David Caspers
20 December 2024
IST 664, Natural Language Processing
Final Project Report

A Natural Language Processing Approach to Analyzing Political Rhetoric and Inferring Ideological Alignment

Introduction

Understanding political alignment is a critical aspect of analyzing public discourse. Politicians' statements often reflect their ideological positions, with subjective language providing valuable insights into their personal or partisan views. This study examines how the language used in public statements can reveal political alignment, focusing specifically on identifying and analyzing subjective content. Through this approach, an attempt is made to use subjective statements to rank politicians along an ideological spectrum, bridging linguistic analysis with ideological evaluation.

The broader implications of this work are substantial. Automatically determining where politicians lie on the political spectrum offers several key benefits: reducing bias and effort in manually analyzing large volumes of political data, improving transparency by illustrating how closely politicians' rhetoric aligns with their principles, and tracking shifts in rhetoric over time to provide valuable insights for voters, journalists, and analysts. By concentrating on subjective statements, this study aims to capture the opinions and values embedded in political rhetoric. Embedding-based ranking methods are employed to reveal ideological alignments, ultimately addressing the central question: How do politicians' words reflect their ideological leanings?

Exploratory Analysis

About the Data

This project utilized two primary data sources and one reference source to analyze political discourse and infer ideological alignment. The **NewsSD-ENG corpus**, sourced from Francesco, et al 2023.[1] This corpus provided labeled examples of subjective and objective sentences manually extracted and annotated from British newspaper and magazine articles that dealt with law, civil rights, economics, and other controversial political subjects. In the annotator's case, subjective statements reflected personal opinions, feelings, or tastes, while objective statements were defined as everything else. This dataset was critical for training and evaluating the sentence

[1] Antici, Francesco, et al. "A corpus for sentence-level subjectivity detection on english news articles." *arXiv preprint arXiv:2305.18034* (2023).

classifier to distinguish subjective content, which forms the foundation of the analysis. The articles contained the following fields:

- Sentence: the actual text that was annotated
- Label: SUBJ or OBJ
- Solved Conflict: Booleen value which reflected whether the annotators conflicted in their label. The majority vote was used to assign the class label regardless of disagreement.

The second data source, the **VoteSmart API**[2], provided public statements such as tweets, interviews, and speeches from members of Congress during 2022–2023. These statements served as the primary corpus for analyzing political opinions, offering real-world examples of language used by politicians to express their viewpoints. By focusing on subjective content within these statements, the project aimed to capture meaningful insights into ideological leanings. Statements from 430 politicians consisting of 37,092 subjective sentences were collected. A total of 105 members of the House or Senate were not represented in the scraped data.

Lastly, the **Limited Government Scorecard**[3], developed by the Institute for Legislative Analysis, ranked politicians based on how their voting records aligned with limited government principles. This expert-driven benchmark allowed for a comparison of the project's automated rankings with established evaluations from experts, ensuring a reliable assessment of the proposed methodology. Together, these three data sources provided a robust foundation for identifying, analyzing, and validating political alignment through language.

[2] https://www.votesmart.org/votesmart-api
[3] https://scorecard.limitedgov.org/guides/US-Guide-2022.pdf

David Caspers
20 December 2024
IST 664, Natural Language Processing
Final Project Report

Data Visualization (NewsSD-ENG corpus)

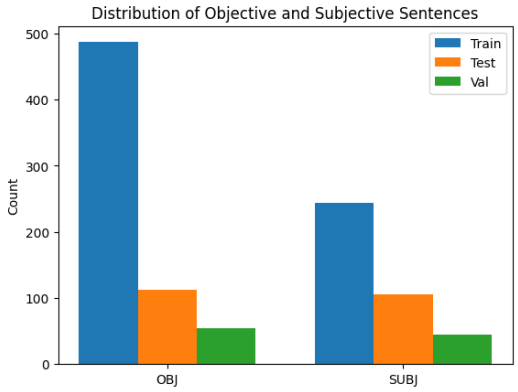

Figure 1 - Sentence Label Distribution Across Data Sets

This bar chart illustrates the distribution of objective and subjective sentence labels across the training, testing, and validation datasets. The training set exhibits a disproportionate number of objective sentences, potentially introducing bias into the model. To mitigate this, label weighting will be incorporated into the loss function during training to ensure balanced learning for both labels.

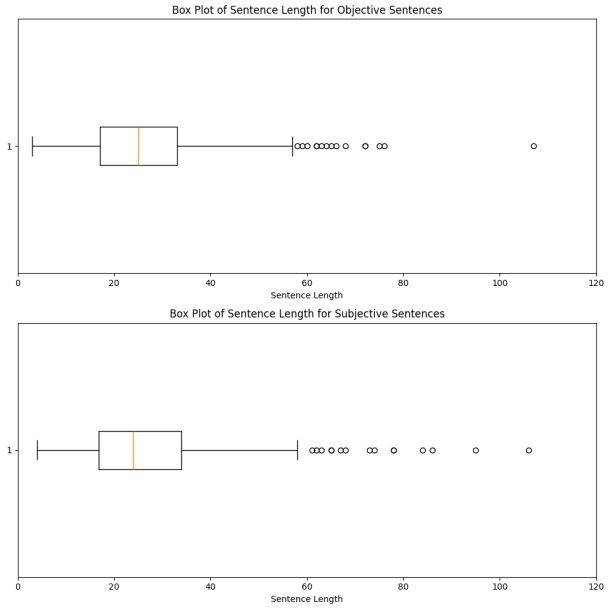

Figure 2 - Sentence Lengths Boxplot (By Label)

The box plots show the distribution of sentence lengths for objective and subjective sentences. Both sentence types display similar patterns, though subjective sentences exhibit slightly greater variability, including longer sentences and more extreme outliers. To address the wide variation in sentence length and avoid bias towards longer sentences, normalization techniques like TF-IDF weighting will be applied during preprocessing.

David Caspers
20 December 2024
IST 664, Natural Language Processing
Final Project Report

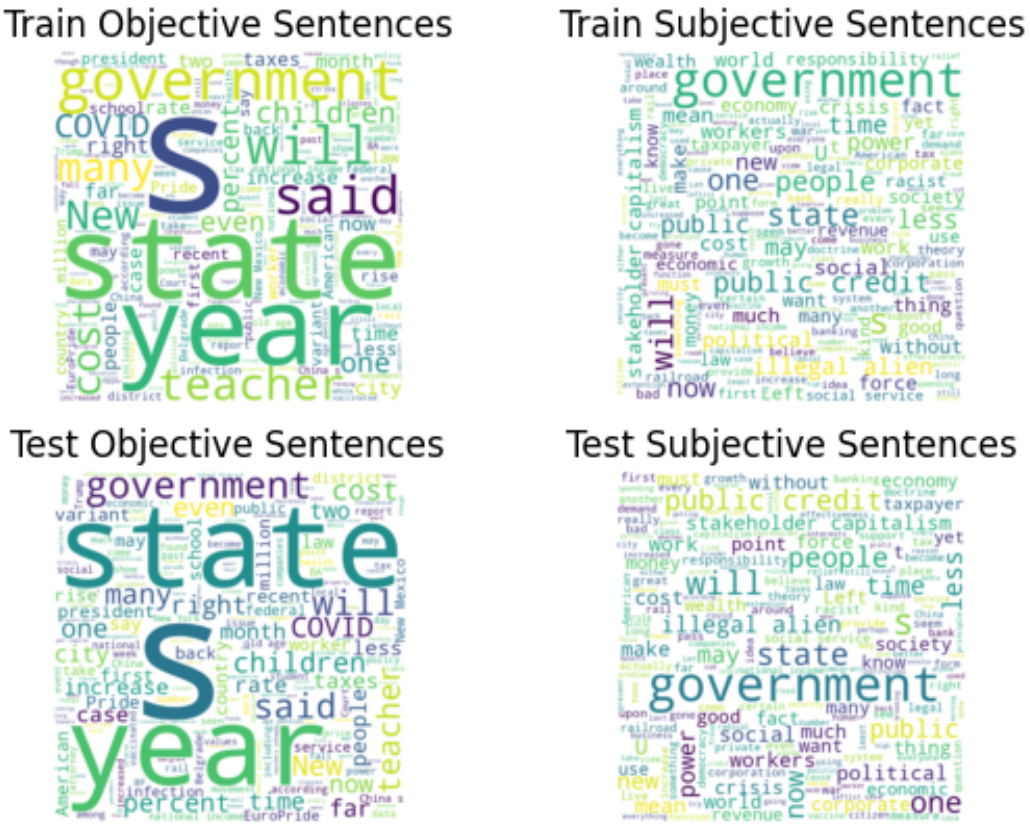

Figure 3 - Word Clouds By Sentence Label

The word clouds highlight the most frequently occurring words in objective and subjective sentences across the training and testing datasets. Objective sentences prominently feature terms such as "state," "year," and "government," reflecting a focus on factual or formal language. Subjective sentences, on the other hand, exhibit a more diverse lexicon with balanced word usage. The consistency of dominant words between training and testing sets for both sentence types suggests robust linguistic patterns within the dataset, enhancing its reliability for model training. However, overlapping terms like "government" in both labels may complicate classification, especially when using simple bag-of-words approaches.

David Caspers
20 December 2024
IST 664, Natural Language Processing
Final Project Report

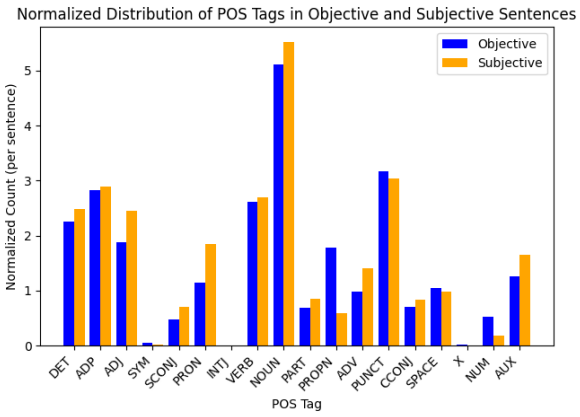

Figure 4 - Distribution of Parts of Speech Tags (Normalized by Number of Sentences)

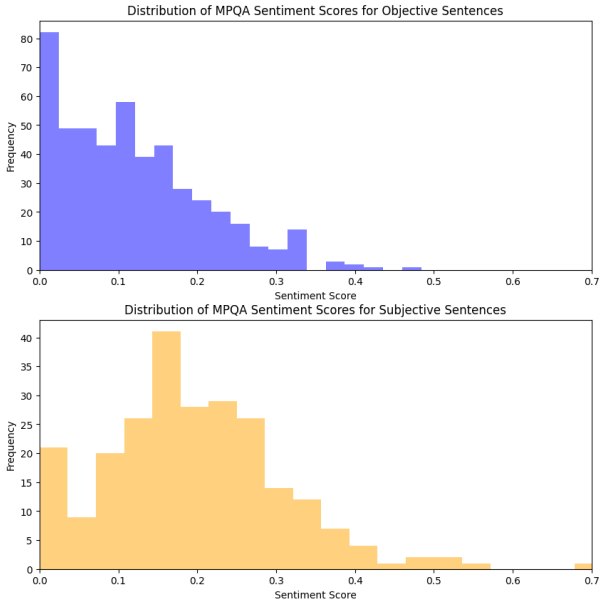

Figure 5 - Histogram of MPQA Scores by Sentence

The bar chart shows the normalized distribution of part-of-speech (POS) tags in objective and subjective sentences. To account for the higher prevalence of objective sentences in the dataset, the counts were normalized by the total number of sentences in each label. The distribution of other POS tags is relatively balanced between the two categories. This suggests that the overall linguistic structure is similar across both labels. Consequently, additional features beyond POS tags may be required for effective classification.

The histograms show that subjective sentences have a sentiment score distribution skewed to the right, peaking around 0.2, while objective sentences are most frequently aligned with scores near 0. Despite the differences in distribution, there is significant overlap between the two, particularly at lower scores. This overlap suggests that while MPQA sentiment scores may provide some utility for classification, they are unlikely to be sufficient on their own for accurate sentence-level differentiation.

David Caspers
20 December 2024
IST 664, Natural Language Processing
Final Project Report

Data Preparation and Transformations

The VoteSmart corpus was collected and processed through a combination of web scraping and data cleaning steps to prepare it for analysis. Using Python's requests and BeautifulSoup libraries, public statements on various topics from the VoteSmart website were scraped. The scraper targeted multiple categories such as "defense" and "energy," retrieving key details including the statement's date, title, politician, subject, and full text. The data was systematically saved to CSV files organized by topic.

After scraping, the data underwent cleaning and organization. Politician names were cleaned and normalized by removing non-alphanumeric characters and spaces, and statements were categorized by year (2022 or 2023) and saved into separate directories. Additionally, statements were grouped into individual files for each politician and year, simplifying the downstream processing. These steps ensured the data was consistently formatted and accessible for follow-on analysis.

Tokenizer (Spacy and BERT), POS Tagging, and TFIDF

The Spacy tokenizer was used to tokenize the NewsSD-ENG corpus into words and assign part-of-speech (POS) tags. For the transformer-based DistilBERT model, the DistilBERT tokenizer was used to preprocess the text data. The DistilBertTokenizerFast from the Hugging Face library was employed to tokenize the sentences into subword tokens, which were then padded and truncated to a maximum sequence length of 512 tokens. Since the "distilbert-base-uncased" model is lowercased, the tokenizer also converted all text to lowercase during preprocessing, ensuring consistency with the pre-trained embeddings. This ensured consistent input dimensions for the transformer model.

The Spacy POS tagging was leveraged in two ways during data preparation. First, POS tags were appended to each word in the text, allowing the model to differentiate between different uses of the same word (e.g., "run" as a noun versus a verb). Second, the POS tags were aggregated into counts for each sentence, creating features that capture the grammatical structure of the text.

To represent the text data for traditional models, a TFIDF Count Vectorizer was used to aggregate words. Unlike simple term frequency-based vectorization, TFIDF adjusts weights based on how frequently a word appears across documents, reducing the impact of common but less informative words. Stop words were not removed from the corpus, as their TFIDF weights inherently diminished their influence while still allowing them to contribute useful contextual information. This approach provided a nuanced representation of the text, retaining potentially meaningful but frequent words while reducing their noise.

David Caspers
20 December 2024
IST 664, Natural Language Processing
Final Project Report

Models (Sentence Classifiers):

Subjective Classification Model Architectures and Features

This study developed two types of Naive Bayes models for classification: Multinomial Naive Bayes (MNB) and Bernoulli Naive Bayes (BNB). MNB is well-suited for text data as it models frequencies, while BNB offers a complementary perspective in text classification tasks by only analyzing presence-absence features. Both models were evaluated with various feature configurations to explore their effectiveness.

Feature variations included a Bag of Words representation (unigram), n-grams (bigrams and trigrams), and combinations of unigrams, bigrams, and trigrams to capture increasingly complex word sequences. Additional experiments incorporated linguistic features, such as appending part-of-speech (POS) tags to words to distinguish different grammatical uses and aggregating POS counts per sentence to capture broader syntactic patterns. MPQA sentiment scores were also included as features to provide a sentiment-based perspective for classification. These variations allowed for a comprehensive comparison of how different text representations influenced model performance.

In addition to the Naive Bayes models, a fine-tuned version of the transformer model "distilbert-base-uncased"[4] was developed and evaluated. DistilBERT, a distilled version of BERT, is a pre-trained transformer model designed to capture deep contextual representations of text while being computationally efficient. Fine-tuning involved training the model on the specific dataset used in this study, allowing it to adapt its embeddings and classification layers to the task of distinguishing between objective and subjective sentences.

To optimize performance while minimizing computational overhead, only the last two layers of the DistilBERT model were fine-tuned, while the remaining layers were frozen to preserve the pre-trained knowledge. Class imbalance was addressed by computing class weights inversely proportional to class frequencies and passing these weights to the loss function during training. The model was compiled using an Adam optimizer and a Sparse Categorical Crossentropy loss function with logits. Training was conducted over 20 epochs with early stopping based on validation performance. This selective fine-tuning approach allowed for efficient adaptation of the pre-trained model to the specific classification task while leveraging its pre-existing rich contextual understanding of text.

---

[4] Sanh, Victor, et al. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." *arXiv*, 11 Oct. 2019, arXiv:1910.01108.

David Caspers
20 December 2024
IST 664, Natural Language Processing
Final Project Report

Model Results:

Model *all Naïve Bayes models use TFIDF*	***Accuracy BNB / MNB***	***Precision BNB / MNB***	***Recall BNB / MNB***	***F-1 BNB / MNB***
Bag of Words (Unigram)	53.42% 49.31%	62.83% 75.24%	53.42% 49.32%	44.95% 33.56%
Bigrams + Trigrams	47.95% 47.95%	43.94% 46.21%	47.95% 47.94%	32.90% 34.32%
Unigram + Bigram + Trigram	65.66% 67.44%	64.98% 64.14%	65.66% 67.44%	65.81% 62.72%
Part of Speech Appended to Words, Unigram, Bigram, Trigram	50.22% 49.77%	54.32% 75.35%	50.22% 49.77%	40.93% 34.53%
Unigram, Bigram, Trigram, and POS Counts	64.84% 48.86%	69.45% 57.92%	64.84% 48.86%	63.12% 33.35%
Unigrams + Bigrams + Trigrams + MPQA Sentiment Score	65.30% 49.32%	69.81% 62.34%	65.30% 49.32%	63.68% 34.31%
Tuned “distilbert-base-uncased”	75.80%	76.73%	75.80%	75.68%

All Naive Bayes models utilized TFIDF, with the Unigram + Bigram + Trigram combination performing the best among them, achieving an accuracy of 67.44% and an F1-score of 65.81%. This suggests that capturing contextual information from adjacent words (e.g., word pairs and trigrams) significantly enhances the model's ability to classify sentences. On the other hand, incorporating additional features like part-of-speech (POS) tagging and MPQA sentiment scores did little to improve performance for the Naive Bayes models, indicating that these features may not have provided enough discriminative power.

David Caspers
20 December 2024
IST 664, Natural Language Processing
Final Project Report

The fine-tuned "distilbert-base-uncased" model outperformed all Naive Bayes models, achieving the highest accuracy (75.80%) and F1-score (75.68%). This highlights the effectiveness of transformer-based architectures in leveraging contextual and semantic information in text, enabling a deeper understanding of linguistic nuances.

Applying Model to Determine Politician's Political Alignment

The VoteSmart politician data was parsed and classified using the fine-tuned DistilBERT model to identify subjective statements within public speeches. The dataset was preprocessed by breaking full-text statements into individual sentences using Spacy's sentence tokenizer, which were then tokenized with the DistilBERT tokenizer, converting text into lowercase word tokens while applying padding and truncation. The model predicted class probabilities for each sentence, and subjective sentences were retained for further analysis. Examples of the classified texts can be found below. Statements from 430 politicians (US House and Senate Members) consisting of 37,092 subjective sentences were collected.

Text: Today Rep. Ilhan Omar released the following statement to commemorate the one year anniversary of the January 6th insurrection.
Predicted Class: OBJ
Class Probabilities: [0.9988122 0.00118786]

Text: "Today marks one year since the attacks of January 6th.
Predicted Class: OBJ
Class Probabilities: [9.990722e-01 9.278190e-04]

Text: I will never forget the experience of fearing for my life, my fellow members, and staff on a day designed to show the strength of our democracy.
Predicted Class: SUBJ
Class Probabilities: [0.00251374 0.99748623]

Figure 6 - Sampled Prediction Results for Rep. Ilhan Omar

Political Affiliation Methodology:

For each politician, the embeddings of all retained subjective sentences were averaged to create a single "representative" embedding that encapsulates their overall rhetorical stance. To compare politicians across the political spectrum, baseline embeddings were established using representatives with opposing political ideologies, as defined by the Limited Government

Scorecard[5]. In this case, Rep. Debbie Dingell (D-MI) and Rep. Ralph Norman (R-SC) were selected.

The embedding of each politician was then projected onto a vector defined by these baseline embeddings, with the resulting projection reflecting the magnitude and direction of their political affiliation. The visual illustrates this methodology, where the embedding of a politician to be ranked is shown relative to baseline embeddings of two politicians with opposing ideologies. The dashed lines represent the projection of the embedding onto the baseline vector, and the magnitude of this projection quantifies the politician's alignment with a particular ideological stance. This methodology allowed for a nuanced ranking of politicians based on the content of their subjective statements, and the results were compared to the Limited Government Scorecard to validate the approach.

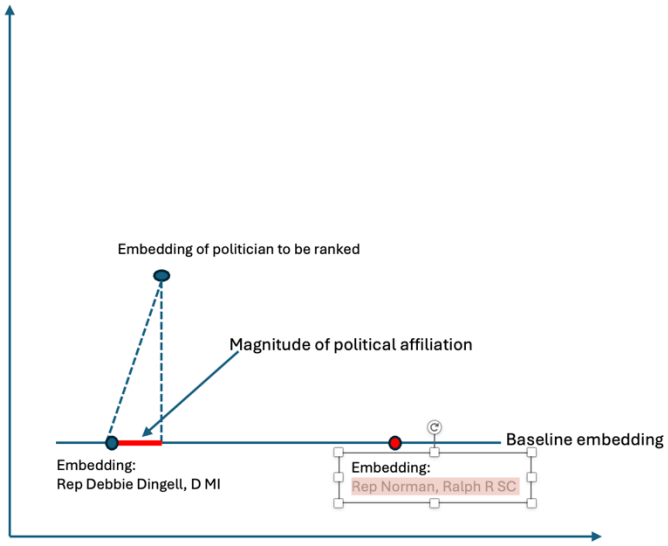

Figure 7 - Illustration of Ranking Methodology

To calculate the projection of a point v_p (the embedding of the politician to be ranked) onto a vector defined by the embeddings of two baseline politicians, v_1 and v_2, the following equation was used:

$$Projection = \frac{(v_p - v_1) \cdot (v_2 - v_1)}{||v_2 - v_1||^2} \cdot (v_2 \cdot v_1)$$

[5] https://scorecard.limitedgov.org/guides/US-Guide-2022.pdf

David Caspers
20 December 2024
IST 664, Natural Language Processing
Final Project Report

Results Analysis

The rankings of the top 45 lawmakers committed to limited government, as defined by the Center for Legislative Analysis's 2022 Report Scorecard, were compared against their corresponding rankings using this embedding-based methodology. Due to limitations in the VoteSmart API, 14 lawmakers did not have any available data and were therefore omitted from the analysis, leaving 31 lawmakers. The limited government list was subsequently reranked to reflect this adjusted subset of lawmakers.

Rankings of Champion of Limited Government House of Representative Lawmakers:		
Politician	**Actual Rank**	**Predicted Rank**
Rep Chip Roy	1	45
Rep Andy Biggs	2	38
Rep Matt Rosendale	3	70
Rep Lauren Boebert	4	17
Rep Greg Steube	5	109
Rep Tim Burchett	6	222
Rep Jim Jordan	7	133
Rep Mary Miller	8	59
Rep Ronny Jackson	9	6
Rep Paul Gosar	10	53
Rep Andy Harris	11	161
Rep Lance Gooden	12	173
Rep Troy Nehls	13	20
Rep Dan Bishop	14	24
Rep Kevin Hern	15	276
Rep Ron Estes	16	171
Rep Randy Weber	17	21
Rep Andrew Clyde	18	51
Rep Debbie Lesko	19	294
Rep Warren Davidson	20	41
Rep Mark Green	21	16
Rep Russ Fulcher	22	97
Rep Tom Tiffany	23	50
Rep Barry Loudermilk	24	10
Rep Jodey Arrington	25	232
Rep Matt Gaetz	26	13

Rep Jim Banks	27	262
Rep Glenn Grothman	28	279
Rep Gary Palmer	29	3

The model's performance was generally poor, with significant deviations between the actual and predicted rankings for many lawmakers. An analysis of the data used for ranking reveals that lawmakers with rankings further from their actual position often had speech data focused primarily on single topics, such as a bill they sponsored. This narrow focus likely resulted in embeddings that were not as representative of their broader political positions or rhetoric. A promising area for future exploration would be to filter the dataset and only compare politicians with robust embeddings generated from diverse and representative speech data. This refinement could improve the alignment between embedding-based rankings and actual rankings.

Limitations and Areas for Future Study

This study faced several inherent challenges and limitations that provide opportunities for refinement and future exploration. One major difficulty was accurately distinguishing between objective and subjective sentences. Determining whether a statement is purely factual or opinion-based often requires a nuanced understanding of context and, at times, can itself be subjective. This ambiguity likely impacted the classification model's performance, as subjective statements often carry subtle tones that are challenging for automated systems to capture without deeper contextual knowledge. This is particularly a challenge in the context of politics, where individuals have an incentive to appeal to large audiences by avoiding explicit statements or framing their statements in a way that maintains ambiguity, allowing them to resonate with diverse groups. Politicians often employ rhetorical strategies that blur the line between objective and subjective language, making the task of automated classification inherently complex.

The embedding projection methodology was also constrained by the quality and diversity of the data. Some politicians' embeddings were derived from narrowly focused speech data, such as those centered around specific legislative bills, which limited their representativeness. This lack of generality in the embeddings reduced the model's ability to provide accurate rankings. Furthermore, the exclusion of 14 lawmakers due to missing data in the VoteSmart API restricted the analysis, leaving a smaller pool of politicians and potentially introducing bias.

Future research could address these challenges by exploring methods to refine the subjective-objective classification process, such as incorporating more advanced contextual models or leveraging external datasets for additional context. To improve the representativeness of embeddings, analyses could be restricted to politicians with quotations that span a certain minimum threshold of diverse topics, ensuring a broader coverage of their rhetorical styles and ideological stances. Additionally, expanding the data sources to include other forms of public discourse, such as interviews or social media posts, could enrich the dataset and provide a more

holistic view of each politician's rhetoric. By addressing these limitations, future work could improve the accuracy and robustness of the methodology, making it more effective in analyzing and understanding political rhetoric.

Conclusion

This study explored the use of natural language processing and embedding-based methodologies to analyze political discourse and infer political alignment. By identifying and analyzing subjective content in politicians' public statements, embeddings were generated to represent their rhetorical and ideological stances. While the fine-tuned DistilBERT model performed well in classifying subjective sentences, the subsequent embedding-based ranking methodology revealed limitations. Significant discrepancies were observed between the predicted and actual rankings of lawmakers' political affiliations, particularly for those whose speech data lacked diversity and focused on single topics. These findings underscore the importance of comprehensive and representative datasets in embedding-based analyses.

Despite these challenges, the methodology demonstrates potential for automating the evaluation of political alignment by leveraging subjective language. With improvements, such as filtering data to ensure robust and diverse embeddings, this approach could provide a complementary tool to expert-driven evaluations like the Limited Government Scorecard. The insights gained from this project highlight both the promise and limitations of NLP techniques in political analysis, setting the stage for future refinements to better capture the complexity of political discourse.