



Classifying Presidential Candidates' Speech Patterns



Brendan Corcoran and Kyana Van Houten
CS 229: Machine Learning

We characterized the speech patterns of leading candidates for United States President to determine which features are most effective at differentiating the candidates. We compiled full-text transcripts of stump speeches, debates, and interviews from the current election cycle. Then, we determined appropriate indicators from the raw text. Our algorithm constructs a model or "fingerprint" for each candidate's speech style which we use to predict the speaker of a new text excerpt.

Feature Selection

We collected campaign speeches and interviews of the last five candidates running in the 2016 Presidential Election. From these text excerpts, we measured features of two categories:

What a candidate speaks about:

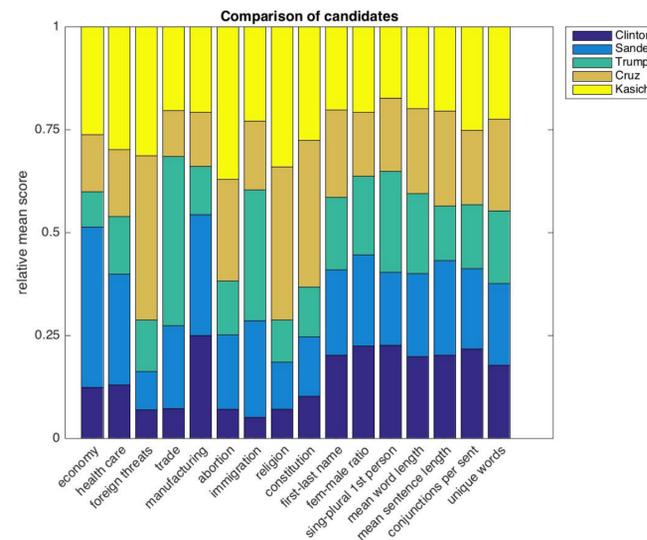
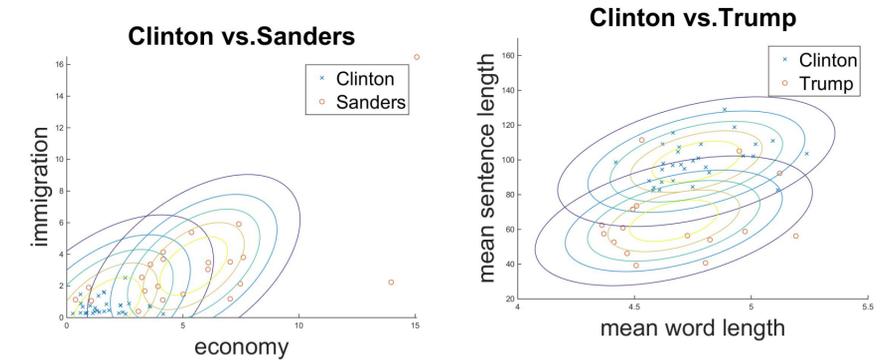
- Economy
- Health Care
- Foreign Threats
- Trade
- Manufacturing
- Women's Health
- Immigration
- Religion

How a candidate speaks:

- Mean word length
- Mean sentence length
- Mean number of conjunctions per sentence
- Number of unique words per total words
- Relative usage of female and male pronouns

Algorithm

We want to teach the algorithm to discern the speaker from these features. We decided to use Gaussian discriminant analysis (GDA), a generative learning algorithm that builds a model of what each candidate sounds like. This takes the form of a multivariate Gaussian where each feature is a dimension. When a new speech is received, the algorithm compares it to each of the candidates' models (Gaussians) and determines which model it is most similar to. We selected this method due to the observed normal distributions of several features and the limited number of training examples available.



Results

We achieved a training error of 29% and a test error of 31% both of which are well below random guessing (80%). We also measured several interesting results about each candidate's speech patterns. For example, Bernie Sanders spoke the most about the economy. Ted Cruz spoke the most about religion, the Constitution, and foreign threats. Donald Trump spoke the most about immigration and had the shortest mean sentence length. Trump tends to use first person singular tense significantly more than Cruz. The two Democrat candidates tend to speak more about manufacturing than the Republican candidates.

I'm a progressive, but I'm a progressive who gets things done.



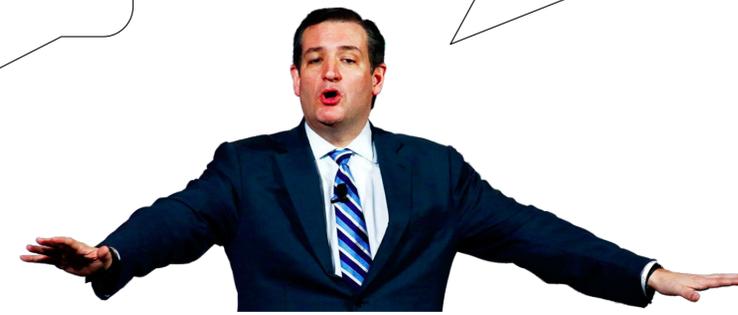
I have spent my life taking on the billionaire class and the special interests.



They're bringing drugs. They're bringing crime. They're rapists.



It is time to reclaim the Constitution.



I seem to be the only adult on the stage.

