Analysis and Prediction using Lexical Data for US Presidential Election Speeches

CULMINATING PROJECT FOR CISC 251 AT QUEEN'S UNIVERSITY

Author: Alex Beamish

1 Introduction

The focus of this analysis can be broken down into several distinct goals. One goal of the analysis is to develop a model that can accurately predict election outcomes based on frequency data for words in election speeches. Since this is a typical supervised learning task, several different types of machine learning algorithms can be effectively applied. While development of models is an important part of this analysis, it is not the only focus. Another goal of the analysis is to find specific words that are associated with winning candidates. A final goal of the analysis is to determine the extent to which deceptive language contributes to a candidate's chances of winning in the election. All of this information can be used to develop strategic insights for candidates. These problems can be effectively tackled by exploring the data using correlation metrics, clustering and comparison among clusters. Data exploration techniques such as clustering and correlation analysis are often thought of as precursors to model development. In this analysis, they not only serve that purpose, but they also provide important insights related to the primary questions at hand.

2 Preliminary Steps

2.1 Data Preparation

In the beginning, various sets of data were combined into a single data table. Each row of the data table contained the word frequency data (see *Section 6.1* for definition) for a single election speech. The rows (records) were labeled with row headers containing the name of the candidate who gave the speech. The columns (attributes), with the exception of the final column, contained all frequency data for a specific word. These columns were labeled with column headers containing both the specific word and the part of speech of the word (see *Section 6.3*). The final column (target attribute) held values of 0 or 1 depending on whether the candidate corresponding to that row ultimately won the election (0 if the candidate lost, 1 if the candidate won). The target attribute was converted from a numerical attribute to a categorical attribute with two possible values (classes) for model development, but not for correlation calculations.

Even though all of the data values were already between 0 and 1, it was not easy to compare attributes because the attributes had different means and standard deviations (i.e they were not standardized). This issue was remedied through standardization (z-score normalization in Knime) prior to clustering and calculating correlations. When developing the prediction models, the data was sometimes standardized (e.g. for the Random Forrest and for the Support Vector Machine), and was sometimes left unaltered (e.g. for the Neural Network).

At several points in the analysis, smaller data tables were created. These smaller tables contained data for only a subset of the attributes. Some of these subsets were based on the part of speech (POS) of the attribute. For example, one of the tables contained frequency data for only nouns. Another subset contained only the data for words for which changes in frequency of use is associated with deception (see Section 6.1 for definition). The subsets were generated in Knime using a regex in the "column filter" node.

2.2 Initial Data Exploration

To begin, the mean value was calculated for each attribute, purely for the purpose of learning about the data. Following this, the mean value of each attribute for all records associated with winning candidates (these will be referred to as *winning* records) was compared to the mean value of the same attribute for all records associated with losing candidates (these will be referred to as *losing* records). This calculation and comparisons of mean word usage rate between winning and losing candidates did not lead to any

interesting results. This suggests that there is no single word that is almost always used by winning candidates and almost never used by losing candidates. If there were such a word, then there would be an extreme difference in the mean frequency of this word between winning candidates and losing candidates. Instead, it seems more likely that there are several words that may be correlated with winning the election, and that winning candidates may use any subset of these words. The mean is simply too crude of a measure to capture the full complexity of the problem, but it is useful as a starting point for further analysis.

3 Clustering and Correlation Analysis

3.1 Clustering

Initially, the records were clustered based on the values of all of the attributes in the main data table. This was done using the k-means algorithm, and k was chosen to be 5 because this value maximized the silhouette coefficient (0.077). In clustering the attributes, one would hope to obtain clusters of speeches that highlight differences in types of speech, candidate personality or lexical strategies. However, partly due to the large number of attributes, the initial k-means clustering did not produce a useful result of that nature. This is evidenced by the fact that the silhouette coefficient is very close to 0 even for k=5. One problem with clustering the data using a large number of attributes is that it obscures the similarities between records on specific subsets of attributes that might be relevant. In an attempt to solve this problem, the attributes were divided into groups based on part of speech (POS), and a separate clustering of the records was done using each subset of the attributes. However even these clustering attempts did not produce a mean silhouette coefficient above 0.1. The difficulty in clustering the data is to be expected since there are many ways to choose words for an election speech, not just a handful of well-defined word-choice strategies. Despite the failure to obtain useful clusters, clustering was a useful primary step in data exploration; at the very least, it suggests that the prediction problem is fairly difficult.

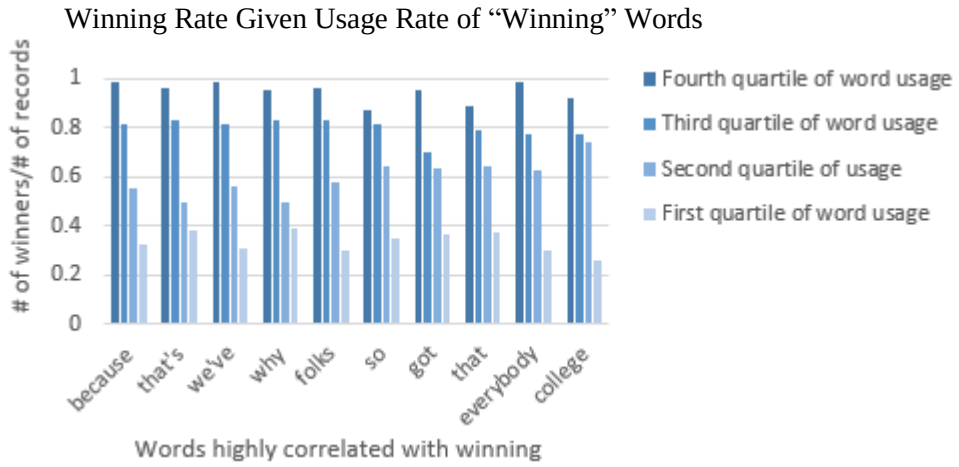
3.2 Correlations

After clustering was attempted, the correlation between attributes was examined by calculating Pearson's product-moment coefficient for each pair of columns. Most of the attributes with the greatest degree of linear correlation were words that form an easily recognizable phrase when compounded. Examples include the words "wall" and "street", "health" and "care", and "middle" and "class" (correlation values of 0.94, 0.92, 0.88, respectively) . Sometimes it is beneficial to remove attributes that are highly correlated with other attributes, as this can decrease the total number of attributes. However, attributes were not removed in this analysis so as to avoid information loss.

Correlation of attributes with the target attribute (i.e. the election result) was also examined, once again using Pearson's product-moment coefficient. One interesting result that came out of this was that the word "Obama" was the word most negatively correlated with winning the election. The word "Clinton" was also negatively correlated with winning, while the word "Mccain" was positively correlated with winning. The obvious explanation for these observations is that the candidate who ultimately wins the election will typically mention the other (loosing) candidate in their speeches, while the candidate who ultimately loses will typically mention the other (winning) candidate. After these correlations were noticed, candidate names were filtered out of the list of attributes using a regex. Including the names of candidates would have decreased the validity and usefulness of the prediction models.

After Pearson's correlation coefficient had been calculated between each attribute and the target attribute, the attributes were ranked from most positively correlated with the target attribute to most negatively correlated (see fig. 1 in Section 6.2). Following this, the 10 words most positively correlated with the

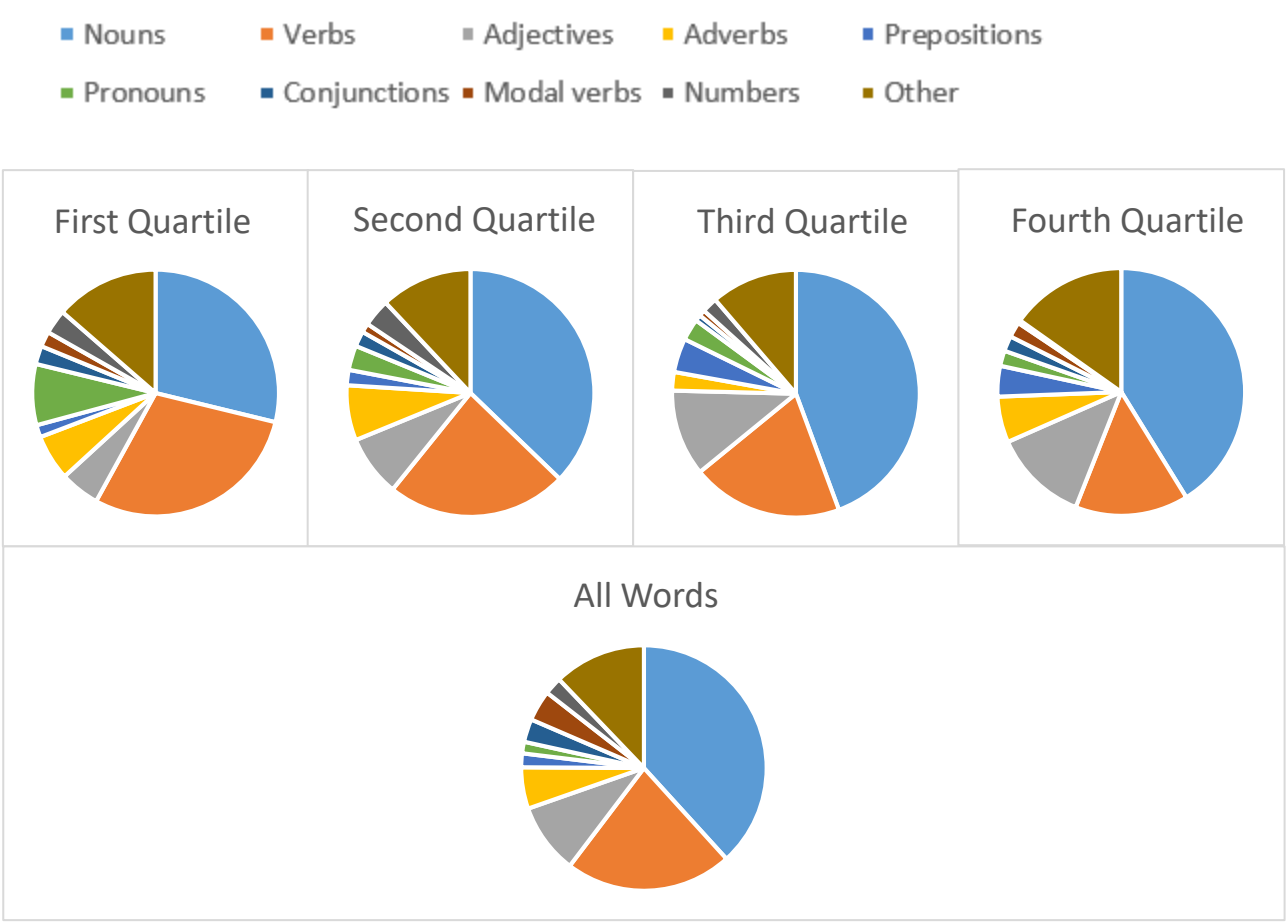
target attribute were selected for further analysis. For each of these words, the records were divided into quartiles based on the usage rate of the word in that record. The first quartile contained the records with the lowest usage rates for that word, while the fourth quartile contained the records with the highest usage rates (note that a record can be in different quartiles for different words). Since the data was standardized at this point, the different quartiles correspond to different distances from the mean usage rate of the word. For each word, and for each quartile for that word, the rate of winning records in that quartile was calculated. The rate of winning records is the number of winning records divided by the total number of records in that quartile. The results of these calculations are displayed in the following bar chart.



The data in the above bar chart can be thought of as the likelihood that a candidate wins the election given the quartile they fall into for usage of a specific word. Of course, the usage rate of a single word is a ridiculously small amount of data to use for the prediction of an election result, but the starkness of contrast between quartiles for words highly-correlated with the target attribute is remarkable. In *Section 3.5*, the above data will be used to derive strategic insights.

3.3 Part of Speech Analysis

Each word in the data set had already been labeled with a tag (see *Section 6.3*) denoting theappropriate part of speech (POS). A natural question is whether the distribution of POS (the relative number of nouns, verbs, adj., etc.) is different for a sample of words that are more correlated with the target attribute than for a sample of words that are less correlated. If it is different, then that would indicate that words of a certain POS might be more predictive of winning, and this could have strategic implications for political candidates. The following pie charts show how the relative number of different types of words changes as one samples words from different quartiles of correlation with the target attribute. Note that the quartiles in this case are different than the quartiles in *Section 3.2*; here the first quartile is the quartile of attributes most negatively correlated with the target attribute, while the fourth quartile is the quartile of attributes most positively correlated with the target attribute. The POS distribution of the entire data set is also shown in the following figure.



The above pie charts clearly show that the proportion of nouns in the third and fourth quartiles is larger than the proportion of nouns in the first and second quartiles. In other words, there are a relatively high number of nouns that are highly positively correlated with winning the election, and relatively fewer nouns that are highly negatively correlated with winning the election. The opposite is true of verbs. These results are analyzed in *Section 3.5*.

3.4 Deceptive Word Analysis

To examine the role that deception plays in US election outcomes, the correlations between "deceptive" words (see *Section 6.1* for definition) and the election outcome were compared to the correlations between "non-deceptive" words and the election outcome. To accomplish this, the main data table was split into two tables: one for the deceptive words present in the main table and one for the non-deceptive words present in the main table (see note in *Section 6.3*). Then, for both data tables, Pearson's product-moment coefficient was calculated between each word and the target attribute. The results show that the deceptive words have an average positive correlation value of 0.143 and an average negative correlation value of -0.104, while the "non-deceptive" words have an average positive correlation of 0.116 and an average negative correlation of -9.29×10^{-2}. This indicates that, on average, varying the rate at which one uses "deceptive" words is likely to have a larger impact on the election outcome than varying the rate at which one uses "non-deceptive" words. Since variation in the use of "deceptive" words is associated with deception, this strongly suggests that a candidate's use of deception in their speeches has a significant impact on their chances of winning the election.

3.5 Strategic Insights

After attempting clustering, calculating correlations, and examining the roles that factors like POS and deceptiveness play in determining election outcomes, it is possible to derive some strategic insights for political candidates.

Firstly, if one examines the correlations between all attributes and the target attribute, one might notice that the words “because”, “why” and “so” are among the words most positively correlated with winning the election. One explanation for this is that these words might indicate that the candidate is offering an explanation of some sort, and this might make the candidate seem intelligent or knowledgeable. Another notable fact about the list of words highly positively correlated with winning the election is that a few of these words are somewhat informal. Examples of these “informal” words include contractions like “that’s” and “we’ve”, and colloquial words like “folks”. Perhaps these words make the candidate come across as more likeable or friendly. In addition to looking at the words positively correlated with winning, it is important for candidates to look at the words negatively correlated with winning as well, as this can give a clue about which words to avoid. Interestingly, two of the words most negatively correlated with winning the election are “government” and “elected”. This could imply that voters negatively judge candidates who talk too much about politics. Further analysis would be required to verify any of these speculative theories, but they all seem plausible.

The POS analysis showed that the proportion of nouns is higher for a sample of words with a high positive correlation to the target attribute compared to the proportion of nouns for a sample of words with a high negative correlation, and that the opposite is true for verbs. This means that for many nouns, increasing usage rate can increase a candidate’s chances of winning the election. On the other hand, for many verbs, decreasing usage rate can increase a candidate’s chances of winning. There are many possible explanations for this phenomenon, and one should be careful not to speculate too much about it because the POS analysis omits information about specific words, which might be highly relevant. However, it should be noted that nouns often carry more substantive information than verbs, and repeatedly using a certain noun might reinforce this substantive information. Repeatedly using a certain verb might not have the same effect, and might make the candidate seem repetitive and boring.

The final strategic insight is that deceptiveness likely increases a candidate’s chances of winning the election. Whether a candidate ought to be deceptive in their speeches is not the focus of this analysis; that is a moral issue for the candidate to contemplate on their own. However, candidates should be aware that choosing to not be deceptive at all could put them at a disadvantage. On the other hand, this analysis does not examine the effects of attempting to be deceptive but either failing to actually deceive people or having hidden truths uncovered by another person or group. Failed deception could have a significant negative effect on a candidate’s winning chances. More research is needed to examine both the benefits and the risks of deception for candidates.

4 Classification Models

4.1 Supervised Learning Models

Several classification models were developed in order to predict the outcome of the election based on the word frequency data in the main data table. Each of the 3 models was developed using a different machine learning algorithm. In each case, 5-fold cross validation was used to ensure that the model’s performance was consistent. Using multiple different algorithms ensures that any peculiarities in the results of individual models are identified. The following table shows information about each of the 3 models, including basic performance metrics.

Model Type	Parameters	Correct Predictions	Incorrect Predictions	Prediction Accuracy
Multi-layer Perceptron (MLP)	Hidden Layers: 2 Hidden neurons per layer: 10	340 (279 winners, 111 losers)	41 (31 losers predicted to win, 10 winners predicted to lose)	90.5%
Random Forrest	Split Criterion: Information Gain Ratio Limit on Tree Depth: 15 Number of models: 100	354 (281 winners, 73 losers)	77 (69 losers predicted to win, 8 winners predicted to lose)	82.1%
Support Vector Machine (SVM)	Overlapping Penalty: 2.0 Kernel Type: Polynomial Power: 1.0 Bias: 1.0 Gamma: 1.0	413 (282 winners, 131 losers)	18 (11 losers predicted to win, 7 winners predicted to lose)	95.8%

Prediction accuracy is not an ideal metric due to the fact that there is not an even number of winning records and losing records in the dataset. In fact, approximately 67% of the records are winning records. Due to this imbalance, precision, recall, and the F-Measure are more useful as performance metrics than prediction accuracy. These metrics are calculated for the SVM model in *Section 4.2*.

4.2 Evaluation of Models and Next Steps

All three models were fairly successful in predicting election results. The SVM had the highest classification accuracy (95.8%) out of the three models. The precision for the SVM model is 0.96 and the recall is 0.98. These values can be used to compute the F-Measure, which is 0.97 for the SVM model (see *Section 6.1* for formula). The fact that the F-Measure is very close to 1 indicates that the SVM model performs very well on the data.

Theses initial models could likely be improved through the use of ensemble techniques such as bagging and boosting. In general, combining the strengths of multiple models will improve performance, and this is the main purpose of ensemble techniques. Another interesting step would be to train models using only

words of a certain POS, or only words associated with deception. This might provide further insight into the predictive power of certain types of words compared to other types. Neither of these steps were taken in this analysis due to time constraints, and due to the large amount of time and effort exhausted in the data exploration phase.

5 Discussion

The models developed were obviously successful in predicting election outcomes, but could be further refined. Now, it is important to take a step back and interpret the results in a larger context. It is important to consider the goals one has when analyzing the speech data. If one's goal is to simply obtain a predictive model for past US elections, then considering all types of words used in election speeches makes sense. However, if one hopes to generalize results to elections other than US presidential elections, or even to future US presidential elections, one must carefully consider details such as the role of certain types of words in speech, and particularly in election speeches.

For example, nouns often convey specific details of the subject at hand in a speech. It might seem natural to think of some nouns as proxy variables for specific topics (e.g. the noun "Iraq" conveys a lot of topical information when used in an election speech). However, there is more to the picture than that because multiple different nouns can represent the same thing. Different words referring to the same thing can have very different connotations, and can therefore elicit different reactions from the public. It can be difficult to determine whether a speech's success is owed to the choice of topics and the opinions of the speaker, or to the choice of specific words.

This brings us to a limitation of this analysis, which is that this analysis does not uncover the reasons why specific words are correlated with winning or losing the election. In fact, the models developed in this paper are probably not very generalizable, partly due to the issues mentioned above, and partly due to the fact that the model was trained on data from a single country, and from a relatively short historical period.

On a related note, one should not draw any conclusions about cause and effect from this analysis. Simply because words are correlated with winning does not necessarily imply that a candidate will be more successful if they use more of these words. There could be an unknown underlying factor that makes people more likely to use certain words, and also makes them more likely to win elections. Nevertheless, the models developed in this analysis clearly suggest that words are powerful predictors of election results. Political candidate would certainly do well to examine the data before deciding what to say in their next speech.

6 Appendix

6.1 Definitions and Formulas

Word frequency/rate of word usage = the number of times the word was used in the election speech divided by the total number of words spoken in the speech

Deceptive words = words for which a change in usage rates is known to be associated with deception

Precision = True Positives / (True Positives + False Positives)

Recall = True Positives / (True Positives + False Negatives)

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

6.2 Figures

Words highly (linearly) correlated with election victory

Most positively correlated		Most negatively correlated	
Word with POS tag	Pearson's product-moment coefficient	Word with POS tag	Pearson's product-moment coefficient
because_CS	0.5158955463563536	he_PPS	-0.25312381140048396
that's_DT	0.4951799239392825	case_NN	-0.2580743279550563
we've_PPSS	0.4446639741799087	dollars_NNS	-0.25826518927803005
why_WRB	0.43313242917013134	life_NN	-0.262964994391641
folks_NNS	0.3912334460040334	most_QL	-0.26343586093680255
so_CS	0.381141065102543	would_MD	-0.26971907288008407
got_VBD	0.3556595962008113	increase_NN	-0.2746565287123468
that_CS	0.35164566815873954	to_IN	-0.2793240299735532
everybody_PN	0.3441453755691101	spending_NN	-0.2816088987834434
college_NN	0.33640631488493045	among_IN	-0.2892922655478836
decade_NN	0.33145580612201175	the_AT	-0.29441834747453133
what_WDT	0.32848936535195233	will_MD	-0.30035765137987436
laughter_NP	0.32795937764733113	of_IN	-0.3133051602951578
work_VB	0.3230486560265187	elected_VBN	-0.3335005256709765
roads_NNS	0.31460689718769413	government_NN	-0.3536341346725045
you_PPSS	0.3135612122283796	greater_JJR	-0.3547631588941269

Note: Candidate names are omitted from this list of words

6.3 Other Notes

For the POS analysis, the following table shows the tags for common POS. Not all tags are included in order to save space.

Category	Tags
Nouns	NN, NP, NS
Verbs	VB, VBD, VBZ, VBG
Adjectives	JJ, JJT, JJS, JJR
Conjunctions	CC, CS
Adverbs	RB
Pronouns	PP, PPO, PPS, PPSS, PN
Prepositions	IN
Adverbs	RB
Modal Verbs	MD
Numbers	CD

The data in 'deceptiondocword.csv' was not used because the values in it represented the number of occurrences of words rather than usage frequency, and because most of the important data in it was already contained in the main data table (but converted to a frequency, which is more useful). The list of deceptive words in 'deceptionword.csv' was still used in the analysis.