooks

[1] wescraping_rachel_maddow.ipynb
[2] webscraping_tucker_carlson.ipynb
[3] data_cleaning.ipynb
[4] pew_issue_data.ipynb
[5] word_count_funcAndViz.ipynb
[6] Combined_limited_dictionary_method.ipynb
[7] GoogleNewsVectors.ipynb
[8] BERT_Topic_Modeling.ipynb

Data Files

[9] Tucker_transcripts.tsv (output of webscraping)
[10] Maddow_transcripts.tsv (output of webscraping)
[11] Carlson_cleaned.tsv (output of data cleaning)
[12] Maddow_cleaned.tsv (output of data cleaning)
[13] Pew_Issue_Polling_2022.xlsx
[14] pew_ratings.csv

Collaboration Log

https://docs.google.com/spreadsheets/d/1O4Pr8__4ZTiP6MndpWa4Py1bfZ7UP2lanOW-0mg4onk/edit#gid=0