Predicting Supreme Court Votes Through Conversational Dynamic Features

Tara Balakrishnan, Paula Kusumaputri, Luda Zhao
{taragb, paulaksp, ludazhao}@cs.stanford.edu

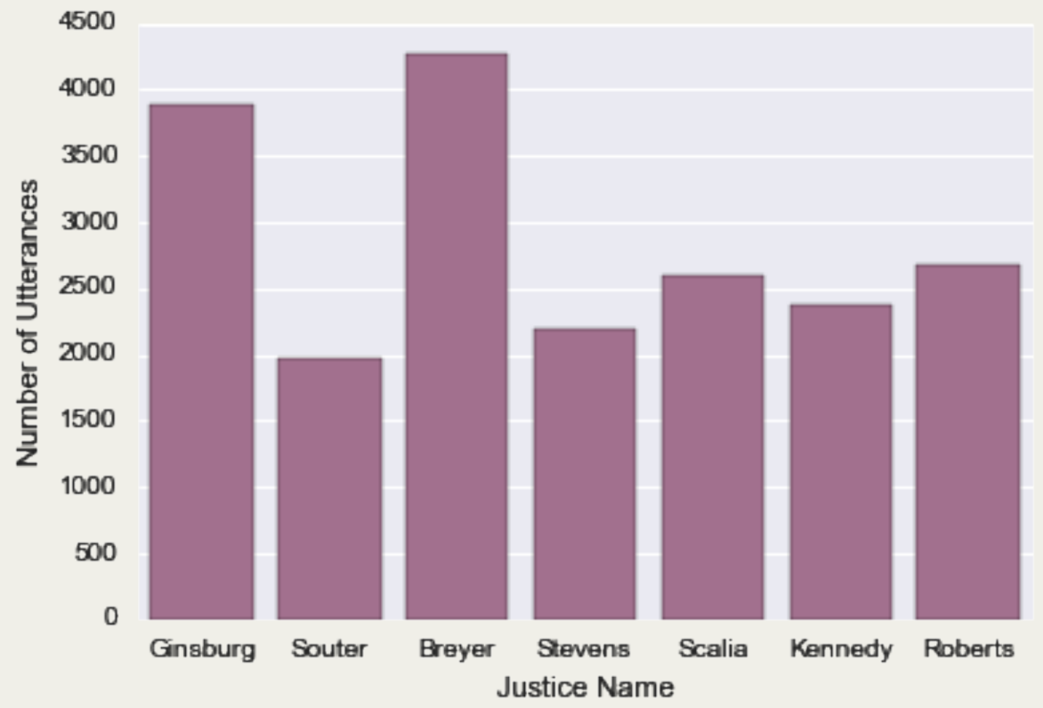
Introduction

It is incredibly valuable to citizens to understand the makeup and workings of the federal government. In order to gain insight into government workings, we want to understand if we can predict the vote that a specific Supreme Court Justice will cast given their participation in the oral argument for a case. As we do not rely upon past voting history, this problem also helps us more generally understand how to interpret conversational dynamics. We answer questions such as can we tell what a person thinks about a subject given the use of certain words, the number of times they interrupt someone, or the sentiment that they exhibit?

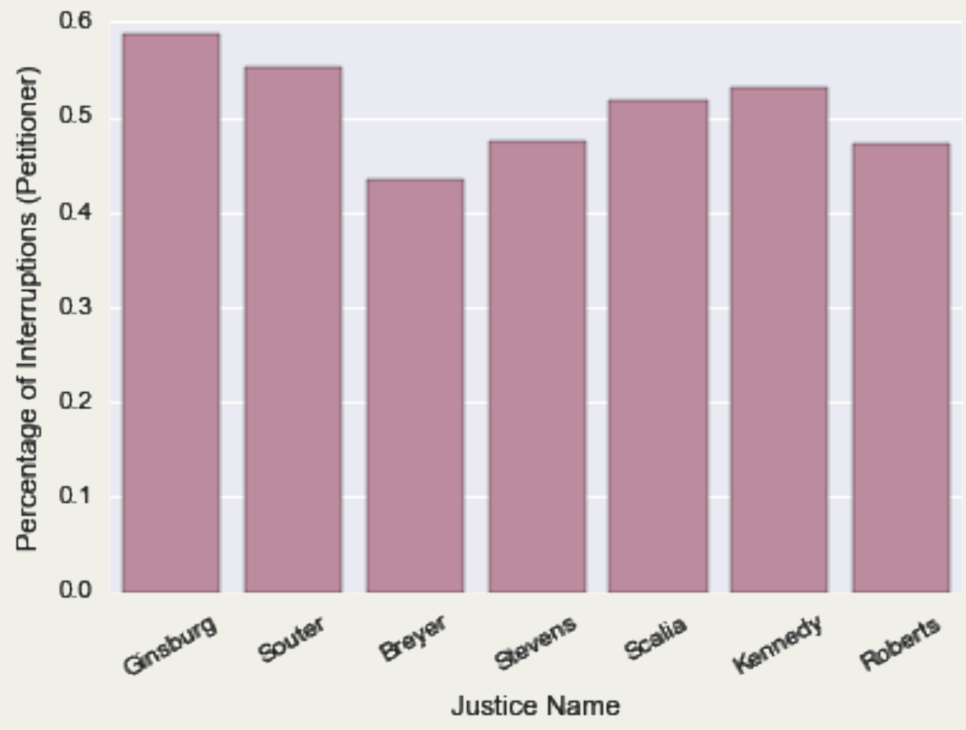
Dataset

- collection of oral arguments, from 204 Supreme Court cases, over a 5-year period
- oral agument split into series of utterances, each utterance marked with metadata: case id, utterance id, speaker, presentation side, indicator if utterance is part of same conversation
- dataset has **51498** utterances, making 50389 conversational exchanges between 7 Justices and 311 respondents/petitioners
- additional meta-data: case outcome, voting results for each Justice involved in case, gender annotation

This graph (right) illustrates the spread between the different Justices in terms of how many times they speak. The number of utterances directly correlates to the number of word features we see.



Historically, interruptions have thought to be an extremely good predictor of a Justice's tendency to disagree with the side he/she is questioning. We are looking to quantify this decades old theory and learn how useful # of interruptions are as predictor of agreement.



This graph (left) illustrates how likely a Justice is to interrupt the petitioner compared to the total number of interruptions that they have made. This gives us an estimate of how often they tend to agree with the petitioner compared to the respondent.

Feature Extraction

We can divide our feature set into detection of 4 broad categories:

1 BASIC LINGUISTIC MARKERS

We extracted simple metrics, such as the average length of the utterances, the average length of the utterances, the percentage of the time the justice is interrupted or interrupts, and the presence of "hedge" words (words that indicate uncertainty). We also tagged each word with a part of speech, from Stanford NLTK's POS tagger, and included both unigram and bigram POS features.

2 SENTIMENT ANALYSIS

We classified each utterance in a case with a positive, negative or neutral sentiment score using VADER sentiment classifier which was pre-trained on social media data. We also look at counts of sentiment reversals in order to discover how a Justice's sentiment trends over the course of a case.

3 IDEOLOGICAL BASIS

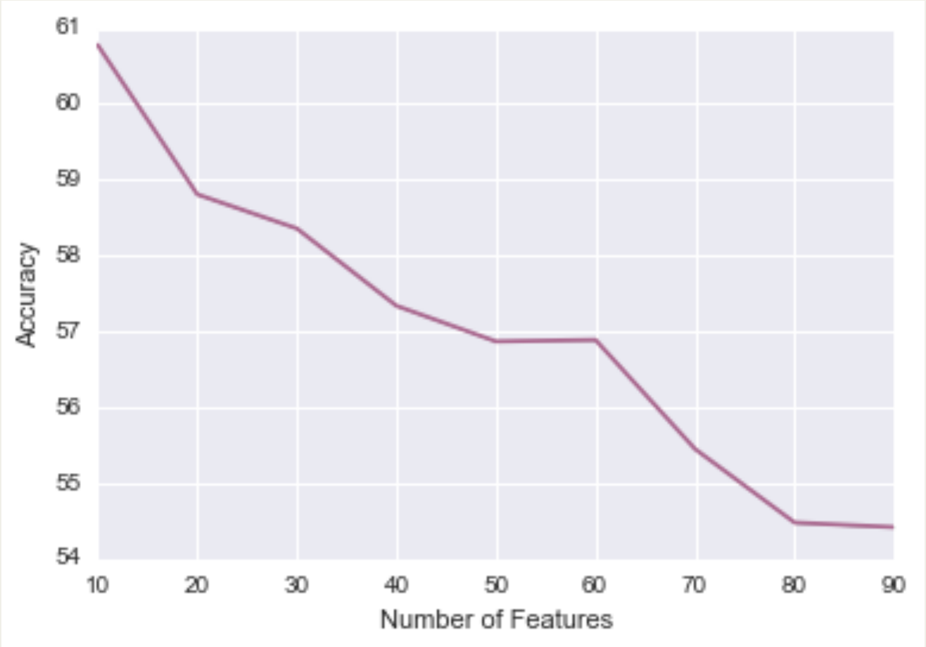
We iterated upon the traditional "bag-of-words" classifier to build an n-gram extractor that stores some spatial information of words in addition to their frequency. We applied feature elimination to get the best set of n-grams for each length n.

4 TEXTUAL SIMILARITY

We used the word2vec model developed by Google to produce word embeddings for a text input. This indicates similarity between words and phrases from the oral argument,when comparing the projection of the word/phrase into multi-dimensional vector space.

Feature Elimination

Because the feature vector that we built for each case contains over 7000 features, we faced risk of over-fitting to our training data. We further improved the performance of our model by reducing the size of our feature vector. Using Recursive Feature Elimination, we reduced the training errors of each model, by limiting the model to those features that have the largest weights, and removing those that are redundant, conflicting, or lead to overfitting.



Models

1 NAIVE BAYES

We started out with unigram feature vector, ran on a Naive Bayes Model. We then iterated upon our feature selection to increase accuracy by first including all extracted features into feature vector and then only the top weighted features in the vector using RFE.

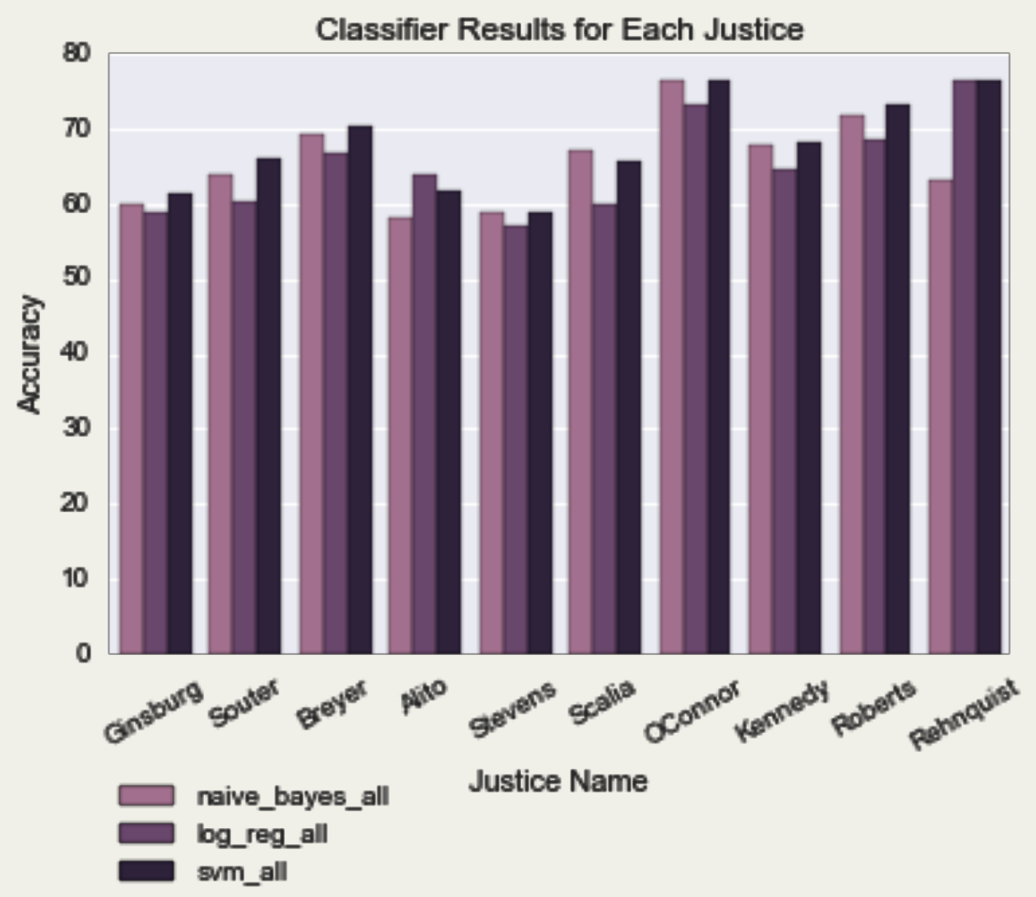
2 LOGISTIC REGRESSION

Implemented L1 and L2 penalized logistic regression for various values of C.

3 SVM

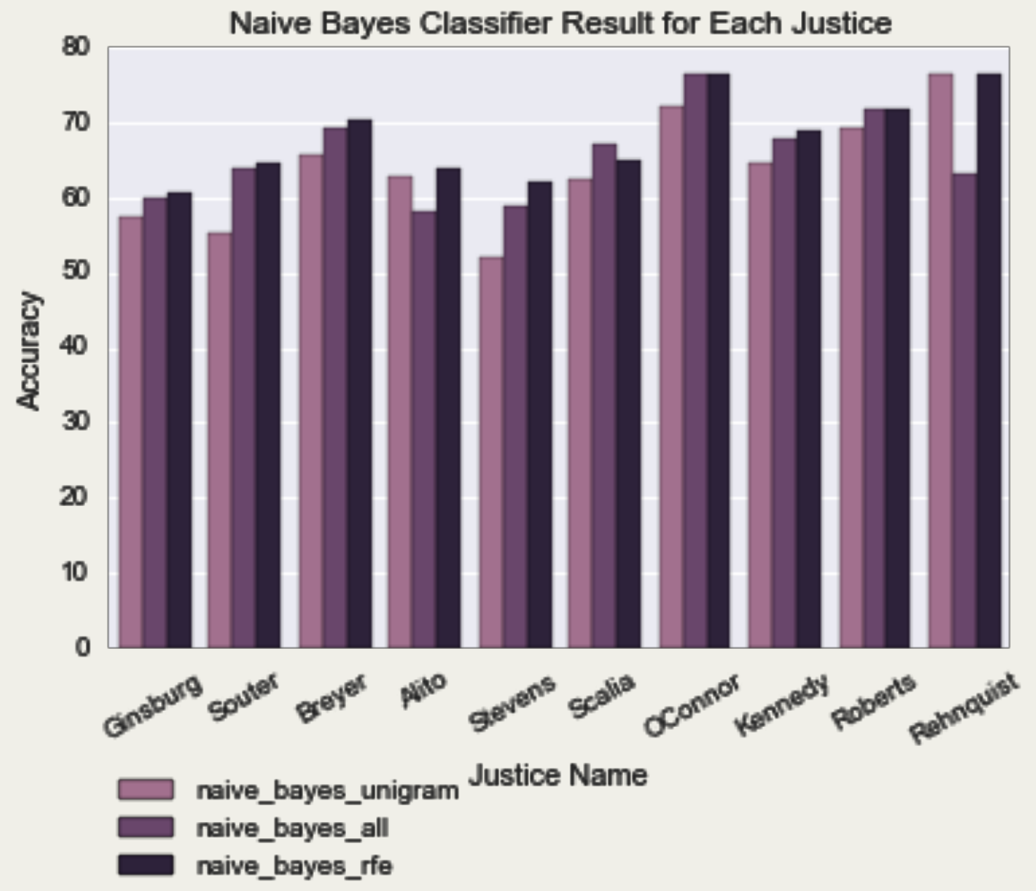
To determine which kernel function to use for SVM model, we evaluated the performance of several different kernels on the entire feature vector. From experimental attempts, the kernel that yielded smallest error was Radical Basis Function (RBF) kernel. When training an SVM with RBF kernel, we examine C and gamma, spaced exponentially for parameter tuning and obtained highest accuracy for C = 0.01 and gamma = 0.1

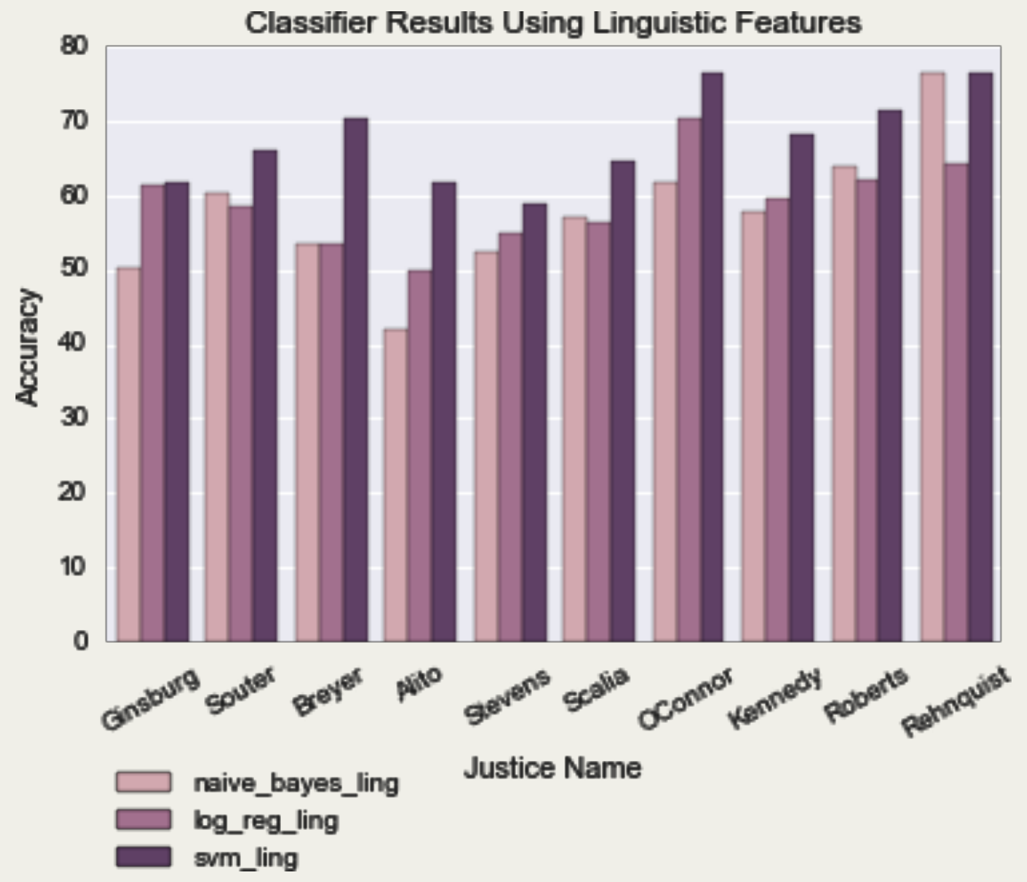
Results + Analysis



Graph (left) depicts prediction accuracy using most of the extracted features (the results still haven't been combined with the linguistic features).Logistic Regression performed poorly likely because we have many strongly correlated features, while SVM performed relatively well due to the soft margin.

With Naive Bayes, we consistently achieved higher accuracy when utilizing reduced feature vectors (RFE). Although we extracted thousands of features to begin, we could generate an accurate prediction using only 10 features.





Graph shows cross-validated accuracy using only linguistic markers (interruptions, sentence frequency or length, presence of hedge word indicating uncertainty) and the average sentiment scores using NLTK Vader sentiment tagger.

Conclusion

Without knowing the past voting histories or appointments of each of the Justices, we achieved around 60-65% accuracy on average for all the judges using a purely text-based understanding of an oral argument. We could still improve on this average by performing hyper-parameter tuning using the recursively eliminated feature vector, as this way we would be tuning on a set of features that isn't overfit to the training data. Additionally our models hasn't yet combined all features, which could increase accuracy. Yet, our success thus far nevertheless informs us that conversational dynamics are indeed prevalent in Supreme Court cases.