Predicting President Party Affiliation of State of the Union Addresses

Molly Hutchinson msh87409@uga.edu
Frank Wills frank.wills@uga.edu
Department of Computer Science, University of Georgia

Abstract

Each year of a US President's term in office they give the State of the Union Address to Congress, detailing information about the nation and the President's agenda. We trained several classifiers on a corpus of State of the Union addresses to predict the political party affiliation of the president giving that address. Through these models we hoped to examine whether it was possible to determine the party of a president based on their addresses and to compare how each model performed on a generally formal corpus such as the State of the Union address. We found promising results, with accuracies consistently well above random chance.

1. Introduction

The State of the Union Address is an annual report by the president of the United States given to the U.S. Congress near the beginning of the calendar year. Initially, it started as a written report but eventually evolved with the help of radio and television to the live broadcast format. The speech provides the opportunity for a president to review the prior year, give a preview of their agenda for the upcoming year, and give an update on the current state of the nation.

In this project we hoped to take transcripts from State of the Union addresses and analyze their word use to attempt to predict the political party of the president giving that address. There has been a lot of work on language processing to pick up on partisan bias in less formal political discourse, but in our project we hoped to analyze more formal speech. One possible use for the models in the future could be to analyze new State of the Union addresses to detect whether a speech leaned more towards a certain political party in the event that a president were to be relatively non-partisan. Additionally, it could also help to provide insights on other speeches or works – not specifically the State of the Union address – and could even be used to compare and contrast current US political parties with other non-US political parties or even future parties. As the United States political partisanship starts to extend to have more of a global reach, the ability to track US partisanship is becoming increasingly important.

We obtained the transcripts from the "State of the Union Corpus (1790 - 2018)" Kaggle dataset from Rachael Tatman, which contained only the text files for the addresses. From there, we created a spreadsheet that matched the files containing the transcript text with columns that contained the year the address was, the president it was associated with, and the political party of that president. In total, the number of State of the Union addresses in our dataset was 227.

To become usable with both our Naive Bayes and K-Nearest Neighbor models, we first needed to process our transcripts. First, they were tokenized and divided into chunks of 1,000 tokens to be fed to the classifier. Stopwords and punctuation were also cleaned out of the text so that they would not interfere with the training of the model. In some experiments, the political parties present in the dataset that were not 'Republican' or 'Democrat' were moved into a category titled, 'Other'.

Several experiments were run utilizing the Naive Bayes Multinomial and Bernoulli classifiers as well as the K-Nearest Neighbors Brute Force, KD-Tree, and Ball Tree algorithms. Some experiments ran with grouped categories, 'Democrat', 'Republican', and 'Other', while some ran with all of the presidential party categories including 'Democrat', 'Republican', 'Democratic-Republican', 'Federalist', 'Whig', and 'Unaligned'. In total, 10 different experiments were run; experiments consisted of 5 experiments for each Naive Bayes or K-Nearest Neighbor model with grouped categories and 5 experiments for each model with ungrouped categories.

Our model accuracies were all fairly high, all performing well above random chance. Accuracies for all of experiments that were performed are summarized in Table 1. Table 2 shows our True Positive rates, which were all well above their associated proportions of the corpus and test set, showing that our model performed very well on our dataset, even if some 'Other' True Positive rates suffered from an overall lack of data associated with them. Even still, our models performed well. The best accuracy we found from our testing was from our grouped Multinomial model, which achieved an accuracy of 86.8%. Overall, our conclusion was that our experiments were effective at predicting political party given a State of the Union address. Better balancing of the testing and training data party proportions may be necessary given our relatively small selection of data, but overall our model is generally effective at predicting parties based off of State of the Union speech content.

The remainder of this paper will provide more detail on our data, the processing of our data, how our experiments were developed and how they performed, and the results found in these experiments.

2. Related Work

Our work falls under the umbrella of linguistic analysis of partisan political discourse. This is a widely studied area, especially gaining traction in recent years as partisanship becomes an increasingly important part of the United States political scene.

Professor Jacques Savoy has been an important force in the field of linguistics analysis for several decades. In his 2016 paper "Vocabulary Growth Study: An Example with the State of the Union Addresses"(Savoy, 2015), he analyzes the observed "vocabulary growth" of the speeches against the expected vocabulary growth based on linguistics at the time. Vocabulary growth is the number of new words introduced by the President in each address that were not used in previous addresses. What Professor Savoy found is that there were broad ranges of time where fewer words were introduced than expected punctuated by eras where more words were introduced than expected. The eras of high vocabulary growth tended to coincide with major shifts in American History. President Franklin Delano Roosevelt for instance served during an incredibly tumultuous time in American history and had a higher vocabulary growth than expected. Meanwhile the presidents of the early to mid 1800's saw a low vocabulary growth which Professor Savoy tied to these presidents giving similar arguments and generally expressing similar rhetoric. Professor Savoy also mentions that it is possible that the differences in political opinion between the first two presidents (George Washington and John Adams) and the second two presidents (Thomas Jefferson and James Madison) could have contributed to a high vocabulary growth between them. However, this growth was not detected in the results of his experiment and was determined to not be an explanatory factor. This is relevant to our research as the effect of a president's political views and affiliations on the text content of their addresses is important to the effectiveness of our models. If there is not a significant linguistic difference between the differing parties then the models should be incapable of distinguishing between them. Professor Savoy's work also demonstrates that through vocabulary growth the State of the Union addresses gradually introduce new tokens rather than completely changing the vocabulary or remaining stagnant between addresses. This continuity is important for the functioning of the models, as without some variation and some overlap a model has no footholds to learn and evaluate a corpus. Professor Savoy's work differs from ours in that, while he touched on differing opinions, his work does not directly engage the topic of partisanship. It covers a general linguistic analysis of the State of the Union addresses, but does not go any further in-depth into the specific attributes of a presidency than mentioning general policy topics. His work also applies mostly retroactively and focuses on the historical relevance of the model output, relying heavily on the context and time period of the address. While our models are also tied to the time period and history of the US, the political partisanship has slightly less to do with any given context because periods of US political party control are relatively evenly spaced. Neither Democrats nor Republicans have held the presidency for more than four consecutive terms, and, as Professor Savoy mentions, that sixteen year period starting with President Lincoln was generally repetitive in terms of State of the Union speeches. As such, our models offer the opportunity to be applied more readily to future addresses without the retrospective knowledge of historical context.

In the paper "Evolution of U.S. Presidential Discourse over 230 Years: A Psycholinguistic Perspective", Xueliang Chen and Jie Hu investigate the linguistic attributes of a large corpus of US presidential speeches and discourse(Chen & Hu, 2019). This corpus consists of The Grammar Lab's collection of speech transcriptions for each president from Washington to Obama, amounting to over 3 million words and including the State of the Union addresses. They implement the Linguistic Inquiry and Word Count (LIWC) program to parse through the corpus and determine how frequency of different psycholinguistic techniques appear. LIWC attempts to track a wide variety of psychological states, such as how often a text appeals to authenticity or whether it focuses on the future, present, or past. The process for doing so is exceptionally complex and involves encoding years of linguistic theory into an automated process. Most importantly to our work however, Dr. Chen and Dr. Hu analyze these psychological state frequencies with respect to the party of the president in question. This allows them to compare the trends of Republican and Democrat presidents to use certain psychological states more often than the other, as well as how to track how the two comparatively have changed throughout the past 230 years. What they found through the LIWC analysis is that there are some differences between Republican and Democrat presidential rhetoric, particularly in that authenticity has been gradually growing in Democrats over the years while Republicans have only seen a spike in authenticity in recent years. This difference could be key to whether our models can distinguish between the parties, although our models approach the text in different ways. LIWC has the benefit of analyzing words in the context of the speech, while our models disregard the order or placement of the words. LIWC was built by classifying words into categories somewhat manually and relies on several sub-classifiers to analyze the text, while our models are purely trained through a single layer and automated fit against the corpus. The reduced complexity may influence our model's ability to pick up on the subtle differences in psychological states, but we are also only looking for one metric so the reduced complexity of our goal may counter that out. Overall our research will hopefully show if the subtle differences shown to be present by Dr. Chen and Dr. Hu can be detected through classification models in a way that differentiates political party.

In their paper ”Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014”, Alix Rule, Jean-Philippe Cointet, and Peter S. Bearman propose a method of tracking lines of discourse and shifts of priority across the history of the State of the Union(Rule, Cointet, & Bearman, 2015). Their method is designed to be applied in an automated fashion, analyzing the text content of the addresses in context. Their method also makes use of co-occurrence tracking, where tokens are given different classifications based on the co-occurrence of other relevant words. The example they give is how ”confederacy” appears with different words before and after the events leading up to the civil war. Prior to the war the word was used to describe the collective state governments that preceded the US and was seen with words related to that, while during and after the war it related to the seceded states and was therefore occurring alongside words like ”union”. The co-occurrences are used to build network clusters and group related communities. Their algorithm's results divide the history of the address into long periods they call durées where discourse remains fairly consistent. They determined that the start of the shift towards what most resembles modern political discourse really began roughly around 1917 with the entry into World War 1. They also found that generally liberal ideologies that dominated the mid to late 1900's and 2000's began entering discourse intermittently before then, but weren't solidified until later in the century. Another major discovery through their model was that the State of the Union address has remained fairly stable with regards to sentiment throughout the years, changing in long, gradual shifts. It is uncommon for there to be a huge jump between years. They build river networks to further interpret and track the flow of discourse, showing how different durées of discourse relate to each other sequentially. Our work relies on these co-occurrences and shifts in discourse being consistent within parties. The analysis of the changing meaning of a term or changing party opinion of a topic could be a threat to the performance of our model. The change is not necessarily bad so long as the parties remain distinct and do not overlap too much across time. As opposed to their method, which looks at eras of discourse relatively independent from party, our models hope to detect which party the discourse belongs to. If streams of discourse are too similar between parties, our models will struggle. Their work also purely applies in a sequential sense, being directly linked to the change over time. Our work also must consider the time period, but only in the context of how the parties might have changed throughout in their vocabulary use and rhetoric.

The paper ”Predicting Political Party Affiliation from Text” by Felix Biessmann, Pola Lehmann, Daniel Kirsch, and Sebastian Schelter is probably the closest analog to our research(Biessmann, Lehmann, Kirsch, & Schelter, 2016). The authors examine the ability of a linear regression model to predict political leanings and affiliations of text gathered from speeches, Facebook posts, and manifestos on a corpus of German text. Unlike the US, German politics involves many more active parties, and the parliamentary system can lead to different complex situations than the US three-branch system. The texts are analyzed with sentiment analysis, which divides key words into different sentiments. This allows the models to be trained on a somewhat abstracted corpus of information and diminishes the effect of different vocabularies across text. Linear classifiers were then trained and the results recorded for each category. What they found is that the classification of party affiliation could be done with an accuracy higher than random chance and that the accuracy was not significantly diminished depending on the medium from which the text was gathered. They also determined that the length of the text was important to the accuracy of the models. One important note that is made is that the classifiers fall under the bag-of-words approach, which assumes features are uncorrelated. This means the content must be interpreted given the caveat that the corpus of text violates this rule and is assuming the effect will not invalidate the results. Overall the authors of this paper chose to take a wide swath of sources across a very small time frame, while our research hopes to take a very specific source across a wide time frame. The wide variety of parties in German politics also makes comparing their results to our mostly bipartisan government difficult. Still, their research is relevant in that the techniques they use are similar and the application is nearly the same. We do not directly use sentiment analysis in our research, but as shown in the paper by Dr. Hu and Dr. Chen, the sentiments of American political parties only slightly diverge (Chen & Hu, 2019). The sentiment analysis step may or may not inhibit the model's effectiveness. Finally, it is important to note that the text in this research was in a different language, although there should be many linguistic parallels between German and English.

Continuing in the vein of foreign partisan analysis, ”Predicting Party Affiliations from European Parliament Debates” by Bjørn Høyland, Jean-François Godbout, Emanuele Lapponi, and Erik Velldal covers their process on developing SVM models to predict party affiliation from discourse in the European Union Parliament(Høyland, Godbout, Lapponi, & Velldal, 2014). The study restricted the corpus to speeches given by members of the 5th and 6th European Union Parliament which spanned the terms from 1999 to 2004 and 2004 to 2009. An SVM was trained on the 5th Parliament corpus using the 6 main political parties of the time as classes. The SVM classifier was then evaluated using the corpus of 6th parliament speeches and got decent results, scoring total accuracies of around 0.55 which are good for a six party system that is more evenly distributed than the US. Their planned future work involved investigating whether the bag-of-words learning from their model could be transferred to other models in a similar domain. One major difference between our study and theirs is that they trained the data on a completely different set than they attempted to classify, while our test and training sets came from the same source. This could have introduced error in the form of shifts in parliament makeup and discourse. Another difference is that this study investigated across several members working at the same time, whereas only one president can exist at one time. This means there would be one role in our study against the many members of the European Union Parliament that each have their own focus and identity within a given period. As mentioned towards the end of the paper, the members of the Parliament each have their own issues and priorities, and as such efforts have to be taken to compare the content of their speeches without sacrificing links between them. In that way our research and accuracies may benefit from some aspect of unity regarding the stable position of who gives the State of the Union addresses. Aside from those issues and the inherent differences from studying a different subject body of government, the other significant difference is that, similar to the previous paper, this analysis focuses on a single relatively brief time period rather than training across the entire lifetime of the body of government. That said, the paper does use a similar technique to ours for training and testing the corpus, as well as having a similar domain of formal political speeches. We model some of our analysis on how they decompose the success of their model and also plan to investigate how transfer-of-learning could apply to our models.

”Fake News and Indifference to Truth: Dissecting Tweets and State of the Union Addresses by Presidents Obama and Trump” by David E. Allen, Michael McAleer, and David McHardy Reid employs sentiment analysis and statistical investigation into the tweets and State of the Union addresses given by Presidents Obama and Trump to attempt to gain insight into the differences between the two leaders(Allen, McAleer, & Reid, 2018). The sentiment analysis was done using an R package which grouped key words into the five emotions ”joy”, ”sadness”, ”fear”, ”anger”, and ”surprise” throughout the address. The authors note that while the overall meaning of the address has merit, there is a certain undeniable power that the immediate emotion evoked by a word has in a formal address like the State of the Union that will be broadcast beyond the Chambers of Congress. Both statistics and sentiment analysis agreed that the address and Twitter content analyzed for Trump was in general more positive than President Obama’s content. The authors take great effort to establish that this is not an analysis of the quality of the addresses and express their interest in investigating whether the differences in positivity reflect more on Trump’s self-aggrandizement or the actual state of the country at the time period. After all, the State of the Union address has traditionally been a location for Presidents to outline the threats to the nation and prioritize what problems they will deal with, not just a ground for bolstering the country’s image. Our key takeaway from this paper is with regards to the specific Presidents in question and the statistical differences between their addresses. The transition between President Obama and President Trump was one of the largest in terms of partisanship, and the statements of the latter on Twitter have been widely considered to have expanded the gap between the major parties in the US, so much so that a group of his supporters attacked Congress when it became clear that he would have to give up his position to a Democratic candidate. This event marked an important milestone in Presidential discourse as the former president was banned from most social media sites. By retrospectively analyzing the existing differences from an extreme jump in partisanship we can gain insight into the factors that may influence our models, and because such differences were found from sentiment analysis it is likely that our models have something to pick up on. The statistical analysis is especially relevant to our Multinomial Naive Bayes classifier, as the verifiable presence of differing frequencies of tokens is critical to its functioning.

The 2010 paper ”State of the Union Addresses and the President’s Legislative Success” by Jeff Cummins investigates whether the State of the Union address has an impact on how effective a president will be at accomplishing their campaign goals(Cummins, 2010). The analysis is done manually, defining variables which logically define a president’s success. These are offset by additional variables that indicate any interference that might either help or obstruct the president like military activity or partisanship of the legislation, with final counts being fit to a model using Prais-Winston regressions. What the author found was that the direction of the State of the Union address did have an impact on which policies got implemented during a president’s term, although the influence of partisanship and opposition-party strength almost completely eliminated these benefits in contested periods. The manual nature of the paper makes it difficult to relate to our work, but the insights it offers into the mechanical functioning of the State of the Union address with regards to being a ”road map for a president’s policy” and the delineation of different outside variables helps to give context to the otherwise difficult to interpret and measure aspects of the State of the Union address. The limitations of AI with regards to having a deep understanding of the domain and how it relates to outside context intuitively can be mitigated by limiting our own application of them and supplementing with our own ideas and conclusions. As such, this analysis helps to explain where our models might fall short and guide our investigation into what might be picked up by the automatic statistical analysis of our models instead of the manual analysis done by the author. The author also found that strong partisanship and opposing political party strength diminished the positive effect the State of the Union address had on policy. The author suggested that this would motivate presidents to give shorter, more focused addresses with more bipartisan support in order to increase the odds that the policies they discussed would get passed. This is important to our research because the frequencies at which partisan policy related words appear might decrease in our corpus during that time, making distinguishing between the parties more difficult. That said, it is unclear whether this actually happens in practice, since the analysis is somewhat subjective and does not provide specific empirical evidence for that motivation existing.

In the paper ”Change in Metaphorical Framing: Metaphors of TRADE in 225 Years of State of the Union Addresses (1790–2014)”, Christian Burgers and Kathleen Ahrens investigate how the changing of domains and topics over time have modified what metaphors for trade have gotten used in the State of the Union addresses from 1790 to 2014(Burgers & Ahrens, 2020). Metaphors for trade are broken up by the authors into physical objects, living beings, containers, journeys, and buildings. These are tracked with reference to both what the metaphor concretely represents within a domain and what domains get swapped out over time. The process involves searching for instances where the word trade occurs in the State of the Union corpus and classifying each of the 1,159 cases as belonging to one of the five categories manually. An example of this is given for each case, including the sixth case where trade appears in a literal manner without any metaphorical meaning. Two coders developed to classify instances of trade were trained and compared to evaluate the inter-coder agreement. The resulting agreement between the models was generally "very good" according to the model-specific criteria. The results show that attention given to trade as well as the metaphorical mapping of trade has fluctuated without an evident trend across the history of the State of the Union address. The authors found that these fluctuations seemed to hover around the same values throughout, indicating that the use of metaphors was stable across time. Only use of the word trade as a metaphor for a container increased in use over time. They also found that the most common metaphors were the ones that had a wide array of possible meanings and did not get much more specific than their broad category. Metaphors are interesting with regard to text analysis models because they often result in the same words being used to describe several different domains, despite potentially having conflicting concrete meanings. In this way they can pose a potential roadblock or a potential boon to classifiers. A classifier will run into issues when presented with complex wordplay or non-concrete language because it relies on statistical occurrences of words and often disregards the context and spatial order of the tokens that give the wordplay its meaning. On the other hand, a classifier benefits from several related concepts using the same vocabulary, as this increases the effectiveness of adding weight to those common words when predicting whether those concepts are present. This is not unlike the idea of sentiment analysis, which could be seen as a simplified metaphor that maps several words to one meaning. That said, sentiment analysis would undoubtedly beat out extended metaphor use as far as classifier impact is concerned.

"Classifying Party Affiliation from Political Speech" by Bei Yu, Stefan Kaufmann, and Daniel Diermeier tackles a similar problem to our research, attempting to train classifiers to predict party affiliation based on Congressional speech data(Yu, Kaufmann, & Diermeier, 2008). The speeches were gathered from the House of Representatives and Senate for 2005, and subsequently a series of text classifiers are trained on this corpus. These classifiers were then evaluated in two interesting ways: one evaluation checking how the House of Representatives and Senate models evaluated each other's speeches and the other checking how the two evaluated speeches since 2005. These evaluations respectively intended to demonstrate the person dependency (with the people being the House and the Senate) and time dependency. What they found was that the House model performed well on the Senate corpus, the Senate model performed poorly on the House corpus, and the House model performed better the closer to modern day the speeches were taken. An extremely important point that the authors make is that text classifiers struggle with identifying political support or opposition because politicians tend to stay close to vague terminology and avoid specifically calling out a topic so as not to ruffle feathers among allies or stir unwanted controversy. This makes it difficult for classifiers to make connections that bridge the content, which is vague, with a classification that must be specific and have specific views. As far as direct parallels, this paper makes heavy use of several of the same classifiers as we do, explaining the benefits and drawbacks of the two Naive Bayes classifiers and singing the praises of the SVM model. Of course, as was the case with the previous papers, there is a domain difference between the State of the Union address and Congressional speeches. It is entirely possible the two have a large amount of overlap, but in general the President must exercise more unity and a wider breadth of topics in a State of the Union address than a House Representative or member of the Senate must cover in one of their speeches.

Finally, ”Multi-dimensional register classification using bigrams” by Scott A. Crossley and Max M. Louwerse offers an in-depth linguistic analysis of how four dimensions of a corpus can be shown through factor analysis of bigrams(Crossley & Louwerse, 2007). Bigrams are tokens found spatially adjacent to each other in a text, similar to the co-occurrences studied in ”Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014”(Rule et al., 2015). They found that classifying using bigrams led to strong results for both written and spoken texts. Unfortunately a paywall prevented us from gaining the specific figures and statistics presented in the paper. Several of the four dimensions do not have any bearing on our corpus because of the nature of the State of the Union as a formal address, but the most critical part of the study was how it incorporated both spoken and written works into its corpus. The State of the Union address has been given as both, so an in-depth analysis of how that might differ is indispensable to our work. That said, our work does not consider any spatial context beyond the grouping of words into chunks, so the generalizations made in this paper do not necessarily hold for our work. It is still possible that the distinctions between spoken and written works, or the lack thereof, will have an impact on our model’s ability to categorize the addresses. The fact that the bigram models were powerful regardless of their format should indicate that the format of the address should at the very least not break the model’s ability to predict.

3. Data – Preprocessing

Our data was obtained from the Kaggle ”State of the Union Corpus (1790 - 2018)” dataset uploaded and maintained by Rachael Tatman. The official set does not describe where the data was obtained, but the discussion indicates that the address texts may have been pulled from Project Gutenberg and the White House Website. The list of presidents and their party affiliation had to be manually entered. This data set did not actually include the initial 1790 state of the union, so our first address is George Washington’s 1791 address. The final address is the 2018 address. The addresses vary in length, with the shortest address being the first given by George Washington in 1790 at roughly 1,000 words and the longest being Carter’s 1981 address at over 33,500 words. Since we do not have the first address, the second shortest is John Adams 1800 address with roughly 1,300 words. The average length is just over 4,000 words, and in total the corpus is made up of 932,233 words. Figure 1 plots the frequency of words before processing.

The effectiveness of the models we decided to train depend largely on the distribution of word usage, so it is important to consider the linguistic differences and qualities of the corpus of text.

The State of the Union is a formal address given to the legislators in the US Congress, so the language is generally very formal and policy-related. Addresses tend to use words linked to the values of the United States such as ”freedom” to justify policy, meaning that there are several words which are common across all addresses. Also mentioned are the state

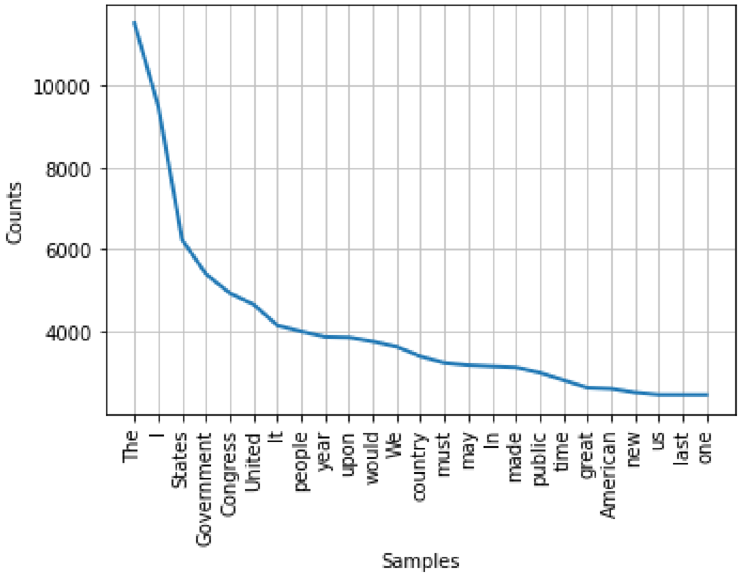


Figure 1: Word Frequencies in the State of the Union Addresses.

of foreign relations. Historically these have been largely between the US and European powers, although in the 20th and 21st century the US has become more of a global power.

When considering the State of the Union address, it is important to also consider the implications of how the role of President has changed over time. For instance, the State of the Union was originally mostly relayed through a written note. Since Woodrow Wilson however, the address has been given in-person and broadcast over the radio to gain public support (Historian, 2021). This represents a growth in the role of the President in the age of radio towards being a nationwide figurehead, as people outside DC could now hear and be persuaded by the president. There is a high likely-hood that knowing the address would be publicly broadcast would shift the linguistic properties from those more common in written word to those more common in spoken word, creating a milestone in the data that could be potentially recognized and used by the model. A similar shift could be seen as the US, and by extension the President, became a larger player in global politics. The State of the Union is now observed by global leaders, meaning that the language has had to shift from an internal correspondence to an official and carefully worded statement. This has no doubt altered the frequency certain words get used in the addresses.

We also must consider how language in general has changed over the past 231 years. When reading through a sample of the corpus from the 1700s and a sample from the 2000s it is painfully obvious when each was written. Words like "depredations" have fallen out of favor in the modern century and rarely appear in modern addresses. More specifically to politics, nations have changed names and the scales of numbers have increased. There would not be much that an early president could have described as existing "in the trillions", and "Persia" would sound out of place in the modern day. Obviously the names of rulers and leaders would date an address even further, as it is rare for someone to remain in power and relevant to the US for more than a few years. These will provide footholds for the

In summary, the vocabularies used in each address in the corpus of text contains many similarities but differ enough to be distinguishable. There is reason to believe that the differences in vocabulary should be tied to time period and potentially party alignment, so it should be an ideal corpus for training text classifiers.

State of the Union transcripts were tokenized and then stopwords and punctuation were removed from the data. The cleaned tokens were divided into chunks of 1,000 tokens each and then count vectorized. The count vectorization was performed in order to get the term frequency of the text. It was then split into testing and training set for our Naive Bayes classifiers using a 75/25 training/testing split.

For our K-Nearest Neighbors classifiers, the data was once again tokenized, cleaned, and then count vectorized. However, the count vectorized data was then converted to a matrix of TF-IDF features. TF-IDF, also known as term frequency-inverse document frequency, is a measure of how important a word is to a document in a collection. In KNN we want to use it in order to find better terms for our analysis and ignore more common words such as 'the' that appear a lot in documents but are not useful in our model.

Furthermore, while the original data has all of the political parties of former presidents present for some experiments we chose to focus only on 'Democrat' and 'Republican' while grouping all of the other possibly options ('Democratic-Republican', 'Federalist', 'Whig', and 'Unaligned') as 'Other'. We do this grouping simply because the majority of what we would consider 'Other' are not all that common in our dataset as they are all relatively old parties and therefore do not appear individually enough for it to be particularly significant with our models. The last president that was not affiliated with either the Democratic or Republican parties was 13th president Millard Fillmore, making the last State of the Union address in the 'Other' category from 1852. Figure 2 shows the State of the Union address distribution by party affiliation. In this figure, of the 227 speeches in our dataset 92 were from Democrats, 89 were from Republicans, 28 were from Democratic-Republicans, 4 from Federalists, 8 from Whigs, and 6 were unaligned. In total, the parties under 'Neither' in the figure only account for 46 of the 227 speeches. Therefore, in some of the experiments 'Democratic-Republican', 'Federalist', 'Whig', and 'Unaligned' are grouped together into one larger category in an effort to reduce the amount of specific categories our model has to try to predict in an effort to get higher accuracy on what we want to focus on – 'Democrat' and 'Republican'.

4. Experiments

The experiments performed on our dataset were carried out in Python using Multinomial and Bernoulli Naive Bayes classifiers as well as Brute Force, KD-Tree, and Ball Tree K-Nearest Neighbors classifiers. The different classifiers that were run did not require significant resources to execute.

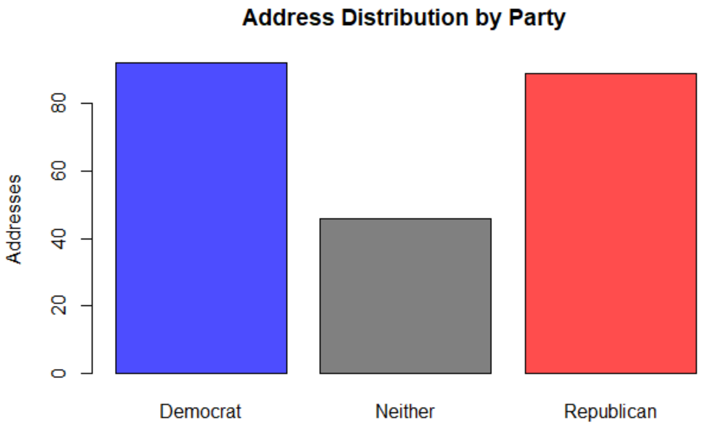


Figure 2: Number of speeches given by each party.

The Naive Bayes classifiers were chosen for these experiments because of our goal of trying to predict party affiliation based on the content of the State of the Union addresses. Naive Bayes is a simple and effective choice for text classification so it was chosen for these experiments.

The K-Nearest Neighbors classifiers were chosen alongside Naive Bayes in an attempt to gain better accuracy. KNN is useful for text classification because of its simplicity, its ability to produce a higher accuracy, and its versatility due to its different algorithms. KNN does best when used in a smaller dataset such as ours and can produce great results, so it was overall a good complement to the Naive Bayes results.

State of the Union address transcripts were tokenized, put into chunks of 1,000 tokens, cleaned of stop words, count vectorized, and then passed into our models. For the KNN models, it is also transformed to a normalized TF-IDF representation. All experiments used a 75/25 training/testing split.

The first two experiments were conducted using three political party categories: 'Democrat', 'Republican', and 'Other'. These two experiments focused around the attribute 'party' which we were trying to predict. In the first experiment, Multinomial and Bernoulli models were fit on the training data and the trained model was used to predict on our test set. In the second experiment a Brute Force, a KD-Tree, and a Ball Tree K-Nearest Neighbors classifiers were used with varying k-values. For Brute Force a k-value of 2 was used, for KD-Tree a k-value of 5 was used, and for Ball Tree a k-value of 10 was used.

The next two experiments were conducted using the full range of political parties to see if our model still worked accurately when presented with more specific categories. The full range of categories include 'Democrat', 'Republican', 'Democratic-Republican', 'Federalist', 'Whig', and 'Unaligned'. Despite the more specific categories, we are still trying to predict the 'party' attribute. These experiments consisted of the same Multinomial and Bernoulli

Model	Grouped Acc.	Ungrouped Acc.
Multinomial Naive Bayes	0.868	0.853
Bernoulli Naive Bayes	0.832	0.780
Brute Algorithm KNN	0.788	0.784
KD Tree KNN	0.846	0.839
Ball Tree KNN	0.791	0.777

Table 1: Accuracies of tested models.

Party	Proportion of Corpus	Proportion of Test Set	Average TP Rate	Max TP Rate
Republican	0.392	0.564	0.782	0.864
Democrat	0.405	0.348	0.938	0.989
Other	0.203	0.088	0.546	1.000

Table 2: True Positive rates of parties vs. expectation based on population distribution.

Naive Bayes models used before for one, and the other used the same Brute Force, KD-Tree, and Ball Tree KNN models from before. Again, the only difference between this set of experiments and the first are the number and specificity of the political parties we are using.

5. Analysis

Our models performed generally very well, each netting accuracies well above random chance. Table 1 shows the accuracies for each model type. There is an accuracy for the ”grouped” models that were trained on the set that categorized any non-Democrat or non-Republican president’s address as belonging to an ”other” party, as well as an accuracy for the models trained on the ”ungrouped” set that kept the smaller parties distinct.

Multinomial Naive Bayes performed the best in both categories, with the KD Tree KNN slightly behind and the other three grouped below. One exception was the Bernoulli Naive Bayes, which jumped up above the bottom three for the grouped data.

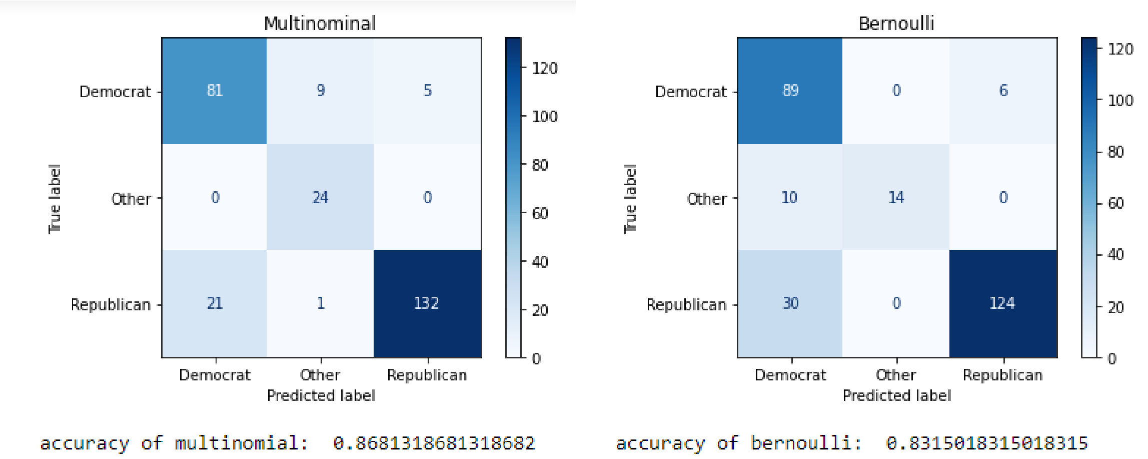
Overall grouped models outperformed ungrouped models, although the extent to which they outperformed differed between models. The Bernoulli Naive Bayes saw a full 5% drop in accuracy when shifting to the ungrouped data set, the most extreme drop of the set. On the other hand the Brute Force Algorithm KNN saw the smallest difference between grouped and ungrouped, only shifting down 0.4%. It is worth noting that the Brute Force Algorithm KNN already had the smallest accuracy for grouped, so the smaller drop may have more to do with never being that high to begin with.

With a perfectly evenly distributed data set we would expect a random model to perform with an accuracy of 0.333 for the grouped data and 0.167 for the ungrouped data. However, our data was not perfectly distributed and based on the distribution of addresses we would expect to see the True Positive rates found in Table 2 roughly fall around the proportions of the test set or corpus if our models were not effective. The ”True Positive Rates” were found by isolating each category and treating the trinary confusion matrices as binary ones, where the model either correctly classified the party alignment of the test chunk or did not correctly classify the party of the test chunk. For example, for the grouped Multinomial Naive Bayes model in Figure 3, 81 of the Democrat chunks were correctly classified as "Democrat", 9 were incorrectly classified as "Other", and 5 were incorrectly classified as "Republican". This means the model got 81 of the 95 correct and would have a True Positive Rate of 0.853. That said, all of our True Positive Rates were well above these values with the exception of some of the "Other" True Positive Rates which suffered from having a lower number of total chunks.

One issue we ran into with the ungrouped models is that the less frequent parties like the Whigs or especially the Federalists (which only includes John Adams's 4 addresses) would have so little data associated relative to the two major parties that the models would be unable to distinguish between them. The Bernoulli Naive Bayes model struggled the most with this, as can be seen in Figures 3 and 4. It performed fairly well for the grouped set, correctly classifying over half of the non-Republican and non-Democrat addresses and never assigning a false "other" prediction. However, in the ungrouped set it incorrectly predicted that every single one of the non-Republican and non-Democrat addresses were from a Democrat president. Bernoulli Naive Bayes models only check for the presence of tokens, independent from the frequencies that they appear at, so it is possible that the Democrats and parties classified as "other" share similar vocabularies but use the words at different frequencies. This is backed up by the Multinomial model outperforming the Bernoulli model with regards to non-Republican and non-Democrat addresses, correctly classifying all 24 of the "other" chunks in the grouped testing set and correctly classifying 14 of the 24 in the ungrouped set. As mentioned earlier in Professor Jacques Savoy's work, the vocabulary growth of the State of the Union addresses has led to new words being introduced into the collective vocabulary of the address over time and words in the vocabulary shifting out of use (Savoy, 2015). It is possible that the Bernoulli Naive Bayes model was relying on the presence of certain tokens in the time period that the non-Republican and non-Democrat parties were active during, rather than the tokens that would place an address into one specific party or the other. In short, the Bernoulli model may have been finding success predicting the time period rather than the party, and since the majority of American political history after the first 75 years has been bi-partisan, the model would have gotten decent scores from assuming older-text came from a third party.

As seen in Figure 6, none of the KNN models were able to outperform the Multinomial model's true-positive rate for "other" parties, although the Brute force and KD-tree algorithm models matched the Multinomial model's record of 14 correct classifications. The parties that they got correct were different, so the models were picking up on different features that successfully classified different chunks.

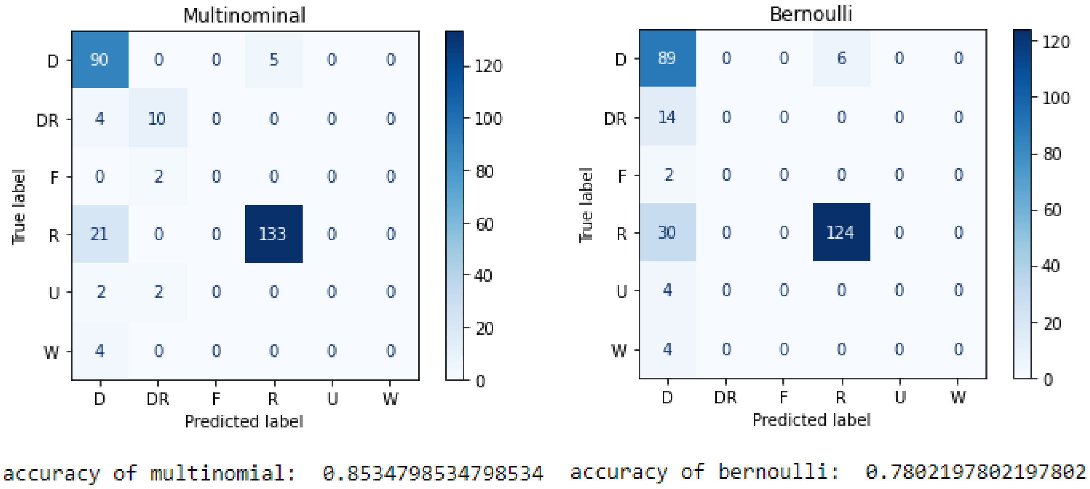
When retesting the values on new randomly selected testing and training sets, there seemed to be some variance in how extreme the differences were. For example, in Table 3 we can see that with a different train-test split, our KNN models did significantly better than the Naive Bayes models. Our grouped Multinomial model, which did the best with our previous test, did not get anywhere close to its grouped accuracy of in Table 1, resulting in a difference of almost 20.2%. This is likely due to our dataset containing high variance, given the large jump in accuracy based on the different training and testing sets. We see this



```
accuracy of multinomial:  0.8681318681318682
```

```
accuracy of bernoulli:  0.8315018315018315
```

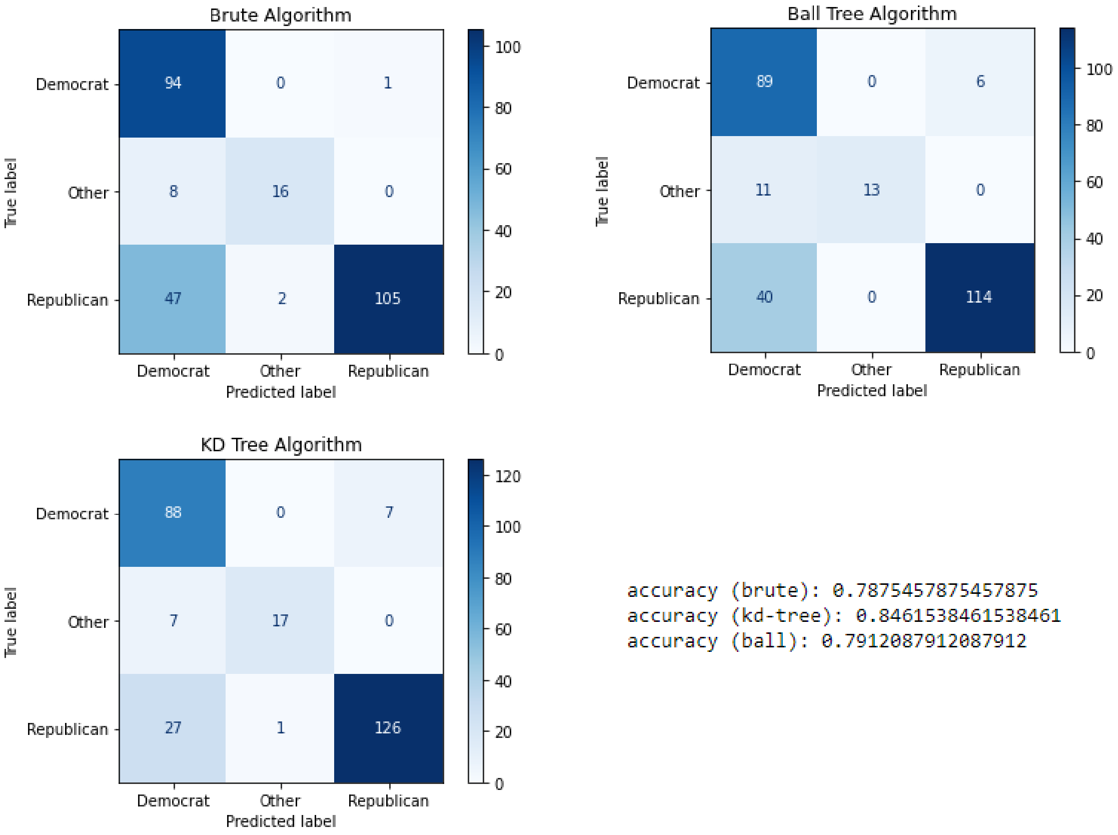
Figure 3: Confusion Matrices for Grouped Naive Bayes Models.



```
accuracy of multinomial:  0.8534798534798534
```

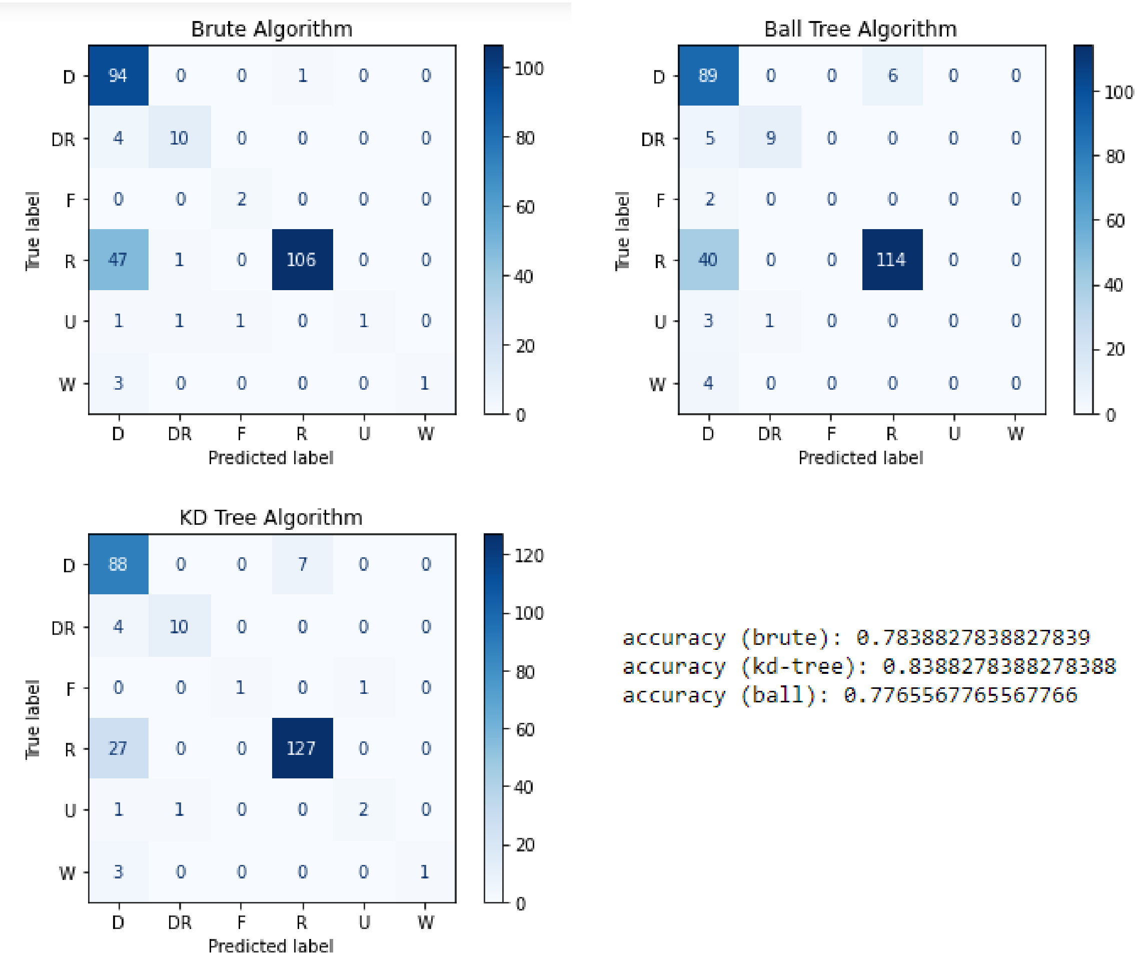
```
accuracy of bernoulli:  0.7802197802197802
```

Figure 4: Confusion Matrices for Ungrouped Naive Bayes Models.



```
accuracy (brute): 0.7875457875457875
accuracy (kd-tree): 0.8461538461538461
accuracy (ball): 0.7912087912087912
```

Figure 5: Confusion Matrices for Grouped KNN Models.



Brute Algorithm

	D	DR	F	R	U	W
D	94	0	0	1	0	0
DR	4	10	0	0	0	0
F	0	0	2	0	0	0
R	47	1	0	106	0	0
U	1	1	1	0	1	0
W	3	0	0	0	0	1

Ball Tree Algorithm

	D	DR	F	R	U	W
D	89	0	0	6	0	0
DR	5	9	0	0	0	0
F	2	0	0	0	0	0
R	40	0	0	114	0	0
U	3	1	0	0	0	0
W	4	0	0	0	0	0

KD Tree Algorithm

	D	DR	F	R	U	W
D	88	0	0	7	0	0
DR	4	10	0	0	0	0
F	0	0	1	0	1	0
R	27	0	0	127	0	0
U	1	1	0	0	2	0
W	3	0	0	0	0	1

```
accuracy (brute): 0.7838827838827839
accuracy (kd-tree): 0.8388278388278388
accuracy (ball): 0.7765567765567766
```

Figure 6: Confusion Matrices for Ungrouped KNN Models.

Model	Grouped Acc.	Ungrouped Acc.
Multinomial Naive Bayes	0.666	0.688
Bernoulli Naive Bayes	0.658	0.604
Brute Algorithm KNN	0.808	0.817
KD Tree KNN	0.804	0.792
Ball Tree KNN	0.770	0.763

Table 3: Accuracies of retested models.

in Table 2, where the party distribution of the corpus is significantly different from the party distribution of the test set. Republican chunks outnumber Democrat chunks in the test set despite Democrat chunks outnumbering Republican chunks in the corpus. Ultimately this just comes down to how the sets are randomly split when the program is run.

6. Conclusion and Future Work

Overall, both the Naive Bayes and K-Nearest Neighbors models were successful in predicting political party alignment based on the content of State of the Union addresses at high accuracies that were all well above random chance.

Further experiments could work to identify whether the models work well with other political content whether it be speeches or even possibly articles, books, tweets, etc. It would be interesting to see how these were classified with our models since as of now it is only looking at these formal political documents, so it could be worth exploring more informal text and seeing how that compares.

It could also be worth exploring different models. Many of the peers we researched made use of Support Vector Machines in their text analysis. Support Vector Machines are extremely powerful for text classification, especially with regards to detecting complex ideas and patterns. Their lack of inclusion in our experiments was more due to time constraints rather than lack of fit. A SVM trained on our data set could very well provide better results than our current models. That said, SVMs generally rely more heavily on lots of training data for each case, and our set is relatively small compared to typical SVM training sets. The word-count is decent, but the complexity of the subject matter means there are lots of potential patterns and cases to pick up on, and each case may only have a few examples. Still, it would be a valuable test to develop.

Another possible route would be tracing which addresses frequently got incorrectly predicted by models to see if there is overlap between the models. It is possible that some presidents made efforts to avoid partisan language and would be difficult for even a trained linguist to distinguish. Self proclaimed centrists like President Eisenhower may fall under this category. If this is the case it may be interesting to approach the problem from the angle of regression, assigning a value to how far Democrat or Republican each president was. The main issue here would be that this value is subjective and could be difficult to determine, especially for presidents prior to the bipartisan era. There is also the consideration that the platforms of the two parties have shifted over time and two presidents very extremely aligned with a party at different times could have very different platforms.

In experimenting further with different training and testing splits with our data, we concluded that our dataset contains high variance, likely due to our dataset being relatively small. Therefore, our model's accuracy varies a lot depending on what testing and training data it is presented with. Due to time constraints, stratification of the testing and training split was evaluated somewhat but ultimately not implemented. From our few tests with this implemented, it seemed to lower the variance and would be worth exploring in a future experiment. Cross validation could also be implemented in order to bring that variance down further.

As mentioned previously, the paper "Predicting Party Affiliations from European Parliament Debates" by Bjørn Høyland, Jean-François Godbout, Emanuele Lapponi, and Erik Velldal incorporates a study of how future efforts may be employed to transfer the learning of the models across different domains. This seems like it would be an effort that could yield good results with our work as well. United States partisanship is becoming a larger issue globally, so being able to transfer the learning to another medium or nation's political speeches could be beneficial.

Finally, we would like to experiment with how interpreting sentiment analysis would alter the accuracy of the models. Sentiment analysis could potentially eliminate some of the time-context from the corpus, as combining different tokens into one sentiment would help alleviate the changing specific names across history. However, whether eliminating such a context would be beneficial is in question, as the accuracy of the non-Republican and non-Democrat parties is potentially linked to the ability for historically significant words and milestones that characterize their collective older time-period. Several of the related papers we investigated were written with regard to the changing nature of the State of the Union address linguistics over time, but others made effective use of sentiment analysis to bridge gaps between different domains. The question may be less of how the accuracy is affected and more how sentiment analysis would open the door for the aforementioned transfer of learning.

References

Allen, D. E., McAleer, M., & Reid, D. M. (2018). Fake news and indifference to truth: Dissecting tweets and state of the union addresses by presidents obama and trump. *Tinbergen Institute Discussion Paper*, *020*(3).

Biessmann, F., Lehmann, P., Kirsch, D., & Schelter, S. (2016). Predicting political party affiliation from text. *PolText 2016*, *14*, 14.

Burgers, C., & Ahrens, K. (2020). Change in metaphorical framing: Metaphors of trade in 225 years of state of the union addresses (1790–2014). *Applied Linguistics*, *41*(2), 260–279.

Chen, X., & Hu, J. (2019). Evolution of u.s. presidential discourse over 230 years: A psycholinguistic perspective. *International Journal of English Linguistics*, *9*(4), 28.

Crossley, S. A., & Louwerse, M. M. (2007). Multi-dimensional register classification using bigrams. *International journal of corpus linguistics*, *12*(4), 453–478.

Cummins, J. (2010). State of the union addresses and the president's legislative success. In *Congress & the Presidency*, Vol. 37, pp. 176–199. Taylor & Francis.

Historian (2021). State of the union address: Us house of representatives: History, art, archives..

Høyland, B., Godbout, J.-F., Lapponi, E., & Velldal, E. (2014). Predicting party affiliations from european parliament debates. In *Proceedings of the acl 2014 workshop on language technologies and computational social science*, pp. 56–60.

Rule, A., Cointet, J.-P., & Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, *112*(35), 10837–10844.

Savoy, J. (2015). Vocabulary growth study: An example with the state of the union addresses. *Journal of Quantitative Linguistics*, *22*(4), 289–310.

Yu, B., Kaufmann, S., & Diermeier, D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, *5*(1), 33–48.