**Introduction**

The United States of America constitutionally protects the right of the people to speak freely. The ability to express oneself freely and openly was so important to the Framers of the Constitution – politicians, mostly – that they listed it before the remainder of the first 10 amendments commonly referred to as the Bill of Rights. Indeed, it is this right that allows for protest – preferably peacefully – to take place, which has led to monumental change throughout the history of this Nation. From Martin Luther King Jr. to the current climate in response to the death of George Floyd in Minneapolis, free speech has defined us as a Nation. But free speech does not infer that everyone speaks for themselves. Unfortunately, some people seem to think, speak, and act a certain way in order to be representative of a group. Call it optics, and no group of people in the country exemplifies this more than politicians.

Anecdotally, it seems as if politicians are likelier than ever to toe the party line. While it is true that party lines shift and change over time, certain fundamental beliefs seem to persist within each party. It is hypothesized by the authors that the tone, context, and language of the party members is pervasive enough to be identified with natural language processing techniques. It seems likely that many of the talking points, relative sentiment and tone of speech, individual or group think, etc., have remained consistent enough over the course of the last 10-20 years that it should be possible to accurately predict which party a Governor belongs to based solely on their inaugural speech.

Why, in the most liberated country on the face of the earth, do modern day politicians choose to toe the party line? Or do they? Surely, they have organic thoughts of their own that deviate from their respective parties. Right?[[1]](#footnote-2)

# **Section 1: Analysis and models**

As stated in the introduction, the goal of this study is to determine if the party affiliation of a governor can be determined by their inaugural address. This section will discuss the data and methodology for the study.

## **Section 1.1: About the data**

In order to study gubernatorial inauguration speeches, it was required that documents be gathered. Our goal was to gather one democrat speech and one republican speech from each of the 50 US states, putting emphasis on elections where party power shifted.[[2]](#footnote-3) This proved challenging. There does not appear to be a “collected works” of gubernatorial inauguration speeches. Perhaps governors are not “important enough” to justify such collections as is done for the President. Additionally, some states have not had a power shift for many years and finding the text of the speeches from more than even ten years ago proved very difficult.

The final collection of speeches came from the following states: Arizona, Connecticut, Illinois, Kentucky, Louisiana, Maine, Massachusetts, Michigan, New Hampshire, and Vermont. Each state is represented by both a democrat and a republican speech that occurred in the last 20 years. No independent Governors were selected; this was not a difficult chore as there have only been four independent Governors since the turn of the century – one of them being Jesse Ventura, a former pro-wrestler.

The 20 speeches were organized into two folders based on the party. To investigate the speeches and gain an initial understanding they were examined first along party lines. With the speeches grouped by party, each collection was broken down into individual words and only the alphabetical words were kept. Additionally, for the initial exploration stop words (it, is, the, etc., as well as the names of the ten states) were removed. Having done this, it was discovered that Democrats used 2,571 unique words in their speeches, 52.12 percent of which were stop words. Comparatively, republicans used 3,478 unique words in their speeches, 54.47 percent of which were stop words. So, while the republicans had more to say, more of it was “fluff”.[[3]](#footnote-4) Below are the 50 most common words and bigrams[[4]](#footnote-5) with democrats on the left and republicans on the right, fittingly.

|  |  |
| --- | --- |
|  |  |

Unigram ranking was nearly identical for the five most common unigrams. Republicans and democrats refer to state, people, new, work and one at nearly identical rates. Indeed, the first major difference between the two lists is the word “together” which appears as the 6th most frequent unigram in the democrat speeches but is only the 20th most frequently used unigram in republican speeches. Accounting for the length of speech differences, one would expect that given the number of times a republican said together that a democrat should have used the word together only 33 times. They used it 52 times; that is a 66% increase over what one would expect. The proclivity of the democratic party to use the word “together” or the republicans’ proclivity not to speaks to the subtle differences in the messaging between the two groups. A similar distinction in unigram frequency is exhibited in the use of the word “want” where republicans use the word at a rate 55% higher than democrats. Differences like this will prove useful when models are trained and tested later.

Bigram frequency appears to be less informative than unigrams. Bigrams appear oftentimes only once or twice in a document – this is likely due to the relatively short size of the documents. "God bless” appears on both the democrat and republican bigram lists as the first and second most common bigrams, respectively. The phrase “God bless” typically is a sign off at the end of a speech. It appeared 10 times for the democrats and 12 times for the republicans – the difference is that the democrats used less words in their speeches. It is unlikely that the presence of the phrase “God bless” has any significant ability to predict the party giving the speech any more so than the phrase “good morning” or another salutation, indeed it is likely that the governors used this phrase because they “felt like they were supposed to”, as alluded to in the introduction.

Aside from the fact that the bigram (God, bless) was more frequent in the democrat speeches than the republican speeches (by an astonishing 0.01%), nothing too enlightening is there, but it does give some additional insight into the speeches. There were 1,040 sentences made up of 21,126 words in the democratic corpus whereas the republican corpus had 1,427 sentences and 30,368 words.[[5]](#footnote-6) Thankfully for the democrats, this is not a popular vote, although typically they wish it were. Kidding aside, there is some level of concern about the disparity between the sizes of the corpuses; it is not tremendous, so the 58% majority case was established as a baseline for classification accuracy. The authors remain cognizant of this disparity.

With preliminary exploration complete, the original unaltered speeches were tokenized into sentences and paired with a label indicating the party of the governor whose speech it came from. These pairs were compiled into a single collection. Each sentence was then tokenized into individual words, and subsequently had non-alphabetical tokens removed. Stop words were left in place, for the time being. Finally, the sentence/party pairs were randomized to ensure no bias is introduced. This process allowed testing via Naïve Bayes classification, which is discussed in the proceeding section.

**Section 1.2: About the models**

As always, classifier models require training and testing sets. However, it is not a simple matter of dividing the sentences into two pieces and then training. A decision must be made as to what will be tested. In other words, how should the classification be done? What features of the sentences could contribute to an accurate prediction model?

The remainder of this section will discuss the general methodology of the training and testing of the data, along with a discussion regarding the features that were chosen. The results of the experiments will be discussed in section two.

While the goal of this part of the discussion is a general run-down of the methodology of the study, this will not be an expose’ on the inner workings of Naïve Bayes classifiers. That being said, the 2,000 most common words from all of the sentences were identified and once a specific feature was selected the following steps were performed:

1. Each of the top 2,000 words that appear in a sentence is assigned a binary label reflective of whether it matches the current feature of choice
2. The collection is then divided into training and testing sets[[6]](#footnote-7)
3. The training data was used to train a Naïve Bayes classifier algorithm within nltk (the Natural Language Toolkit in Python), and the resulting model was applied to the testing data
4. Accuracy was measured, using *n*-fold cross-validation to ensure a good model had been developed
5. Micro-averaged *F1* score was calculated for a better measure of performance of the model.

This sequence of steps was performed for each of five feature sets. The first feature set was the sentences by themselves with no additional specifications other than a check to see if the words were one of the 2,000 most common. The second feature set was subjectivity, utilizing a list of subjectivity measures for all the words. Next, negation was used, where the words were identified as negation words or not (no, not, never, etc.). This was followed by using part of speech tags and finally with bigram pmi measures.

As was mentioned in section 1.1, stop words were left in the sentences initially. This was done so that some of feature sets could be tested with and without stop words in place to see if their removal improved the results. Each experiment was performed with and without stop words to gauge the effect of stop words on the efficacy of the classifier.

Combining different features was explored as well, however the aforementioned repeating of experiments with stop words removed was the only success.[[7]](#footnote-8)

**Section 2: Results**

This section, as described in section 1.2, will discuss the results of the 10 experiments that were performed. As there are essentially five pairs of tests (with and without stop words removed), the results of each pair will be reported together. Please note that accuracies, confusion matrices, and metrics are reflective of 10-fold cross validation, i.e. the accuracy is the mean accuracy for each of the 10 folds in the validation, confusion matrices report percentage[[8]](#footnote-9) instead of count, and metrics are reflective of collective true positives, false positives, etc. across the entire cross-validation. Remember that there is a baseline accuracy of 58% due to the imbalance in the data between democrats and republicans. In each of the following tables, predicted party labels are on top, with actual party labels on the side.

|  |  |
| --- | --- |
| **Base feature** | **Base feature without stop word** |

Accuracy in the base feature set (i.e. no additional features) was improved over merely selecting the majority class. We were more likely to predict republican but were incorrect in our prediction at a similar rate even though there was a significant disparity between actual cases. What is most interesting about the base feature set is that by leaving the stop words included a more accurate model whole but that the majority class, republicans, was predicted less accurately after taking out the stop words. As mentioned previously, republicans use more stop words than democrats do. The logical inference stemming from this acknowledgment is that the stop words are pushing down the remainder of the republican features. In all cases, precision, recall, and F1 were higher for republicans than for democrats. Again, for democrats and for the model in total, removing stop words lower the predictive ability of the model.

|  |  |
| --- | --- |
| **Subjectivity** | **Subjectivity without stop words** |

Subjectivity had similar results to the base feature set. Again, removing stop words improved the ability to predict correct cases but lessened the recall – in other words, predicting the majority class is more likely after removing stop words. In this instance, however, the total accuracy of the model was not reduced upon removing stop words.

|  |  |
| --- | --- |
| **Negation** | **Negation without stop words** |

Negation without stop words was the best model of any of the ten models that were developed. It was the most accurate of any model and had a remarkably high precision for the republicans, topping 0.80. In fact, it was the only model to hit a 0.80 for any metric. As such this model was able to identify the largest number of republican true positives as well. Whether the fact that negation words captured the most republican sentences implies anything about the content of the speeches they were part of and/or reflective of the party is a topic for a different study.

|  |  |
| --- | --- |
| **Part of speech tags** | **Part of speech tags without stop words** |

But for the negation without stop words model – which as previously mentioned was the most accurate model – the part of speech tag with stop words included had the highest F1 value of any model. Its accuracy was in line with many of the other models and shared in the increase of the majority class precision and accuracy increase by removing stop words and the overall decrease in recall and accuracy for the entire model. This is a common theme among many of these models. Again, this model outperformed the baseline of 58% accuracy.

|  |  |
| --- | --- |
| **Bigram PMI** | **Bigram PMI without stop words** |

Measuring bigrams did not significantly increase or decrease the predictive accuracy as compared to the other models. The accuracy results were in line – that is, better than selecting the majority case 100% of the time and roughly the same as the other models. A similar behavior to previous models was observed using bigram PMI as a feature set when considering precision, recall, and F1.

It should also be mentioned that an additional attempt at classification was performed using a support vector machine[[9]](#footnote-10) with only the base sentences. The results were in line with the Naïve Bayes model that was reported earlier in the present section.

It was a bit surprising that the base model with just the sentences and no additional features was nearly 67% accurate. Even more surprising was the general lack of improvement from every other model. The disparity in the precision, recall, and F1 scores between republicans and democrats for every model are perfectly logical due to the imbalance in the data. With this in mind, micro-averaged F1 was considered as it is the better metric for imbalanced data, but because micro-averaged F1 is equal to accuracy, consider the accuracy as the primary measure of the effectiveness of the models.

Also, notice that in every case the number of democrat speeches that were predicted to be republican rose when stop words were removed, and that fewer republican speeches were predicted to be democrat speeches. At first glance, these results are counterintuitive since the republicans had a higher percentage of stop words in their speeches. However, it is (speculatively) probable that this effect on the false positive and false negative rates is moot once a more balanced collection of sentences is acquired and used for modelling.

**Conclusion(s)**

It seems that, based on the results of section two, republicans are likelier to be “of one mind” than democrats. This is not an outrageous conjecture as 10 different tests unilaterally predicted republican speeches more accurately than democrat speeches. However, as was mentioned the collection of sentences is significantly swayed in favor republicans. To truly measure how well these models could predict political affiliation of a sentence, a more balanced collection of sentences is needed. This could also lead to further areas of study.

One aspect that was not explored in the analysis is sentiment. It was conjectured over the course of the investigation that the sentiment of each sentence could influence the political affiliation classification. Politicians are not well known for being polite towards their political opponents; the inherit sentiment of each statement, whether implied or explicit, might make for an interesting feature. If this turned out to be accurate, then the technique could be used both for and against politicians when they claim that a statement was taken out of context, misinterpreted, or what have you.

It was also suggested that the models (or at least a couple of them) be applied to a fresh speech, to see how well that speech might be predicted. The idea being that each sentence could be labelled by the model and then a simple majority would indicate the party of the governor. This was not explored further, however, due to the same imbalance in the current training set.

NLP has been used to detect fake news; a likely extension of that work is applying those models and methods here. Sentences could be tested using fake news classifiers to determine if they were factual statements or lies – within context, of course. The potential benefit is not just automated fact checking – an increasingly necessary service these days – but also a further distinction between the political parties based on an additional “truthiness” feature.

Many amorphic and even some more fully-fledged thoughts and projects have sprung out of this work. Alas they will be left for future projects.

1. The authors have made a concerted effort NOT to introduce any personal beliefs or party/ideological stances in the examination. [↑](#footnote-ref-2)
2. E.g. Kentucky in 2015 when a republican replaced the outgoing term limited, democratic governor and Kentucky again when a democrat was once again elected in 2019. [↑](#footnote-ref-3)
3. A t-test would be required to determine if the mean number of stop words per 100 words was statistically different. [↑](#footnote-ref-4)
4. Trigrams were also looked at, but none were more than 0.01% frequent, with the only remotely interesting ones being “past two years” and “last two years”, both from the republican speeches. [↑](#footnote-ref-5)
5. It is worth noting that a 2,467-sentence corpus is by normal standards not a large enough dataset with which to base the study. The authors acknowledge this, but as this is somewhat of an introductory examination this corpus will serve as a base upon which to grow the study once speeches for all 50 states can be acquired. [↑](#footnote-ref-6)
6. In the initial stages, the ratio of training to testing data was varied to determine which cut point provided the best accuracy. Ultimately, 90%/10% training/testing was selected [↑](#footnote-ref-7)
7. If for no other reason than the authors’ lack of Python skills at the time of writing [↑](#footnote-ref-8)
8. This percentage is reflective of the proportion of the total feature sets THAT were true positive, false positive, etc. [↑](#footnote-ref-9)
9. After significantly different data preparation. This analysis was done with R instead of python. As no new results were achieved the details of this analysis are not included [↑](#footnote-ref-10)