Text and Tabular Data Analysis of American Political meetings along Time and Space

Bar Avraham and Adam Uziel (baravrah,adamu)@post.bgu.ac.il
Department of Software and Information Systems Engineering
Ben-Gurion University of the Negev Beer-Sheva, Israel

1 Introduction

Government meetings play a pivotal role in deliberation, decision-making, and policy shaping, offering insights into the political landscape of a given region. In today's data-driven world, researchers and policymakers leverage these platforms to uncover hidden patterns, monitor evolving trends, and gain deeper insights into complex dynamics. The LOCALVIEW(6) dataset represents a vast compilation of government meeting records, incorporating video, text, and tabular data. It encompasses a range of years and geographical locations across the United States, providing a comprehensive snapshot of political activities. This expansive dataset captures not only political discourse but also provides a unique opportunity for in-depth exploration. Within this study, our exploration of the LOCALVIEW dataset focus in on text transcripts and tabular data from meetings. By employing a multi-pronged approach that amalgamates natural language processing, machine learning, and visualization techniques, our aim is to uncover latent topics, trace sentiment trends, and predict political dynamics.By utilizing the Latent Dirichlet Allocation (LDA) model (5), latent themes within the dataset are revealed. Subsequently, our in-depth emotion analysis employs techniques such as cluster analysis, dimensionality reduction, and graph analysis to paint a comprehensive picture of the emotional intricacies underpinning the meetings. Leveraging the advanced capabilities of the BERT model (3), we predict the state of origin with remarkable precision, accurately discerning the source state of each political meeting. Furthermore, our exploration into the relationship between racial demographics and recent election voting patterns delves deep into the importance of race demographic features as opposed to general time-varying demographic features. As we navigate through this study, we aspire to shed light on the evolving nature of political discourse, offering valuable insights and contributing to the broader understanding of American political dynamics.

2 Related Work

The analysis of text and tabular data, especially in the context of political meetings, has garnered significant attention in recent years. The article "Data Is Plural"(4) mentions the development of the LOCALVIEW database by Soubhik Barari and Tyler Simko. It emphasizes the importance of structured analysis of political meetings, shedding light on the potential of topic analysis in this domain. A comprehensive article on Towards Data Science(1) delves into the nuances of applying topic modeling techniques to political texts, offering insights into the challenges and potential of this approach. A study titled "PoliBERT: Classifying political social media messages with BERT"(7) leverages the pre-trained BERT model to classify political messages on social media platforms, emphasizing the model's potential in understanding political discourse in the digital age. The research "Emotion Analysis of Ideological and Political Education"(2) integrates BERT for emotion analysis in ideological and political education, showcasing the depth of insights that can be derived from such an approach.

3 Dataset

The LOCALVIEW dataset comprises over 139,616 videos and transcripts from government meetings spanning 2006 to 2022. It encapsulates insights from 1,012 locations and 2,861 government bodies across the US. This vast dataset enables large-scale analyses of policy domains and temporal patterns. The standardized transcripts promote easy comparisons, and the dataset's ongoing updates ensure its relevance. In our study, we utilized this dataset, with a focus on textual components, sourcing data from CSV and JSON files through LocalView web service URLs.

3.1 Data Pre-processing

To prepare the data for meaningful analysis, we applied the following pre-processing steps:

Text Cleaning: We started by ensuring data quality by removing rows where the text content was missing, resulting in a final dataset comprising 99,190 rows. Subsequently, we standardized the text data by converting it to lowercase, expanding contractions, removing possessive markers, eliminating punctuation, and handling common stopwords. This resulted in a cleaned DataFrame, ready for analysis.

Lemmatization: In order to discover meaningful topics within the text, we transformed words into their root forms. We employed the spaCy library for lemmatization on the 'caption text clean' column, leading to a more interpretable representation of textual content.

Forming Bigrams and Trigrams: We created bigrams (sequences of two adjacent words) and trigrams (sequences of three adjacent words) to enhance topic modeling accuracy by capturing multi-word expressions and nuanced relationships. Treating these sequences as single units revealed complex and context-rich topics. Then we selected the top 500 n-grams and substituted them in the text column with versions where their constituent words are connected using underscores.

Filtering for Only Nouns: Nouns are likely indicators of topics. For example, 'The curriculum overhaul is underway, and teachers are implementing innovative techniques.' In this case, the focus of the discussion is clearly on 'curriculum overhaul' and 'teachers implementing innovative techniques.' The supporting words contribute context to these core concepts. By filtering for nouns, we refined the text to include terms that distinctly outline the central theme of each topic.

4 Methods

In this section, we outline the methodologies employed to extract insights from the LOCALVIEW dataset.

4.1 LDA Model

The core objective of topic modeling is to identify latent topics within the dataset. We utilized LDA, a probabilistic generative model, to uncover the underlying themes within the text. Through the gensim library, we constructed a document-term matrix, which serves as the foundation for LDA. We constructed an LDA model with 20 topics and showcased the most prevalent words within each topic. We determined the optimal number of topics through a combination of coherence scores)Figure 14 in the appendix) and trial and error.

Sometimes, words that are ranked as top words for a given topic may be ranked high because they are globally frequent across the text in a corpus. Relevancy scores help prioritize terms that belong more exclusively to a given topic. This can enhance interpretability further. The relevance of term w to topic k is defined as:

$$r(w, k|\lambda) = \lambda \log(\phi_{kw}) + (1 - \lambda) \log \left(\frac{\phi_{kw}}{p_{kw}} \right)$$

where ϕ_{kw} is the probability of term w in topic k and $\frac{\phi_{kw}}{p_{kw}}$ is the lift in the term's probability within a topic to its marginal probability across the corpus (this helps discard globally frequent terms). A lower λ value assigns more importance to the second term, which emphasizes topic exclusivity. After experimenting, we determined that $\lambda = 0.2$ was the optimal choice, and we showcased the top 15 relevant words for each topic. Utilizing these words, we formulated appropriate titles for each topic. Notably, certain topics exhibited ambiguity or lacked informative content. Moreover, it's important to highlight that certain meetings might involve discussions spanning multiple topics. To tackle these issues, we introduced an extra column that enumerates the pertinent topics for each entry, organized in descending order of relevance. This multifaceted strategy ensures a thorough exploration of the topics existing in the dataset. Furthermore, each entry in the dataset was allocated the most suitable topic for the next visualizations.

4.1.1 Visualization

Temporal Evolution of Topics: An interesting aspect of the dataset is the temporal evolution of topics over the years. By aggregating topic assignments and counts over time and normalizing them based on the total number of meetings in each year, we generated line plots that illustrate the changing prevalence of topics. The highest relative count in each year was highlighted to emphasize pivotal shifts in topic trends. For example, Figure 1 illustrates the temporal evolution of the "Public Health" topic. We observe a significant increase in topic prevalence starting from 2019, peaking in 2020, followed by a gradual decline. This trend could indicate a heightened focus on public health issues during 2019-2020, possibly related to global events such as the COVID-19 pandemic, followed by a return to more typical levels. In the context of the "Education" topic, Figure 4 reveals a distinctive peak in 2011, This trend could be attributed to various factors. One plausible explanation might involve educational reforms that occurred during that year. It's worth considering that political discourse in the United States, where these meetings took place, is influenced by a wide range of events, including elections, social movements, economic developments, and policy changes.

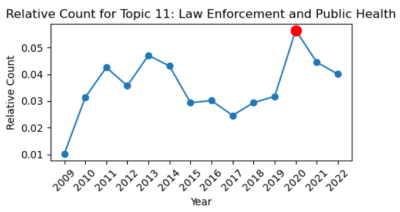

Figure 1: Public Health

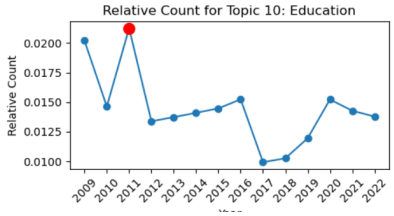

Figure 2: Education

Word Clouds were generated to visually represent the most relevant terms within each topic. By extracting relevant terms and their associated relevance scores, we constructed word clouds that emphasize the most important terms for each topic. This visual representation provides a quick and intuitive way to grasp the essence of each topic. Figure 3 presents words from Topic 2: Housing, while Figure 4 presents words from Topic 11: Education.

Figure 3: Housing Word-Cloud

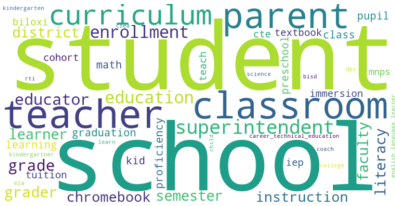

Figure 4: Education Word-Cloud

4.2 Exploring Discussions on Equality

In addition to uncovering topics through topic modeling, we also sought to investigate the prevalence of discussions related to the concept of equality within the meetings. To gain insights into the temporal and geographical trends of equality discussions, we performed the following analysis: We began by quantifying the mentions of the term 'equality' in the cleaned caption texts of the meetings. For each transcript, we calculated the normalized count of the term 'equality' by dividing the number of occurrences by the total number of words in the text. This resulted in a normalized score representing the degree of focus on equality within each transcript. To prepare the data for analysis, we applied a logarithmic transformation to the normalized scores. This transformation enhanced the visibility of changes in discussions on equality while addressing the impact of extreme values. To visualize the temporal evolution of equality discussions, we grouped the data by year to calculate the average of the transformed equality scores for each year. We then created a line plot (Figure 5) to visually represent the trend. The upward trend in the plot indicates an increasing focus on discussions related to equality within the analyzed meetings.

Additionally, we investigated geographical variations in equality discussions. We grouped the data by both year and state to calculate the avrage of the transformed equality scores for each state-year combination. We then generated interactive geospatial maps to visualize these variations across different time periods (2009-2013 (Figure 6), 2014-2018 (Figure 7),

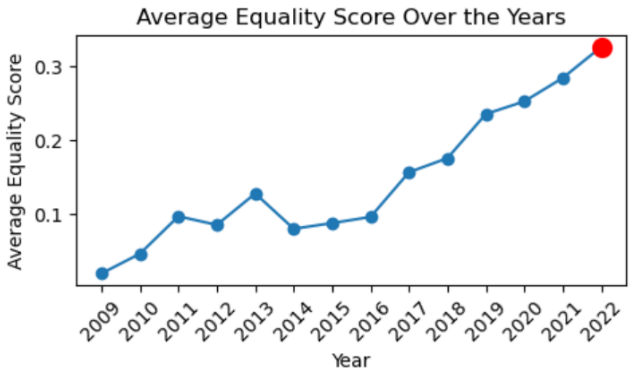

Figure 5: Temporal Trend of Average Equality Score

2018-2022 (Figure 8)). The increasing density of points across these maps as the years progress visually signifies a growing emphasis on discussions related to equality.

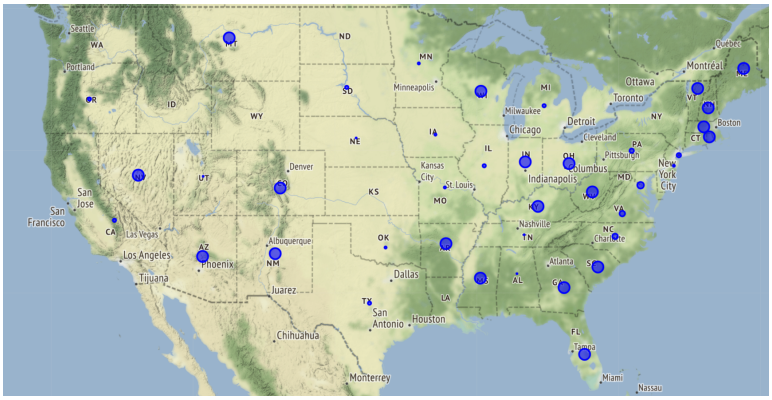

Figure 6: Equality Score 2009-2013

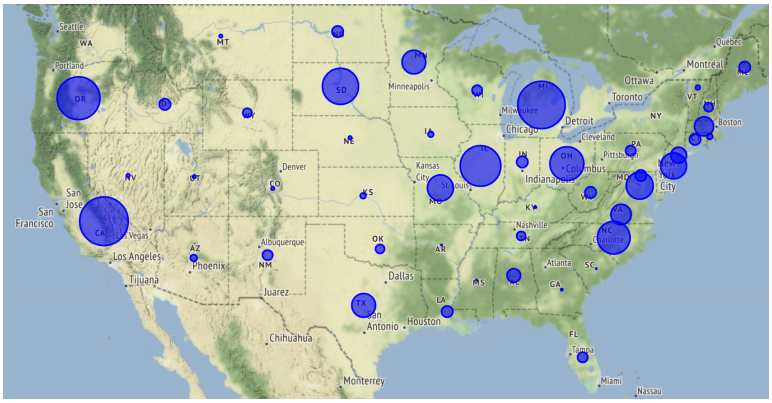

Figure 7: Equality Score 2014-2018

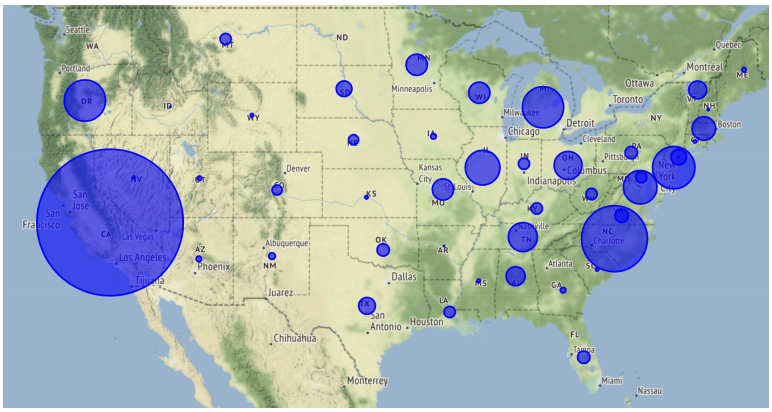

Figure 8: Equality Score 2019-2022

The following subsections make use of state-of-the-art transformer models on the meetings' caption text. Those models treat their input sequentially and thus the inflections of words and stopwords are critical. Thus, the only pre-processing steps taken on the input text in this section were removing punctuation and converting letters to lower-case.

4.3 Emotion Analysis

In this section, we delve into the underlying sentiments expressed during political gatherings. Recognizing the intricate tapestry of emotions in such meetings provides a nuanced understanding of the political climate and discourse. The changes in those emotions were examined in different places (states) and at different times (years). To achieve this, we employed a state-of-the-art emotion prediction model, specifically the "twitter-roberta-base-emotion-multilabel-latest" from Cardiff NLP. This model, fine-tuned on multilabel emotion data, allows for the simultaneous prediction of multiple emotions from a given text. By processing the transcripts of each meeting through this model, we were able to assign a set of emotions, offering a comprehensive emotional profile of the discussions.

4.3.1 Emotion Analysis over Years

Initially, we calculated the frequency of each emotion in each year, and normalized these frequencies by the number of meetings in each year. Figure 9 shows the changes in the normalized frequencies over the years for each emotion. Over the years, political meetings have exhibited a dynamic range of emotions. "Anticipation" has seen a steady rise, peaking in 2020, suggesting growing expectations or uncertainties in political discourse. While "joy" was dominant in 2006, it declined until 2010, hinting at possible shifts in political sentiment. Interestingly, "optimism" remained relatively stable from 2011, indicating consistent hope or confidence in future outcomes. Emotions like "anger" and "disgust" peaked around 2009-2010, perhaps reflecting turbulent times or contentious issues. On the other hand, emotions like "fear" and "pessimism" remained low but had slight upticks in certain years, suggesting underlying concerns. The sporadic presence of "trust" and "love" might indicate specific events or themes during those years. Overall, these trends provide a glimpse into the evolving emotional landscape of political meetings over time.

4.3.2 Emotion Analysis over States

We created weights that reflect how much one emotion dominates political meetings in each state. In the analysis of the frequency distribution of different states and emotions using bar plots, two problems were encountered. First, some states appear more frequently in the dataset than others. Second, some emotions appear more frequently in the dataset in general, regardless of state. To create normalized and unbiased state-emotion pairs weights, we calculated the weights using tf-IDF in such a way that the "documents" represent the states and the "words" represent all instances of emotions associated with those states. The TF-IDF vectors for state's emotions present a rich representation of the weighting of different emotional expressions across different states.

After obtaining a vector representing each state by emotions, we wanted to utilize PCA to reduce the dimension of each vector and in this way visualize the state-emotion space with a scatter plot. Figure 10 shows the 2D emotion space PCA visualization of the states.

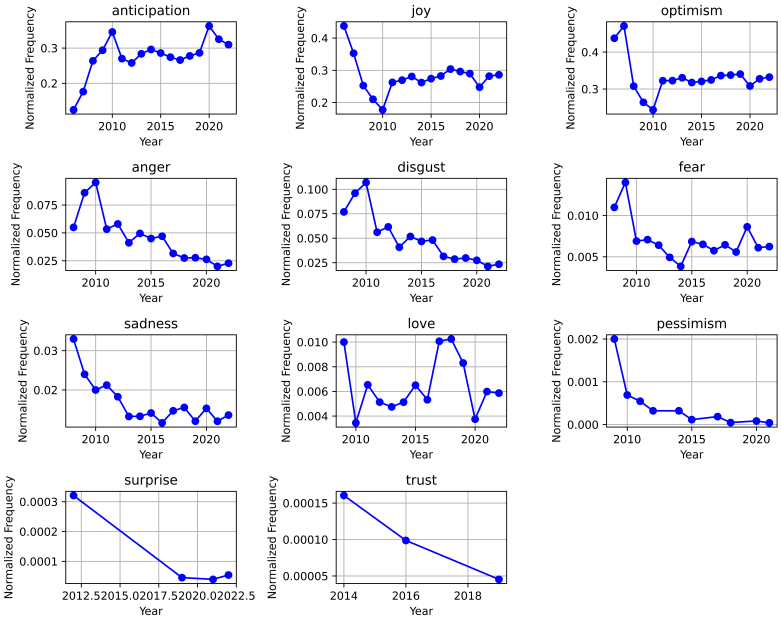

Figure 9: Emotions' normalized frequencies over years.

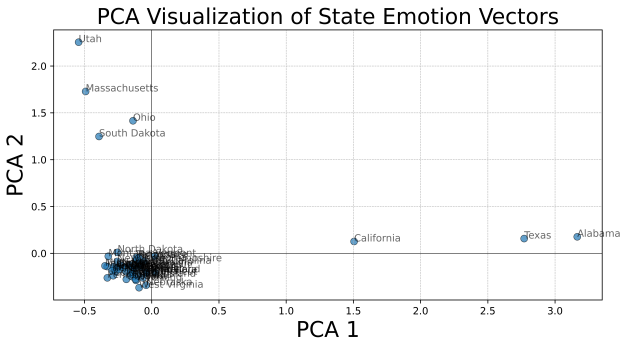

Figure 10: 2D PCA visualization on States' emotion vectors

3 clusters of states can be seen in the PCA plot. To provide stronger evidence for the existence of those clusters, we conducted two cluster analysis methods - Kmeans and Hierarchical Agglomerative Clustering, based on the well-known cosine similarity metric in information retrieval. Before running the 2 clustering algorithms, we performed an elbow plot to determine if 3 is actually the optimal k (number of clusters). From the elbow plot, figure 11 shows that the number 3 is chosen as the optimal number of clusters because at this point, the reduction in distortion begins to level off, indicating diminishing returns from adding more clusters. The dendogram plot created from Hierarchical Agglomerative Clustering can be seen in the appendix. Both KMEANS and hierarchical clustering with three clusters yielded three identical state clusters:

1. cluster 1: Texas, Alabama and California.
2. Cluster 2: Ohio, Massachusetts, South Dakota and Utah.
3. Cluster 3: All the other states.

According to the captions of the political meetings, 'trust' and 'surprise' were the rarest emotions. When inspecting cluster 1, Texas, Alabama, and California are the only states with relatively high tdidf weights for the emotion 'trust', all other states have a weight of zero. Ohio, Massachusetts, South Dakota and Utah were the only states with relatively large tdidf weights for the emotion 'suprise', all other states had weights of 0. An analysis which excludes those states could reveal more complex emotional relationships and clusters in the remaining states.

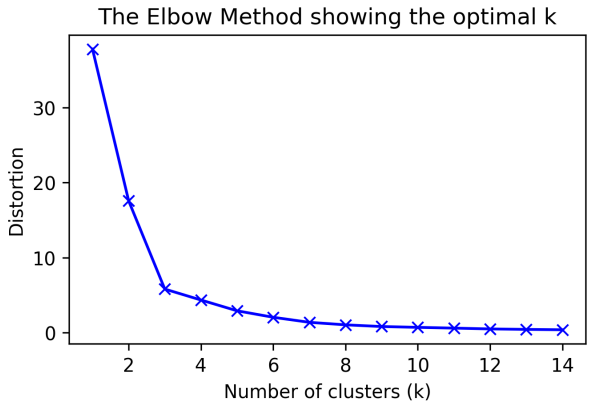

Figure 11: Elbow Plot

To find anomalous states, we constructed a graph of states where each edge represents the strength of cosine similarity between the two states. Then, we plotted the minimum spanning tree, which connects all nodes in the graph so that the total sum of edge weights is minimized. Figure 12 shows the MST graph plot. It can be seen that almost all of the edges are connected to Alabama. Therefore, this state can be considered anomalous in terms of the emotions hidden in its political meetings.

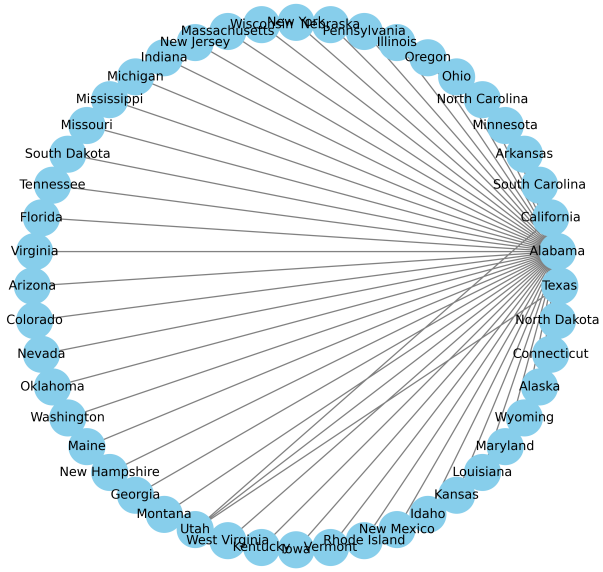

Figure 12: states MST Plot

4.4 State Prediction With BERT

In this section we predicted the state where each political meeting occurred, using only the caption text of each meeting. In order to model the problem, we used a multiclass classification approach. In order to create a realistic setup, we divided the training and testing by time. The training examples were taken from 2020 and 2021, and the test examples were taken from 2022. We filtered out states with fewer than 100 examples. As a result, there were 28460 training examples, 10057 testing examples, and 37 classes (states). A fine-tuning was performed on the pre-trained 'bert-base'

model, which was extracted from Hugging face. This pipeline was tailored to capture both the semantic richness of the political discourse and the unique characteristics of each state's political meetings.

Figure 15 (in the appendix) shows the confusion matrix of our model's prediction on the test set.The model demonstrated a high discriminative capacity, as evidenced by an average one-vs-all ROC AUC score of 0.983, indicating a strong ability to distinguish between states. Overall accuracy on the test set was 81.73%. The precision, recall, and F1-score varied across states, presented in the appendix2, reflecting the model's varying degrees of proficiency for different classes. For instance, states such as California, Illinois, and New Jersey achieved high precision and recall rates, suggesting the model's robustness in identifying unique features of their political discourse. Conversely, states with fewer samples or perhaps less distinctive textual features, like Connecticut, exhibited lower precision. The macro average precision and recall were around 0.80, pointing to balanced performance across all classes, which is particularly noteworthy given the class imbalances inherent in the dataset.

4.5 Votes Prediction with Demographic Features

In this section, we delve into the intricate relationship between racial demographics and the recent election voting patterns for the Democratic party. For each unique location where political gatherings took place, we collated data on general population counts across various years and juxtaposed it with the racial composition of these areas. Our hypothesis postulated that racial demographics might offer more predictive power than mere population metrics over the years. To validate this, we employed regression analysis complemented by SHAP plots, aiming to discern the relative importance of these features. We gauged the efficacy of our regression models—Decision Tree, Random Forest, and XGBoost—using metrics such as MSE, MAE, and R^2, to determine their aptitude in predicting vote shares for the Democratic party based on the aforementioned features. We conducted hyper parameter-tuning for each model with 10-fold cross validation. table 1 shows the results of each model over the metrics. Figure 13 shows the SHAP plots of the XGBOOST model. The Shap plots of the other 2 models can be viewed in the appendix.

Table 1: Evaluation metrics for different models

Model	MAE	MSE	R^2
Decision Tree	0.13	0.03	0.21
Random Forest	**0.12**	**0.03**	**0.32**
XGBoost	**0.12**	**0.03**	**0.33**

The three models—Decision Tree, Random Forest, and XGBoost—demonstrate comparable performance in predicting election vote results for the Democratic party, with Mean Squared Errors (MSE) of 0.03 and Mean Absolute Errors (MAE) around 0.12-0.13. XGBoost and Random Forest slightly edge out in terms of the coefficient of determination R^2, suggesting a better fit to the data's variance. Across all models, the racial demographic features (acs_2018_black, acs_2018_white, and acs_2018_hispanic) consistently rank as more influential based on SHAP values compared to general demographic features like population counts from different years. This prominence indicates that while overall population provides context, the racial composition is a more pivotal determinant in predicting Democratic party vote outcomes. In essence, the models emphasize the significance of racial demographics over general population metrics in forecasting election results for the Democratic party.

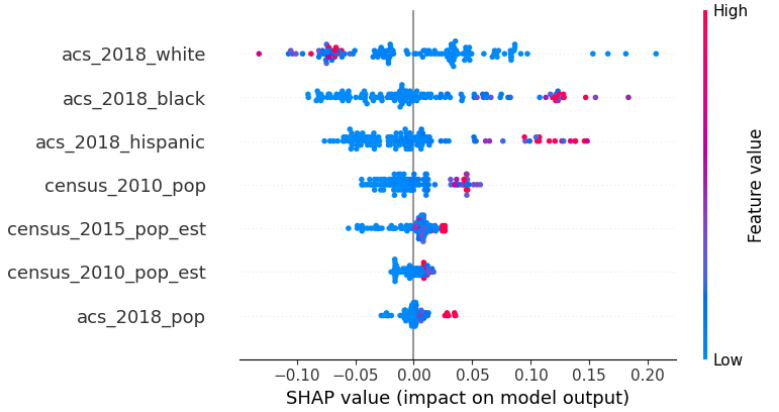

Figure 13: XGBOOST SHAP Plot

5 Conclusion & Future Work

In our study we embarked on a comprehensive exploration of the LOCALVIEW dataset to provided a unique lens to understand the political landscape in United States. Our methodology was multifaceted, combining natural language processing, machine learning, and visualization techniques. We employed the LDA model for topic modeling, uncovering underlying themes and analyzing discussions on equality. The emotion analysis was enriched with cluster analysis, dimensionality reduction, and graph analysis, offering a comprehensive view of the emotional undertones in the meetings. Notably, our state prediction, leveraging the BERT model, yielded remarkable results in pinpointing the origin state of each political meeting. Additionally, we delved deep into the relationship between racial demographics and recent election voting patterns, highlighting the significance of race demographic features against general time-varying demographic features. For future work, there are several avenues to enhance our study's methodologies and findings. One avenue is to enhance the LDA model by incorporating word embeddings that capture semantic similarities between terms. This could lead to more accurate topic assignments and interpretations. Exploring specific topics in greater detail or analyzing the sentiment of discussions around pivotal events or policy shifts could offer deeper insights. Moreover, investigating how geographic and social factors influence political discourse could provide intriguing insights into regional variations. In general, as the field of natural language processing and machine learning progresses, we anticipate improved accuracy and depth in our analyses, allowing us to uncover even more nuanced aspects of political dynamics.

6 Acknowledgment

The authors gratefully acknowledge the assistance provided by ChatGPT in generating and refining the content presented in this paper. For a more comprehensive analysis of the code underlying this study, please refer to this GitHub repository[1].

[1]https://github.com/BarAvraha/Project

References

[1] BEHESHTI, N. Topic modeling with political texts. *Towards Data Science* (2023). Accessed on [Insert Date Here].

[2] BERKOVITZ, S., MAZUZ, A., AND FIRE, M. Open framework for analyzing public parliaments data. *arXiv preprint arXiv:2210.00433v2 [cs.SI]* (2023).

[3] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] IS PLURAL, D. Data is plural - 2023-03-29 edition, 2023. Accessed on [Insert Date of Access].

[5] PRITCHARD, J. K., STEPHENS, M., AND DONNELLY, P. Inference of population structure using multilocus genotype data. *Genetics 155*, 2 (2000), 945–959.

[6] SIMKO, T. Localview, a database of public meetings for the study of local politics and policy-making in the united states. *Nature* (2023).

[7] SPECIFIED, N. Polibert: Classifying political social media messages with bert, 11 2020. Accessed: 15.8.2023.

7 Appendix

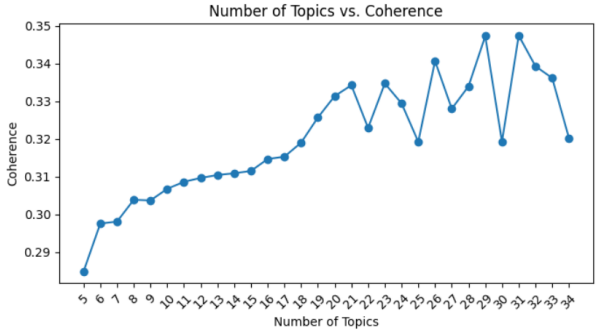

Figure 14: Coherence Score for Number of Topics

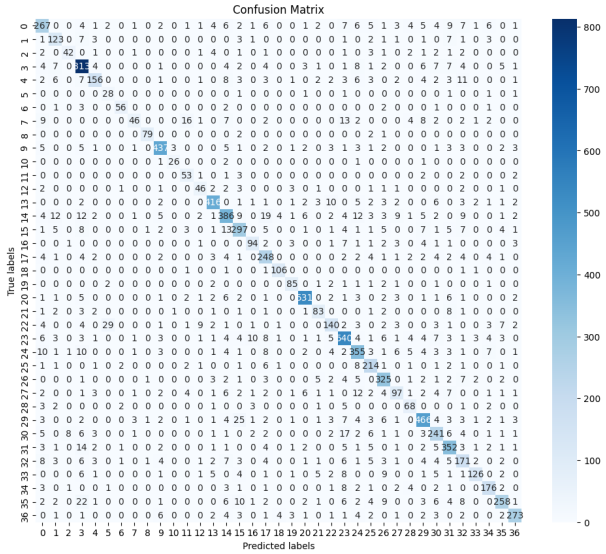

Figure 15: Confusion matrix of our predicted states

Table 2: Precision, recall, accuracy and F1 for achieved by the fine-tuned BERT on each state separately

	precision	recall	f1-score	support
Alabama	0.76	0.74	0.75	360
Arizona	0.72	0.79	0.75	155
Arkansas	0.79	0.64	0.71	66
California	0.86	0.92	0.89	885
Colorado	0.83	0.69	0.75	227
Connecticut	0.45	0.82	0.58	34
Florida	0.89	0.85	0.87	66
Georgia	0.81	0.40	0.53	115
Idaho	0.95	0.94	0.95	84
Illinois	0.91	0.90	0.91	485
Indiana	0.87	0.84	0.85	31
Iowa	0.63	0.83	0.72	64
Maine	0.68	0.69	0.68	67
Massachusetts	0.91	0.89	0.90	466
Michigan	0.79	0.75	0.77	517
Minnesota	0.77	0.79	0.78	374
Mississippi	0.81	0.75	0.78	125
Missouri	0.73	0.83	0.78	298
Nebraska	0.91	0.95	0.93	111
New Hampshire	0.83	0.85	0.84	100
New Jersey	0.94	0.92	0.93	575
New Mexico	0.76	0.77	0.76	108
New York	0.76	0.64	0.69	219
North Carolina	0.82	0.85	0.84	634
Ohio	0.78	0.81	0.80	437
Oklahoma	0.77	0.87	0.82	245
Oregon	0.82	0.88	0.85	369
Pennsylvania	0.69	0.58	0.63	166
South Carolina	0.70	0.77	0.74	88
South Dakota	0.86	0.84	0.85	554
Tennessee	0.74	0.77	0.75	315
Texas	0.79	0.86	0.82	411
Utah	0.65	0.69	0.67	249
Vermont	0.90	0.70	0.79	180
Virginia	0.81	0.84	0.83	209
Washington	0.87	0.74	0.80	351
Wisconsin	0.90	0.86	0.88	317
accuracy		0.82		10057
macro avg	0.80	0.79	0.79	10057
weighted avg	0.82	0.82	0.82	10057

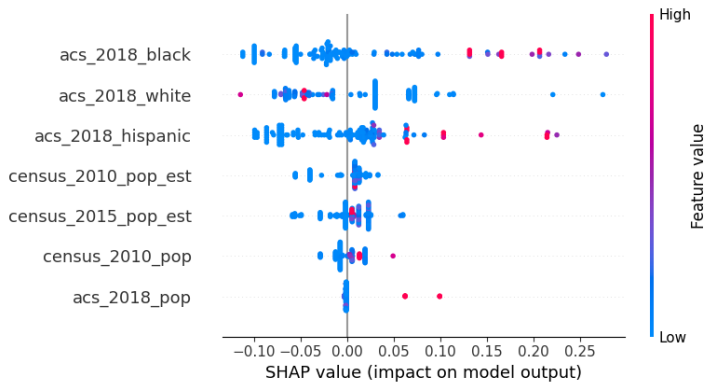

Figure 16: Decision Tree SHAP Plot

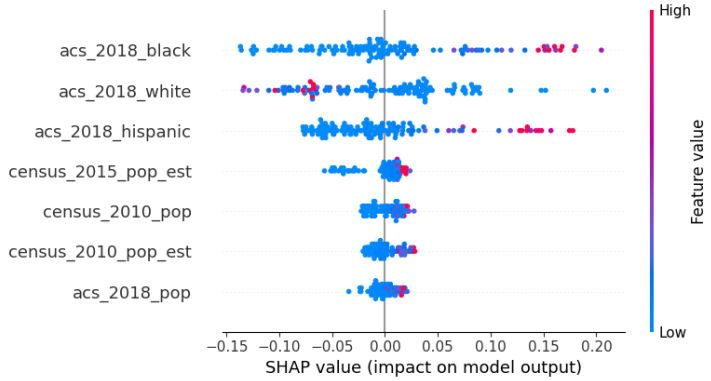

Figure 17: Random Forest SHAP Plot