Abstract

This paper is an exploration into the effect of prosodic and lexical features in political speeches on the success of presidential candidates. Past work has been done detailing different features from speeches that indicate a candidate's persuasiveness and charisma. This project applies the tools of spoken language processing to attempt to extend that research into the domain of actually predicting a candidate's political success. For the data, a corpus of speeches is used, taken from presidential campaigns from 1952 to the present and tagged with a label indicating whether the candidate won or lost. Features such as speech rate and pitch information, MFCC features, and unigram word counts, are extracted from the data and used as input. A baseline model using a multilayer perceptron network with all the features managed to get an accuracy higher than random guesswork, while a slight improvement over the baseline was achieved with a long short-term memory network using only MFCC features in the time dimension for each input. Statistical methods failed to find significant differences between the datasets of the winners and the losers, but the models were able to perform slightly better than expected. Future work could extend the dataset to include primary candidates and could also explore the lexical features of the speeches in more depth.

1 Introduction

Many Presidents are remembered through their speeches. Abraham Lincoln is remembered for the Gettysburg Address, and Franklin D. Roosevelt for his fireside chats. John F. Kennedy has his quote “Ask not what your country can do for you—ask what you can do for your country.” More recently, Barack Obama even inspired an article titled “6 Times an Obama Speech Made Us Stop and Say ‘Wow.’” Every four years in the United States, political candidates travel around the country to host rallies and give speeches to win votes. Being a great orator seems to be an important part of being President, and we wondered if there was a difference in the campaign speeches given by candidates who eventually won the Presidential nomination versus those who lost.

We focus particularly on searching for the existence of distinctive prosodic features by examining candidates’ nomination acceptance speeches at their parties’ national conventions a few months before the election. Every four years, during an election year, political parties meet for a few days to officially nominate a Presidential candidate, in addition to setting party platforms and guidelines. The acceptance speeches nowadays last close to an hour, and their purpose is to establish the “what is at stake in the forthcoming election” and “build majority support” (Galston). We collected these speeches from 1952 through the present day, taking one speech if a candidate ran for a second term, resulting in a dataset consisting of 11 winning candidates and 17 losing candidates from the Democratic or Republican party (with the exception of Gary Johnson, the 2012 and 2016 Libertarian candidate).

With the potential discovery of distinguishing features, our goal was to use them to build a classification model that, given a speech, will predict whether the speaker won the Presidency or not.
2 Related Work

There has been much research on the relationship between prosody and affective meaning, defined by Oxford as “the personal feelings, attitudes, or values of an author or speaker inferred from their words and/or nonverbal behavior” (Scherer 1996; Wichmann 2002). As Strangert writes, “These studies in which prosody is seen as the main contributor to expressing ‘the meaning’ of the message have great relevance for the present work. In addition, prosody itself—speech rate, intensity, fluency etc— is important for characterizing a speaker, as well as expressions of emotions” (2005). We particularly focus on work done on the effects of prosodic features on perceived qualities like charisma and persuasiveness.

For acoustic features, two of the earliest papers we found dated back to the 1970s, and both studied the effects of speech rate on persuasiveness. Both concluded that a higher speaking rate corresponded to a higher perception of intelligence as well as higher persuasive abilities (Apple, 1979; Miller, 1976). One of the studies also examined pitch and found that the lower the pitch the speaker had, the more persuasive he seemed. They hypothesized that lower speech rate lead listeners to think the speaker is less persuasive, fluent, and emphatic and that increased pitch suggest the speaker is nervous (Apple, 1979).

We also found many studies relating prosody to politics. One study conducted by Andrew Rosenberg and Julia Hirschberg gave subjects political speech segments from the nine candidates running for the Democratic party presidential nomination in 2004 and had them indicate their level of agreement with a set of statements about each speaker. In the results, subjects were relatively in agreement about the charisma of a given speaker. The speaker of the subject greatly influenced the charisma rating, as did the genre (when delivering stump speeches, speakers were rated as more charismatic). They found that mean, standard deviation, maximum, and minimum F0 for all male speakers as well as speaking rate “positively influenced ratings of charisma.” Also, speakers with more variation in pitch and sound intensity were rated as more charismatic.

In addition to auditory features, we looked at studies that incorporated lexical analyses, like the aforementioned Rosenberg and Hirschberg paper included lexical properties. On top of standard features, the researchers included “the number of words in the token, ratio of function to content words, pronoun density, and a measure of lexical complexity.” They found that the longer the speech, the more charismatic the speaker appeared. Interestingly, they also discovered that higher usage of first person personal pronouns (I, me, my, etc.) also influenced perception of charisma.

There is also a study that uses a tagged corpus of political speeches with labels such as \{APPLAUSE\} and \{BOOING\} to measure persuasion. Their hypothesis is that audience reactions can be used to measure the speaker’s persuasiveness and that they could, for example, predict sentences that would trigger applause. They conducted many experiments with the corpus; the first was testing if the transcripts could be used to distinguish Democrat from Republican, which they found was possible with only four-sentence chunks. The second was attempting to predict sentences that preceded a positive reaction like applause or standing ovation. They found “persuasive impact of speeches are quite general and, as shown in the cross-classification results, to a certain degree independent from the party of the speakers.” This classifier was used for the 2008 election year and showed that both Obama’s and McCain’s speeches had highly persuasive content, with Obama’s scores being higher. With these results, they concluded that there is the possibility of predicting the impact of a text (Strapparava, Guerini, and Stock 2010).

We did not find any previous work that attempts to specifically determine political success based on prosodic and lexical features. The idea to combine the extraction of prosodic features with deep learning to determine political viability seems to be a new area of research.

In summary, multiple studies have shown that prosodic and lexical features affect political speeches. For example, a higher speech rate, lower pitch, and more varied pitch contribute to increased persuasiveness and charisma, and on the lexical side, more first person personal pronouns and longer speeches also contribute positively.

3 Approach

Our project is an attempt to predict the success of American political candidates based solely on information gained from campaign speeches. The
project uses a variety of different features as input to multiple different models to help inform the predictions.

3.1 Feature Extraction

We used a wide variety of features to inform the models. Some of our features were purely prosodic, while others were lexical.

Figure 1 below gives a map of the features being used. As can be seen, raw MFCC features were extracted from each speech. In addition, Praat was used to extract relevant pitch information from the sound files. The pitch features used include average pitch, maximum pitch, minimum pitch, and standard deviation. Praat was also used to extract speech rate information from the sound files. Both speaking rate and articulation rate were used as features. Note that speech rate is calculated using the total time of the speech, while articulation rate is calculated after removing the time taken in silence during the speech (only the time taken up with actual words is used). On the lexical side, we also tried multiple different features. One set of features used was a unigram model of word frequency in the speeches. Additionally, sentiment analysis was performed on the speeches and used as features. Lastly, the length of the speech was used as a feature as well.

3.2 Statistical Evaluation

In order to better understand our features, we performed a variety of statistical tests on the features to attempt to see if there was statistically significant differences between the two categories of speech. Specifically, a Levene test was performed on all of the features except the lexical data to check for differences in variance, and a Wilcoxon Rank-Sum Test was also performed to check for differences in the median piece of data.

These tests were chosen because the features could not be assumed to be normally distributed, which is the assumption necessary for simpler versions of computing statistical differences like a t-test.

In addition, we performed a statistical analysis of the lexical features, which included both looking at the most common words (excluding stop words) and the number of personal pronouns used in the speech. Previous research showed that the use of personal pronouns and number of words affected how people rated speakers’ charisma with $\rho = 0.023$ and $\rho = 0.026$ respectively (Rosenberg and Hirschberg 2005), and since Democratic and Republican speeches could be distinguished given four-sentence chunks, we also thought there would be a difference in word usage between winners and losers.

3.3 Modeling

First, we created a baseline model in which we fed the various features to a multilayer perceptron neural network. The network has two hidden layers, and the output is a binary value corresponding to whether or not the network predicts the input’s orator will win their election. Limited-memory BFGS is used as an optimizer, while the hyperbolic tan function is used to compute activations.

This model used all of the different features as input (MFCC, pitch, speech rate, and lexical). We also implemented an SVM to verify the success of the baseline.

The other main model used was a long short-term memory recurrent neural network. The LSTM was implemented with two layers—a LSTM layer that reduces the output space to two activation values. Then, a dense layer using a sigmoid activation function is used to attempt to classify the data into winners and losers. A dropout value of 0.2 is included in order to guard against overfitting.
The input to the LSTM (as can be seen in Figure 1) is simply the MFCC features. However, the features had to be reshaped for this model. The input is a three-dimensional matrix. The primary dimension is the speech being given. Then, the secondary axis is time, followed by the MFCC features. In other word, the input is a list of the 13 MFCC features for each of 500 time steps in each speech. The 500 time steps were calculated by averaging the more specific MFCC values on smaller time steps from each speech.

4 Experiments

4.1 Dataset

The corpus that is used for this project is a set of 28 campaign speeches from American politicians dating from the mid-1950s to the present day. Specifically, the politicians used are presidential candidates from the two major parties (the only exception is two speeches from Gary Johnson, of the Libertarian party). The speeches taken are the acceptance speeches at the respective conventions after the nomination. The speeches range in length from around ten minutes to longer than an hour, since the acceptance speeches have been becoming longer and longer over the past half a century. The speeches are split into two categories based on whether or not the candidate giving the speech won the presidential election that they were campaigning for at the time of the speech. Transcripts from the speech were also used to help extract the lexical features.

4.2 Evaluation

Our main evaluation metric is relatively simple, given that the main goal of the project is a simple binary classification problem (to predict whether or not the candidate who gave a speech will win the election). Thus, the main metric used was simply the number of speeches classified correctly, divided by the total number of speeches in the test set. Given the relatively small number of politicians that we acquired speeches from, we decided to randomize the results to make them more accurate. In order to do this, we ran batches of tests where in each batch a random subset of the speeches were chosen to be training data, while the rest were reserved for the test set. Multiple of these tests were run, and the average success over the entire batch was used as our primary metric of success.

The testing was performed using k-fold cross-validation, in which the dataset is partitioned two different subsets. One subset is used for training, while the other subset is used to test the results. This process is repeated on different complementary subsets in order to insure that the test results will generalize well to other data. We used this method on our model because the nature of our task would easily allow for overfitting. The use of this method helped us have more confidence that the results are generalizable.

4.3 Results

The results of the prosodic statistical analysis can be seen in Table 1. Listed are the $\rho$-values of each test. Note that in no case are any of the features statistically significant in between the winners and the losers datasets (normally the $\rho$-value must be below 0.05 for significance). This could be a result of the relatively limited nature of the dataset, small differences in between datasets are more meaningful with large sample sizes.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Levene $\rho$-value</th>
<th>Wilcoxon $\rho$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>0.81466248888</td>
<td>0.869226298639</td>
</tr>
<tr>
<td>min.</td>
<td>0.155490150916</td>
<td>0.621359861767</td>
</tr>
<tr>
<td>max.</td>
<td>0.44899683343</td>
<td>0.48043379148</td>
</tr>
<tr>
<td>std. dev.</td>
<td>0.477712626031</td>
<td>0.79584967464</td>
</tr>
<tr>
<td>speech rate</td>
<td>0.568009372003</td>
<td>0.943747492813</td>
</tr>
<tr>
<td>artic. rate</td>
<td>0.789363072652</td>
<td>0.72423624736</td>
</tr>
</tbody>
</table>

Table 1: Praat Features Statistics

The results of the lexical analysis can be seen in Tables 2 and 3. The only words that are highly different between the two are the word one (seen more often in the winners dataset) and the word president (seen more in the losers dataset). As can be seen, the difference in between the two sets is not significant. In addition, the prevalence of personal pronouns in the winners and losers datasets were also not too different (0.061096 in the winners to 0.060547 in the losers).

The results from the baseline and LSTM models can be seen in Table 4. Note that the baseline model, using all of the features, achieved an accuracy of 55.3%. The LSTM model using 500 MFCC samples improved slightly on the baseline model. It achieved an accuracy of 56.7% on the test candidates. However, the LSTM using 5000 MFCC features actually performed less well than the baseline, with an accuracy of 53.3%. We spec-
Given the numerous complex factors that exist during a presidential election. In addition, the nature of the data also makes the task quite difficult. The dataset used more than 12 gigabytes of audio data, and extracting MFCC features took huge amounts of time (over 10 hours) as did training some of the models. However, the results from this project do seem to suggest that prosodic features can have an impact on the success of a candidate. Given that the problem is a binary classification problem, an accuracy of 56.67% on the test set is not extremely promising, but the test methodology (using speakers not in the training set to test the models) helps ensure that the models were not just learning the speaking characteristics of specific candidates and recognizing those candidates. Thus, the results do seem to show that there exists a (albeit tenuous) connection between a speaker’s prosodic and lexical features and their success in the election.

Moving forward, we would like to see more research done on this topic. One of the major changes we would implement is expanding our dataset and project scope. A problem with our current sample is that there is no way to increase the number of winners, since there is a fixed number of Presidents with audio files of their nomination acceptance speeches. To solve this, we can change our scope to look at classifying candidates as they vie for their party nominations, such that winners are defined as those who win the party vote as opposed to the entire presidency.

In addition to gathering more data, we would also make the data better quality. For this project, we extracted audio files from YouTube videos that included noise such as applause and music at the beginning and ends. We did our best to trim the speeches right when the speaker began and ended, but it would be more robust in our analyses to have data that we know was cut exactly at the same start and end times relative to the speech since the RNN is based on time-oriented data; this way, we know that one time sample in one audio file corresponds to the same point in time in another audio file. Instead of trying to remove the noise or letting it sit in our data as it currently does, what we could also do in a future model is incorporate the noise in our features or at least normalize the MFCCs rather than using raw data. In transcripts, we could include tags such as `{APPLAUSE}` and use them as a measure of persuasiveness, as was done in another study we found for our background research.
Our current model only uses MFCC features in the RNN, but we would like to combine multiple neural networks to also include lexical features as a component in classifying a speech, since our current model requires time-spliced data. In addition, there can be an attempt to model the lexical features more fully. Our investigation of the lexical features did not reveal many differences between winners and losers, but there were interesting differences in collectivism ("person" is the top word in the winners, while "people" is among the top words in the losers) that could be more fully explored in future models. In general, it would be interesting to perform a more comprehensive review of political candidates in general, in order to more fully inform candidates moving forward.

Though this paper did not yield remarkable results, it is still promising given our dataset and sample size, and there are a myriad of ways to improve and move forward.

6 References


