On the Classification of 2016 Republican Presidential Candidates Using Debate Transcripts

Matthew Scicluna

Wednesday, December 23, 2015

Abstract

We collected transcripts from each of the four televised Republican debates and converted the word frequencies of each of the candidates' debate statements into a low dimensional representation using LSA. We then built a classifier which attempted to distinguish statements made by each of the leading presidential candidates to Jeb Bush. We compared the accuracy of each the classifiers to see which candidate is easiest to distinguish from Bush. We then build a single classifier using the previous classifiers which can determine, given a statement, which candidate most likely said it. We illustrate this technique by providing as a demonstration a function which allows user inputted documents to be visualized in two dimensional latent space along with each of the candidates statements.

Introduction

The 2016 Republican primaries has been covered extensively in the media recently, especially with the addition of celebrity and businessman Donald Trump. The news cycle is full of stories about Trump and other high profile candidates like Governor Jeb Bush, Senators Marco Rubio and Ted Cruz and retired Neurosurgeon Dr. Ben Carson. A recurring story in the press is that the addition of candidates like Trump, who is perceived by many to be lacking substance, is causing the debates to resemble predictable stump speeches rather than spirited arguments. We would like to measure the predictability of candidates' lines in the debates to see if the press is correct in this assertion. While judging statements based on informational content is beyond the scope of this study, we can restrict ourselves to looking at the predictability of each of the candidates speaches? as a proxy for informational content. The more predictable a candidate is, the less informative content they have. We can see the predictability of the candidates based on whether we can predict their speeches accurately using a simple model, which in our case will be logistic regression.

We compared each of the aforementioned candidates to Bush, since he is seen as the most mainstream establishment candidate. Classifiers were trained for each candidate to distinguish that candidate from bush, thus producing four sets of classifiers. The accuracy of each classifier was compared to see whether they are significantly different, and if so, which ones were the most accurate. Finally, we combined the four classifiers to build a single classifier which can predict which of the five candidates said which debate line.

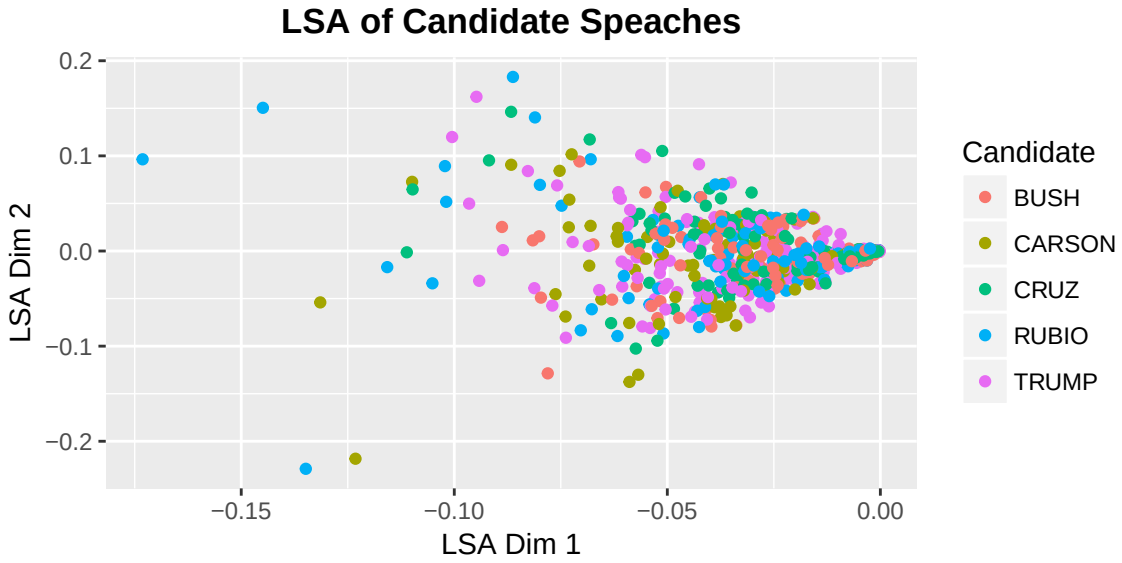

Figure 1: Each of the Candidates Statements in 2D Latent Space

Methods

So far debates between the candidates have been relatively common, with long running transcripts freely available online. We collected transcripts for each of the televised Republican debates from the Time magazine website using Python and the Beautiful Soup package. We inputted the data into R and cleaned the transcripts by applying transformations like removing stop words, stemming words, removing case sensitivity, and removing sparse terms. We then built a term document matrix and applied LSA to it. Once this was done we fit a logistic regressions to the lower dimensional representation of the term document matrix to try to classify speeches based on if they were said by each of the candidates or Jeb Bush.

We illustrate this technique by building a function called *visualize* which can project the each statement into two LSA dimensions. The user can input their own phrases into the function arguments. Each word in each inputted statement will be compared to each word which appears in any of the debate transcripts. Each partial match will be recorded and the statement will be converted into word counts and then projected into the lower dimensional space, along with the candidates?? real statements. It is important to note that for our classifier, we are using much more than 2 dimensions in our latent space, so the following should only be seen as a demostration.

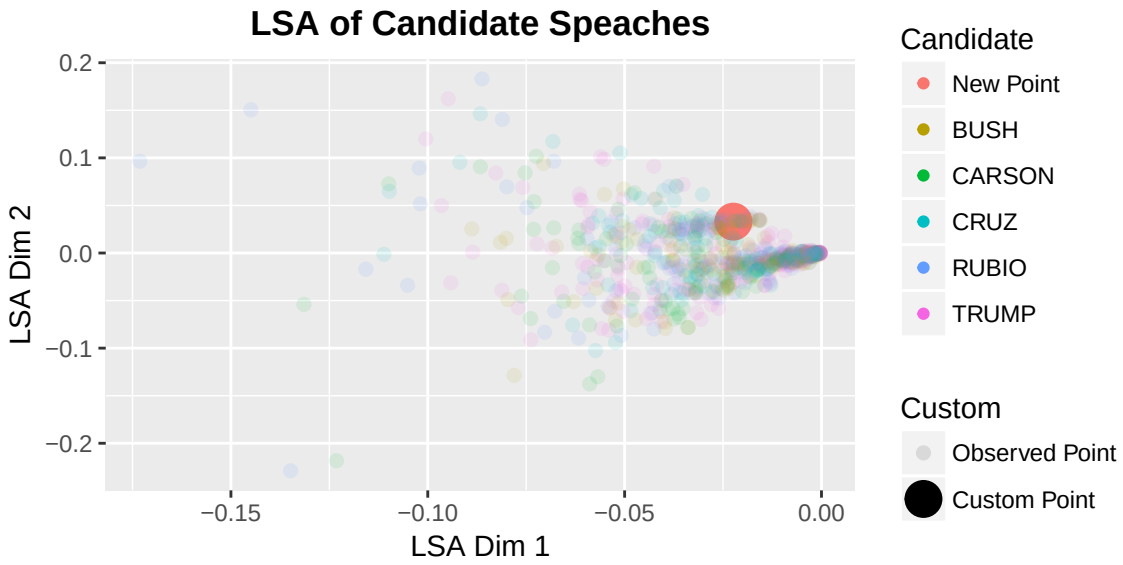

Figure 2: Visualization Function Demonstration

Results

We trained 30 seperate classifiers using stratified random sampling and leaving out 20% of the statements to be used as a test set each time. The accuracy of each model was recorded along with which candidate the classifier was attempting to distinguish from Bush. It should be noted that the accuracy from each of these classifiers was found to be between 60-80%, which is impresive for a classifier of such simplicity. A one way ANOVA on how much each candidate affected the accuracy of their classifiers was performed and the results can be seen in table 1.

Table 1: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Candidate	3	0.27	0.09	20.74	0
Residuals	116	0.50	0.00	NA	NA

To figure out which candidates were easier to distinguish from Bush, we did follow-up tests on all pairwise differences, controlling for type 1 error using Tukey HSD procedure. We found that Marco Rubio and Donald Trump had the largest disparity between classifier accuracy. while the other differences were not statistically significant.

Table 2: Tukey Controlled Pairwise Comparisons

	diff	lwr	upr	p adj
CRUZ-CARSON	0.07	0.03	0.12	0.00
RUBIO-CARSON	0.07	0.02	0.11	0.00
TRUMP-CARSON	0.13	0.09	0.18	0.00
RUBIO-CRUZ	-0.01	-0.05	0.04	0.98
TRUMP-CRUZ	0.06	0.01	0.10	0.00
TRUMP-RUBIO	0.06	0.02	0.11	0.00

Finally, we see that the combined classifier was not very accurate, as can be seen in table 3. This is surprising since each individual classifier was quite impressive.

Table 3: Confusion Matrix For Overall Classifier. Rows are Predictions, Columns are Reference

	BUSH	CARSON	CRUZ	RUBIO	TRUMP
BUSH	0	0	0	0	2
CARSON	2	4	9	2	9
CRUZ	0	0	1	1	3
RUBIO	8	23	28	11	53
TRUMP	0	0	0	0	0

Discussion and Conclusions

While the overall classifier was not very accurate, each of the seperate ones were, indicating that we can build a reasonably accurate classifier which can distinguish candidates from Jeb Bush using only word counts. We believe that this can be replicated for the other candidates not considered here. We also believe that if we consider more complex models, such as autoencoders, we can produce even more accurate classifiers.

There were significant differences in each of the classifiers abilities to classify each of the candidates from Bush. Perhaps not surprisingly the biggest difference in classifier accuracy is between the classifiers of Marco Rubio and Donald Trump. This indicates that the differences between Trump and Bush and Rubio and Bush were the most different. This is not very surprising since Jeb Bush and Marco Rubio are considered to be in the same establishment faction of the GOP, while Donald Trump is considered to be the biggest outsider.

Future directions include finding better proxys for candidates' informational content, and developing a better overall classifier, perhaps using a neural network with a 5 way softmax output. More advanced techniques in data sampling and modelling should be considered.

Overall, we do see that candidates are quite easy to distinguish based on their word counts. This misfortunately supports the claims in the media that the debate lines are predictible. While this may lead to bad domestic policy in the United States, at the very least it is interesting to look at with a statistical lens.