Investigating United States Political Polarization via Natural Language Processing

—

***Charlie Mollin**, Cornell University, Ithaca, NY — cdm225@cornell.edu*
***Matthew Maitland**, Cornell University, Ithaca, NY — mjm638@cornell.edu*
***Seth Rizika**, Cornell University, Ithaca, NY — sr794@cornell.edu*

ABSTRACT Political polarization is a popular topic in political discourse right now. This paper takes an investigative approach to political polarization within the context of presidential speeches. Our methodology is rooted in NLP techniques: tokenization, word embeddings, bayesian probability, and zero-shot modeling. We expect to find notions of polarization in at least individual-party era analysis (not based on current definitions). Data is sourced from a dataset containing 992 presidential speeches spanning six party eras in the United States. Using this data, analysis on individual party era polarization and total polarization is measured both against current party definitions and individual party era definitions. Through this, we created a scoring mechanism rooted in tokenization and probability of tokens occurrence in respective party speeches. Both individual speeches and presidents were scored via this method. In zero-shot learning, we use OpenAI's GPT 3.5 Turbo model to classify speeches based on modern definitions of political party lines. Our results show some polarization; in aggregated measurements of polarization over time we find party skews in correct directions, confirming political polarization. However, in implementation of word embeddings, we do not find polarization. Zero shot modeling was effective in identifying specific eras where presidents from each party frequently gave speeches that more closely aligned with the sentiments of the opposite party. This enabled us to identify certain political, social, and economic explanations for the phenomena. Our findings highlight the nuances of both polarization and shifts in rhetoric over time, quantifying the complexity of political dynamics in U.S. history.

INTRODUCTION: Political polarization is a hot topic among political pundits, everyday citizens, and government decision makers in the United States. Markers of political polarization have increasingly been more salient in the United States lately, with little consensus between parties on many issues, like the border, immigration, and terrorism management. This leads to the question of how to quantify these trends. There are ways in which polarization can be derived; through congressional voting patterns, policy decisions, and other mechanisms. Presidential speeches, we noted, effectively represent party agendas overtime — e.g., a democratic president will give a speech that echoes their party platform. Speeches are inherently NLP friendly, and we found them suited for analysis to gauge the extent of political polarization.

This paper introduces methods of scoring via tokenization, word embeddings, and Z-score computation to grade speeches and presidents on a political spectrum. Additionally, we run Zero-Shot analysis through an OpenAI API call to predict speech parties via the GPT 3.5 Turbo model. Compared to existing literature, we found no research that uses our method in tokenization/speech scoring, however zero shot modeling is fairly common practice. In zero-shot learning, we leverage existing, pre-training large language models (LLMs) in order to generate labels for each data point without fine-tuning the model.

Existing research delves into analyzing

presidential speeches. We draw inspiration from [1] Benoit et. al in analyzing differences in word use between parties — a backbone of our scoring method. [2] Finity et. al (2021) uses direct NLP analysis, like tf-idf vectorization, to find stylistic sentiment differences in speech. We mirrored this study in some effect through vectorizer use to find semantic differences in speech via NLP. In [3] Liao et. al (2021), they take a sub-category and subtask approach to their sentiment analysis using RoBERTa, though this did not relate closely to our practice with presidential speeches, it inspired our party-era-specific analysis. [5] Puri et. al (2019) investigated zero-shot for text classification tasks using pre-trained language models. The study highlights the simplicity and accuracy of using such an approach for text-related tasks, even without any access to the model's training data. Finally, [6] Zavattaro et al (2015) analyzes speeches through machine learning in a lens of communication to the public. This inspired our motivation to use governmental communication — presidential speeches — that are readily available to the public.

Given change in topics over time, in each political party, we hypothesize that there will be noted polarization between the two parties — whether all time (aggregate) or in respective political eras. We investigate polarization over time both measured against current party definitions and individual party eras in expectation of finding polarization.

We note the importance of our work in that US government policy affects every American. This is direct research into policy and idea changes that affects the day-to-day of all Americans. Given our two party system, most Americans are affected by the policies and policy changes rooted in these speeches. Adjacent to other studies in political polarization — such as those with neural networks — our study contributes a new angle to political research.

This research contributes to the political science community, as it furthers knowledge in party change and evolution in the United States. Adjacent to other studies in political polarization our study measures political polarization from a different angle. There is no consensus method to measure political polarization, we found, so our research contributes to the definition.

METHODS: Our corpus is a dataset of 992 presidential speeches. Speeches span across all political eras and presidents, however speech counts are not equal among presidents. The dataset was last updated four years ago, the temporal coverage start date is 04/29/1789, and the end date is 09/24/2019 — ending with the beginning of the Trump administration. The dataset includes individualized datasets for each party era:

- *The First Party Era: 1792 - 1824*
- *The Second Party Era: 1828 - 1854*
- *The Third Party Era: 1854 - 1895*
- *The Fourth Party Era: 1896 - 1932*
- *The Fifth Party Era: 1932 - 1964*
- *The Sixth Party Era: 1964 - 2019*

Rows in the datasets contain the date of the speech given, the president, the party, speech title, summary, transcript, and the URL source of the speech. Despite not having all presidential speeches, we validate the representativeness of the set as we use speeches found in public domain.

Our methodology rests in four main steps:

1. Tokenization scoring with prior of current party definitions
2. Tokenization scoring with no prior, tokens considered on party era at hand
3. Tokenization scoring via word embeddings (spaCy)
4. Zero-shot learning

Tokenization Scoring:

For the first step of our research we developed a tokenization and dictionary based scoring mechanism. Having reviewed our dataset through exploratory data analysis and crafted our stated question, scoring presidents and their speeches aligned best with our investigation. Such analysis contributes to polarization review over time, as each president is measured as aligning with their party or not. Initially, we separate speeches by party era and party (democratic or republican). We then call a countvectorizer, strip accents, and remove stop words. Then the countvectorizer yields matrices for both parties, these matrices are v-stacked then fed forward into a *bayes_compare_language* function to compare differences between each party's vocabulary across speeches. The function, sourced from [4] *FightingWords* examines the usage rate of each word or n-gram as opposed to raw counts, then uses a smoothing dirichlet distribution prior on the vocabulary items to analyze how one party uses a word more than another. It outputs the associated z-score of a word, showing that the word is used more by one party than the other.

Additionally, we incorporate an informative prior of the sixth party era speeches. This allows for guidance in probability analysis and influence of our prior, aligning our results alongside current party definitions — to see how parties have changed, or polarized, relative to current party definitions. In the second step we do not incorporate a prior (or a non-informative prior), thus likelihood of tokens is solely driven by the data at hand, giving us another angle to analyze polarization.

We then build 200 token dictionaries for each party based on the most salient terms found in each through the *bayes_compare_language* results — using 200 terms from republican and democrat results, respectively. Then, simply, we tokenize and loop through all speeches. Speeches are analyzed token by token, if the token is found in either the republican or democratic dictionary, the score of that token is summed along with other found tokens to yield a score for the speech.

> *Democrat tokens have a positive z-score, whereas republican tokens have a negative z-score.*

Summed scores are normalized by the length of the speech and multiplied by 100. We then compile scatterplots, points representing speeches, and flag misclassified speeches — e.g. a republican speech is labeled as democrat. Additionally, we create Kernel Density Estimation (KDE) plots for both party eras and aggregate (adding all party eras together) to visualize polarization over time. KDE plots estimate probability densities, yielding a smoother and more interpretable graph.

In our third step, we analyze our same party dictionaries within word embeddings, via the spaCy medium model. We found this step necessary as before we only considered the token itself, so now we could analyze different usages and contexts of words in speeches. Using our created party dictionaries, we find word embeddings for each term and find the mean across all vectors in respective party dictionaries. From there, we perform the same analysis on the speech at hand, computing the cosine similarity of the speech to each of the dictionaries. Speeches more similar to the republican dictionary were assigned a negative value — for the sake of differentiating the two — and those more similar to democratic were positive. Scores returned were the maxima between the two computed similarities.

Zero-Shot Learning:

The next method, and fourth step, was zero-shot learning. Zero-shot is an approach in NLP that allows models to make predictions for tasks they haven't been explicitly trained on. Unlike traditional supervised learning methods that require extensive labeled datasets, zero-shot learning leverages pre-trained models

to classify a dataset based on context and general knowledge from its training corpus. This method is particularly useful for our study as it enables the classification of presidential speeches according to modern political party lines without the need for specific training data labeled by political affiliation.

By employing zero-shot learning, we can efficiently analyze large volumes of text and identify patterns of political polarization over time. The application of zero-shot learning in this context helps us uncover nuanced insights into the political landscape, providing a mechanism for understanding context-dependent party trends.

To prepare our data, we first had to limit the scope of our dataset. In the initial corpus of 992 speeches, many were neither Democratic nor Republican. Since we are focused mainly on these categories, we dropped any speeches that came from presidents who were not affiliated with either of these parties, leaving us with 867 speeches. The majority of the speeches dropped came from the early party systems. Because we leveraged the OpenAI API, additional preprocessing steps such as tokenization or detailed normalization were not required. The API's built-in capabilities handle these tasks, allowing us to feed the text directly into the model. However, GPT 3.5 Turbo has a 16K token maximum context length, and there were two speeches that exceeded this. In order to maximize the number of tokens we could feed into the model, we used the GPT2 Tokenizer in order to truncate the necessary speeches.

We initially attempted to use the Flan-T5-Large model for our analysis. To test the model, we created two separate datasets. The first dataset was a random sample from our exclusively Democratic/Republican dataset. The second dataset included all speeches from 2000 to the present.

Since we are most interested in the labels that were predicted incorrectly, evaluating the model's accuracy on a single random dataset alone would not be an effective metric. As a result, we deemed a model successful if it demonstrated non-random performance on the fully representative sample dataset and showed significant improvement in accuracy on the modern dataset. Given our focus on the modern definitions of political parties, better performance on the recent speeches implies that the model accurately understands contemporary political affiliations.

We used two variations of prompts in our model evaluation:

1. *Which political party does this speech most closely align with?*
   [Transcript]
   Choices: Democratic or Republican
   Answer:
2. *Does this speech lean more liberal or conservative by modern standards?*
   [Transcript]
   Choices: Liberal or Conservative
   Answer:

The Flan-T5-Large model has a maximum token limit of 512 tokens. As a result, we attempted to feed different sections of speeches into the model, hoping to capture the most salient aspects of the speeches. We tested feeding the first 512 tokens, the last 512 tokens, and the middle 512 tokens of each speech.

Despite the variety of our approaches, including different combinations of input data, prompt selection, and token sections, the Flan-T5-Large model's performance was roughly random across all tests. Therefore, it did not meet the performance criteria we required for our study.

We then determined that our task would likely require a model that can handle all, or almost all of the tokens from any given speech in order to make a proper classification. We experimented briefly with LLaMA models,

designed for longer texts, but experienced similar results to the Flan model.

As we questioned whether zero-shot would be a suitable approach for our task, we found initial success manually inputting our prompted text into the ChatGPT 4 user interface. We used the same prompts and evaluation datasets via an OpenAI API call, using GPT 3.5 Turbo. This method also enabled us to provide a "role" to the model, which, in our case, was:

"You analyze political text and only answer with one word from the given choices."

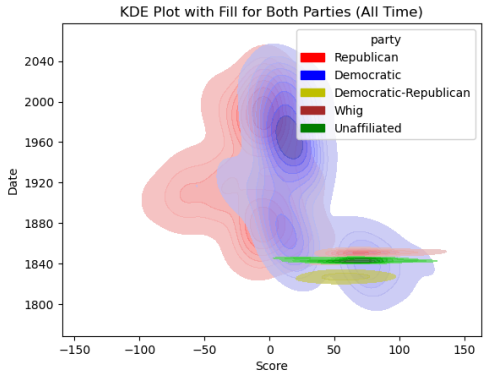
Using this model, we were able to achieve non-random performance when using prompt option 1 across both datasets (71% accuracy), and saw significant improvement from the test dataset that is representative of the entire dataset to the modern test dataset (85% accuracy). These scores met our predetermined criteria for model selection. We did not achieve this criteria for prompt 2, so we elected to only use prompt 1 in our final implementation.

RESULTS:

Tokenization Scoring:

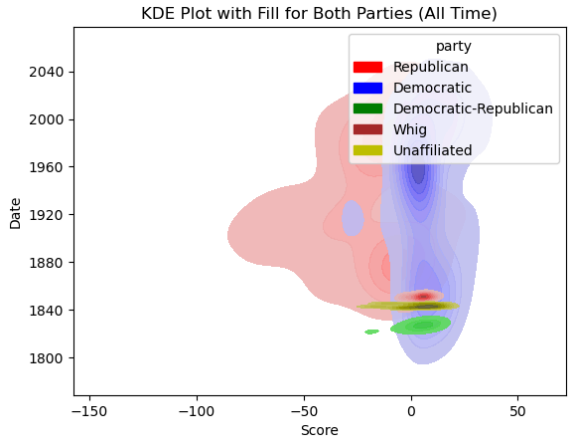
In our tokenization scoring we returned notable trends. In step 1, we found all republican speeches to have a left skew of -1.03, and democratic speeches to have a right skew of 1.28. These skews go in the respective directions of their party values. In terms of summary statistics in classification of speeches and president scores, six republican presidents — Dwight Eisenhower, Ulysses S Grant, James Garfield, Warren Harding, and Abraham Lincoln — scored as democrats. One democrat, Woodrow Wilson, scored as a republican. 29% of republican speeches were classified as democratic, all republican speeches had a median score of -6.8. 12% of democrat speeches were misclassified, and the median speech score for democrats was 15.45. In *Figure U* we note slight divergence of the two parties initially and more convergence later on.

Figure U



In our second step, with no informative prior, we returned a left skew of -1.24 for republican speeches and a left skew of -1.41 for democratic speeches. Democratic speeches had a median score of 5.48, while republican speeches had a median score of -11.02. 16.7% of democratic speeches were classified as republican and 18.1% of republican speeches were classified as democratic. One democratic president, Woodrow Wilson, was predicted as republican. One republican president, Warren G Harding, was predicted as democratic. In *Figure V,* we note slight divergence of the parties later, but a relatively clear split between the two.

Figure V



In the third step, with use of spaCy medium, we returned a right skew of .03 for republican

speeches and a left skew of -.09 for democratic speeches. Republican speeches had a median 72% cosine similarity to the republican dictionary, whereas democratic speeches had a median 74% cosine similarity to the democratic dictionary. 49% of republican speeches were misclassified as democratic and 29% of democratic speeches were misclassified as republican. Seven republican presidents were predicted as democrats: *Dwight Eisenhower, Richard Nixon, Gerald Ford, Ronald Reagan, George H.W. Bush, George W. Bush, and Donald Trump.* Five democratic presidents were predicted as republicans: *Franklin Pierce, James Buchanan, Andrew Johnson, Grover Cleveland, and Woodrow Wilson.* In *Figure W* we see overlap of parties throughout most eras.

Figure W

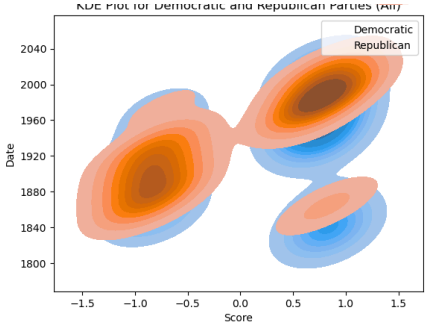


Figure WW

S	*Republican*				*Democratic*			
	Sk	*Wr*	*P*	*Md*	*Sk*	*Wr*	*P*	*Md*
1	*-1.03*	*29%*	*5*	*-6.8*	*1.28*	*12%*	*1*	*15.5*
2	*-1.24*	*18.1%*	*1*	*-11*	*-1.4*	*16.7%*	*1*	*5.48*
3	*.03*	*49%*	*7*	*72&*	*-.9*	*29%*	*5*	*74%*

Zero-Shot Learning:

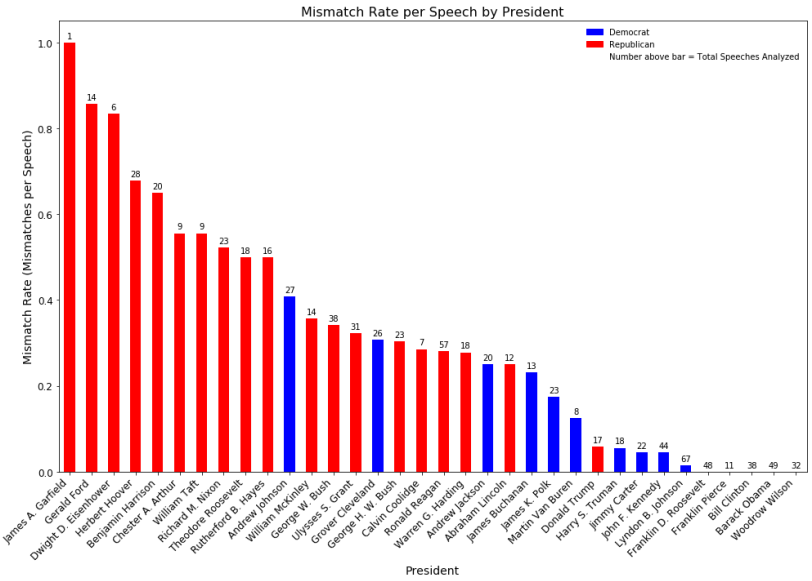
Consistent with our model evaluation score across the randomly sampled dataset, we obtained a 71% accuracy score when we applied the GPT 3.5 Turbo model across the entire dataset (of only Democratic or Republican speeches, 867 speeches total). We expected the model to only output the terms “Democratic” or “Republican”, but there were some erroneous outputs, such as “Federalist”, “Union”, “Congress”, “Territory”, “Administration”, and “China”. There were 60 such instances, so, satisfied with our sample size, we dropped these outputs from our evaluation, leaving us with 807 classified speeches. After this modification, our accuracy across the dataset increased to 77%.

For the evaluation of our zero-shot results, we focus our discussion on the 23% (187 total) of speeches that were misclassified by our model. All of these speeches were either classified as Republican while they truthfully came from a Democrat, or vice versa.

Of these misclassified speeches, 80.3% (150 total) were truthfully Republican but predicted Democratic (Figure X).

Due to differences in the quantities of speeches available to us by each president, we normalized our data to analyze the rate at which each president’s speeches were misclassified. Consistent with the finding that the majority of misclassified speeches were Republican labeled Democratic, we found that the 10 presidents who had the highest rate of misclassified

Figure Y



speeches were true Republican. These presidents, in descending order of misclassification rate, are *James A. Garfield, Gerald Ford, Dwight D. Eisenhower, Herbert Hoover, Chester A. Arthur, William Taft, Richard M. Nixon, Theodore Roosevelt and Rutherford B. Hayes.* It is important to note, however, that we have a limited number of speeches for these presidents. For example, 100% of Garfield's speeches were misclassified, since we only have one speech from him, which was labeled as Democratic. The Democratic presidents with the highest rate of mislabeled speeches were *Andrew Johnson, Grover Cleveland, and Andrew Jackson.* (Figure Y) To better support our primary task, we can organize these results in chronological order of presidential term. In doing this, we find that the majority of truthfully Democratic misclassifications took place between the terms of Andrew Jackson (1829-1837), and Grover Cleveland (1885-1889 and 1893-1897 due to non-consecutive terms). After Cleveland's terms, Democratic presidents gave speeches that leaned Republican by modern standards at a significantly lower rate.

The trends relating to the Republican presidents are less clear. We see an initial spike in misclassified speeches during Garfield and Chester A. Arthur's tenure. The rate of misclassified speeches then decreases after Taft. There is then another spike from Hoover to Ford, before tapering back down once again as we near the present-day (Figure Z).

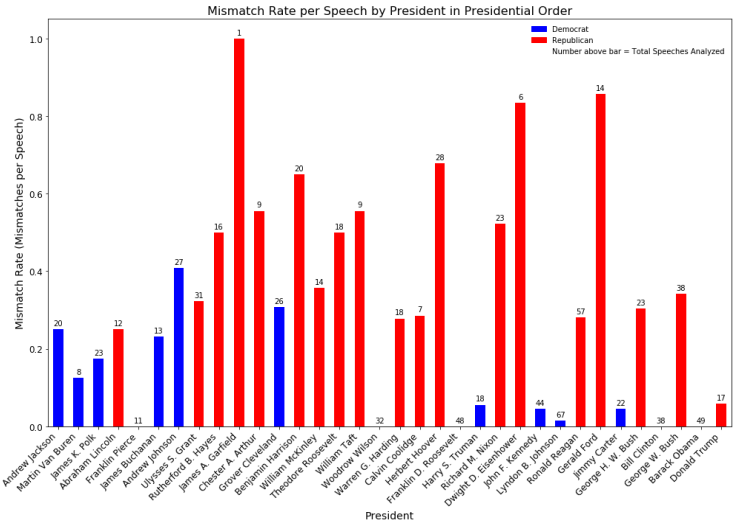
DISCUSSION:

Tokenization Scoring: The three steps of our tokenization scoring method show noted results in polarization.

Step 1:

In analysis of speeches against current party definitions, we find skews of -1.03 for republican speeches and 1.28 for democratic speeches. Given that democratic scores are positive, and republican negative, these skews show a shift over time in scores along party lines — effectively representing polarization. Figure U shows the assortment of speeches in both parties. From the beginning democratic speeches skew much more democratic and near/overlay republican speeches over time. However they maintain their position in the

Figure Z



positive scores, showing polarization and general opposite skew to the republican. As in Figure WW, step 1 has the strongest skew values compared to the other steps, and good speech median scores -6.8 (republican) and 15.5 (democratic), a difference of 22.3. Misclassification rates are relatively low at 29% for republican and 12% for democrat, however they are not the best among all steps. Additionally, five republican presidents are misclassified, three of which served from 1865-1881 (Lincoln, Grant, and Hayes), showing a liberal shift in republican policy during this time period — likely related to policies such as reconstruction and the emancipation proclamation. Overall, these numbers indicate polarization and confirm our hypothesis.

Step 2:

In step 2, using a non-informative prior, we return strong values but don't find overall polarization, rejecting our hypothesis. Figure V shows parties that are similar over time, however the democratic party has fairly low extreme values relative to the republican and skews slightly left. The republican skew of -1.24 shows a skew in the correct direction for republican speeches, but the democratic skew of -1.41 (Figure WW) shows democratic speeches skew conservative over time — rejecting polarization, as we expect to see diverging skews. Step 2 returns low misclassification rates and incidents of president misclassification, showing that presidents largely stay along party lines over time — which we'd expect using a non-informative prior. A difference of 16.48 in median scores can explain why we find no polarization here — compared to step one's difference of 22.3. Interestingly, the one democratic president both this step and the first step misclassified was Woodrow Wilson. Overall, there is only one democrat and one republican misclassified, suggesting no trends. Thus, our results indicate no polarization.

Step 3:

In step three, working with word embeddings, we again find no polarization. In this step we were largely in the dark as to how spaCy medium was trained, so we had no clear anticipation of results. In Figure W, the results are not telling, as both parties overlay each other. Compared to the other steps in Figure WW, step three has least telling and most

"incorrect" values. Both skews go in the incorrect party direction at .03 and -.9. There are seven misclassified republican presidents (49% speeches misclassified) and five misclassified democratic presidents (29% speeches misclassified). Misclassified democrats served from 1853-1921, showing a potential shift in party platform then potentially due to the civil war, reconstruction, and WWI — interestingly, step three also scores Woodrow Wilson as a republican. Misclassified republicans go from Dwight D Eisenhower up to Donald Trump, inclusive of all republicans between them. This could show a shift in republican policy. Our lack of knowledge of spaCy medium's training makes these results less verifiable and interpretable. Overall, step three rejects our hypothesis.

Zero-Shot Learning:

Through our zero-shot learning methodology, we found spikes in the rate at which Republican speeches were misclassified primarily in two distinct time periods: between Garfield (1881) and Taft (1913), and then again between Hoover (1929) and Ford (1977). This first spike, in the late 19th and early 20th century, can be explained by Industrialization and the Progressive Movement. There was rapid industrial growth, leading to significant economic change. The Republican Party, which traditionally supported business interests, had to adapt to the evolving economy. Issues such as trust-busting arose, and presidents during this era were tasked with addressing corruption and the excesses of the new industrial economy. Additionally, during the rise of the Progressive Movement, presidents were tasked with addressing social injustices and regulating big businesses. These shifts in focus and policy likely contributed to the higher misclassification rates, as the rhetoric during these periods more closely aligned with what would be considered modern Democratic principles according to both logic and our zero-shot model.

Similarly, we can map the era with the highest misclassification rates among Democratic presidents to historical trends. According to Figure Z, this era spans from Andrew Jackson's term (1829) to Grover Cleveland's (1897). Within this era are the years leading up to the Civil War, during which the Democratic Party was divided, with Northern and Southern Democrats typically advocating for different policies. This internal division could explain the high mismatch rate, as presidents might have tried to appeal to both groups. Also within our specified era, we see Reconstruction take place, where significant changes aimed at limiting federal intervention were enacted – policies that do not align with modern Democratic principles. Thereafter, during the Gilded Age, was rapid economic growth, industrialization, and corruption. Democratic presidents faced pressures to address labor rights, economic regulation, and corporate power. Many presidents during this time, such as Grover Cleveland, advocated for limited government intervention, which aligns more with conservative principles based on modern definitions.

These findings are particularly interesting, but there are some limitations based on our dataset and the zero-shot approach. In regard to our dataset, it is important to note that we have access to significantly more speeches for recent presidents. As a result, our findings likely become more reliable as our analysis moves closer to the modern era. Additionally, in each of our iterations of developing the model, including the final model selected, we found better performance on democratic speeches. This could represent a trend where Republican speeches tend to regularly include words, phrases, or ideas that resonate with the Democratic party, while Democratic speeches less frequently include similar entities that align

with Republican ideals. The earlier analysis and discussions surrounding the Republican misclassifications continue to be a topic of interest, but it is important to note that some of the seemingly salient trends could be a result of these limitations.

CONCLUSION:

This paper investigates the evolution of political polarization in the United States through the analysis of presidential speeches using NLP techniques. Our analysis revealed several key findings, including evidence of polarization, and eras of US History where presidents often gave speeches that do not align with their party affiliation, based on modern day party definitions.

Our analysis is limited by a few factors. First, the varying numbers of speeches available from different presidents, as well as the fact that the current administration is not included. Since we are concerned with the modern stances of each party, speeches from Joe Biden could be useful to create the most accurate modern party definition. Additionally, we rely on specific NLP methods, each with inherent biases and assumptions, which may have influenced the detection of polarization trends and other results. For example, the training specifics of spaCy's medium model and the preset token limits in zero-shot learning. Additionally, for our zero-shot approach, we are reliant on GPT 3.5 Turbo's training data alone, since zero-shot learning does not involve model fine-tuning.

Further research should consider updating the dataset to include recent speeches and possibly incorporate speeches from a wider range of political figures. It would also be beneficial to refine NLP techniques, perhaps by developing custom models trained specifically to recognize historical shifts in political language, thereby improving accuracy in detecting polarization. Additionally, shifting our tokenization scoring parameters could yield better results. It could also prove beneficial to leverage a more powerful model for zero-shot learning, such as GPT 4, which could improve accuracy and reduce noise, particularly within the Republican Party.

REFERENCES

[1] Benoit, W., Goode, J., Whalen, S., & Pier, P. (2016). "I am a candidate for president": A functional analysis of presidential announcement speeches, 1960-2004. Speaker & Gavel, 45(1). https://cornerstone.lib.mnsu.edu/speaker-gavel/vol45/iss1/3

[2] Finity, K., Garg, R., & McGaw, M. (2021). A text analysis of the 2020 U.S. Presidential Election campaign speeches. In 2021 Systems and Information Engineering Design Symposium (SIEDS) (pp. 1-6). Charlottesville, VA, USA. https://doi.org/10.1109/SIEDS52267.2021.9483735

[3] Liao, W., Zeng, B., Yin, X. et al. An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa. Appl Intell 51, 3522–3533 (2021). https://doi.org/10.1007/s10489-020-019641

[4] Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis*, vol. 16, no. 4, 2008, pp. 372-403. SPM-PMSAPSA.

[5] Puri, R., & Catanzaro, B. (2019). Zero-shot text classification with generative language models. In 3rd Workshop on Meta-Learning at NeurIPS 2019. https://doi.org/10.48550/arXiv.1912.10165

[6] Zavattaro, S. M., French, P. E., & Mohanty, S. D. (2015). A sentiment analysis of U.S. local government tweets: The connection between tone and citizen involvement. Government Information Quarterly, 32(3),

333–341.
https://doi.org/10.1016/j.giq.2015.03.003