Effect of Language on the US Presidential Election: Analyzing Speeches for Semantic Patterns

Prepared By: Sanindie Silva, 20155558

CISC 251

1. Exploration Overview

Speeches are an integral part of a political campaign, where politicians can make promises and present a respectable persona to the greater public. Speeches play an immeasurable role in whether a candidate wins the US presidential election. This study attempts to measure how the language used in speeches correlates to a Presidential win.

This report aims to find:

1. A model with the highest prediction accuracy for the likelihood of winning the US presidential elections.
2. Words associated with winning speeches and any good strategies to win the elections.
3. What role does deceptive language have a role in winning the US elections?

While this report is not the most in-depth exploration of US political speeches, it properly outlines the process of finding the conclusions and appropriately addresses the procedures taken to build a process. For this exploration, KNIME and Microsoft Excel were used to format and develop models.

2. Data Preparation

The data provided was compiled into one CSV file most intuitively. The Speeches acted as the first column's values, the 1000 most frequent words because column headers and the provided matrix became the average frequency of the properties. The win/loss of each of the speeches was put into the last column and acted as the target label.

A similar process was repeated for the deception words, except KNIME was used to create normalized values for the deception words. Later into the project, I debated making a CSV file compiling all the values under the designated POS tags to compare the different types of words. Still, I decided to organize my data with column and row filters within KNIME instead.

3. Biases and Sources of Error

Biases can unintentionally influence certain behaviours, and the models and tools used to understand the raw data need to be accountable for these discrepancies. The raw data provided more information for winning candidates/speeches. At the beginning of this project's process, the top 1000 words' frequency was summed and separated by winners and losers, then it was repeated for the mean of those values. Two graphs were created. From the first graph (Figure 1), comparing the "Total Frequency of Losing and Winning Words," it seems like the winners hold a large overhead margin over the losers, and two defined lines are defining the winners and losers. While in the second graph (Figure 2), which compares the means instead of the sums, the losers say the more common words more often, and past the 20th word, the means muddle together. This early step acted as a reminder for future stages to reduce biases with more winning states' values. Luckily for KINME, the 1000 most frequent word features were already normalized. However, KNIME was used to normalize the deception word features to use it if needed in the future.

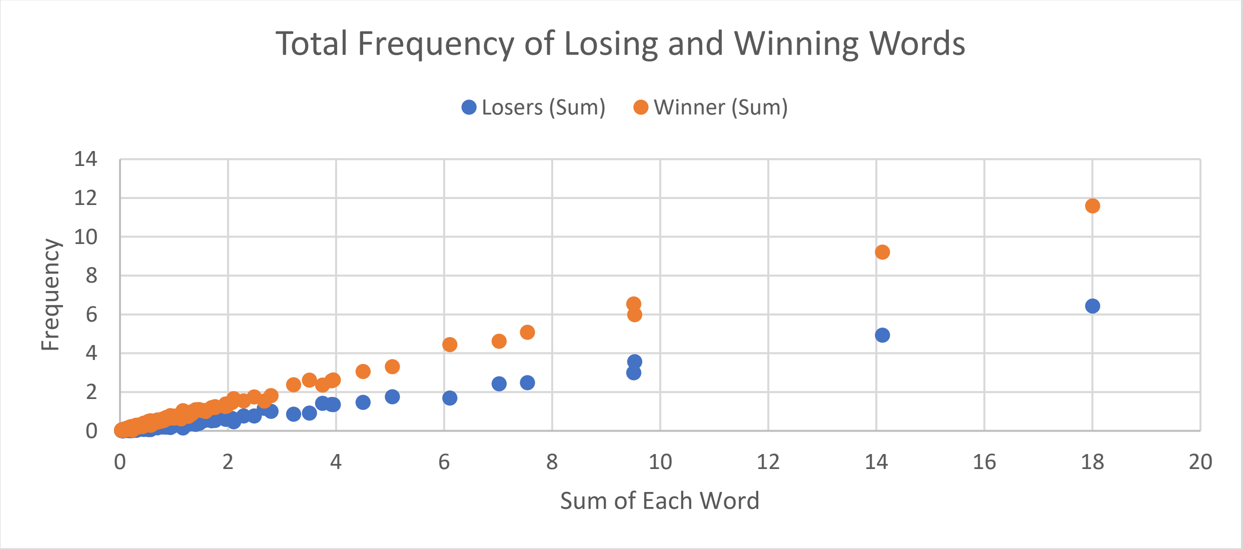


Figure 1: Total Frequency of Losing and Winning words; Scatter plot

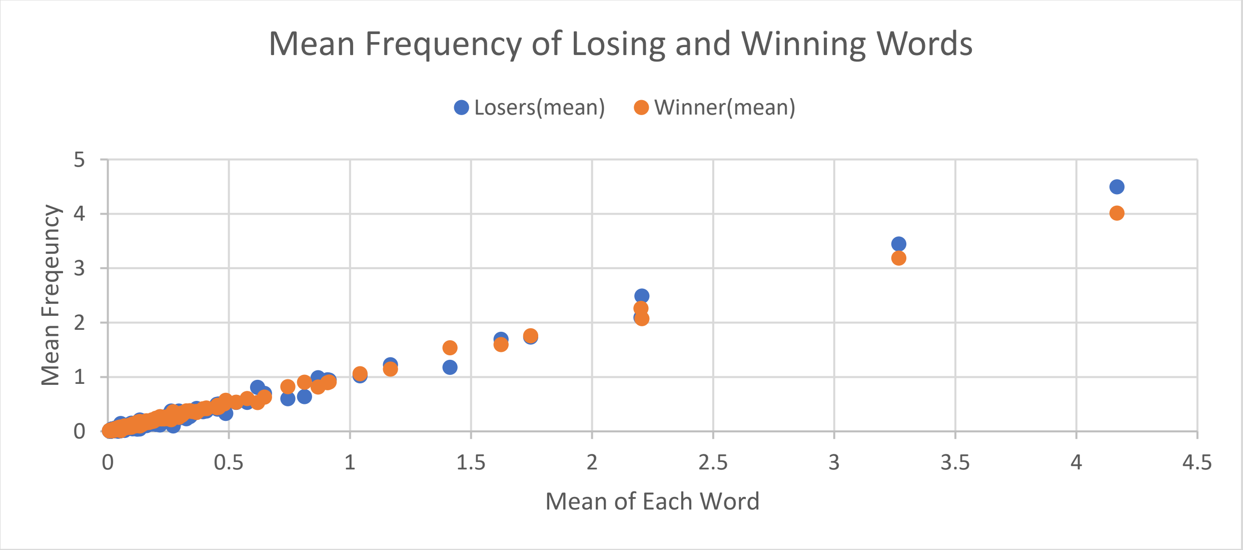
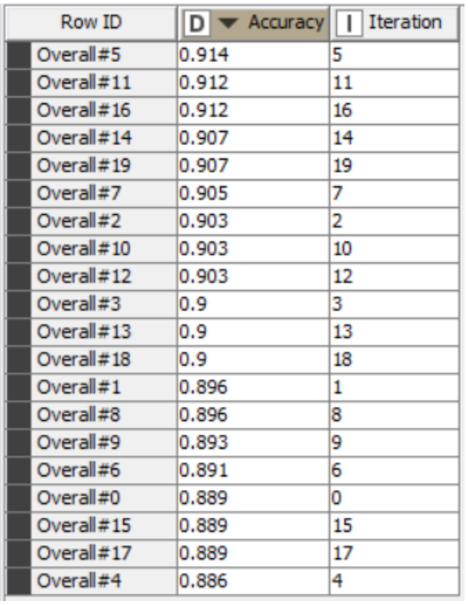


Figure 2: Mean Frequency of Losing and Winning Words; Scatter plot

While KNIME does a good job of remaining consistent through its models, and there isn't a lot of variance between the same models, it can create very different prediction accuracy when random sampling is used. There were two attempts to assist in try and moderate the effect of this randomness. At first, one random seed was consistently used in every model; that way, the same values would be used in each model's training and testing set. Unfortunately, good random sampling for a Random Forest can be atrocious for a Neural Network, and other factors using one random seed did not account for. As a result, the second attempt was implemented. To get the highest possible (somewhat random) accuracy, loops were used in KNIME to perform the same models' calculations multiple times, score the models, and record the results (Figure 4). The "Loop End" node collected the accuracy in a table (Figure 3), and the highest result was used to represent the model's capabilities in Section 6. Predictions.



Row ID	D Accuracy	I Iteration
Overall#5	0.914	5
Overall#11	0.912	11
Overall#16	0.912	16
Overall#14	0.907	14
Overall#19	0.907	19
Overall#7	0.905	7
Overall#2	0.903	2
Overall#10	0.903	10
Overall#12	0.903	12
Overall#3	0.9	3
Overall#13	0.9	13
Overall#18	0.9	18
Overall#1	0.896	1
Overall#8	0.896	8
Overall#9	0.893	9
Overall#6	0.891	6
Overall#0	0.889	0
Overall#15	0.889	15
Overall#17	0.889	17
Overall#4	0.886	4

Figure 4: Accuracy Table from KNIME Loop End node

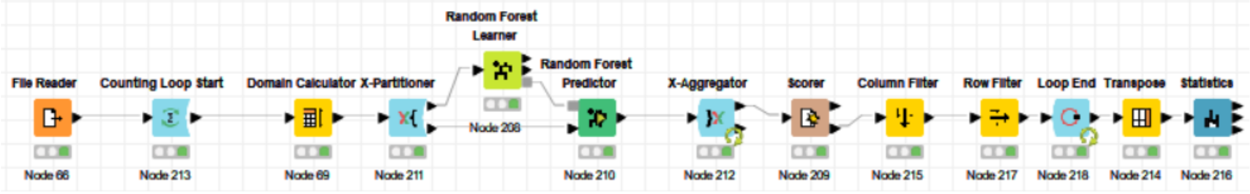
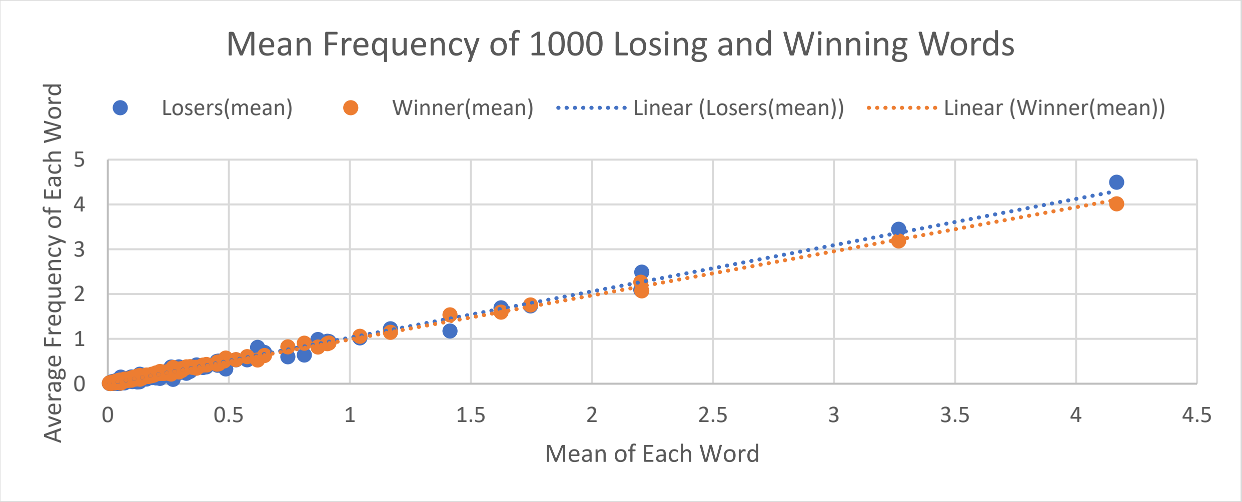
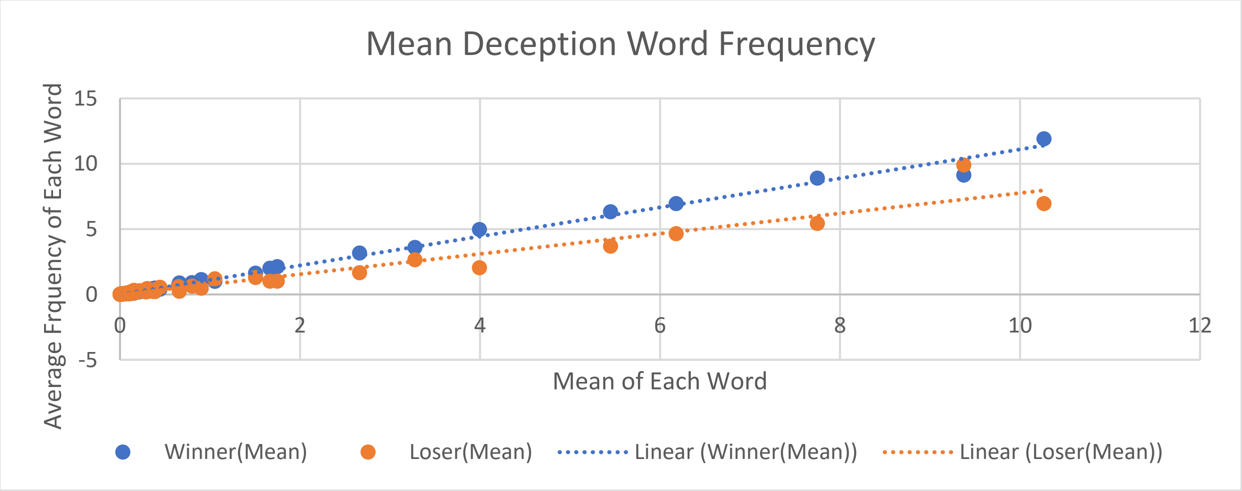


Figure 3: Sample KNIME loop workflow, using random forests

4. Simple Statistics

The following two graphs visualize the difference of the means of the winners and losers for the 1000 word data set and the deception word data set. All other basic statistics found were not useful enough to include, except for the 1000 words' variance, which was: 0.0000009048.

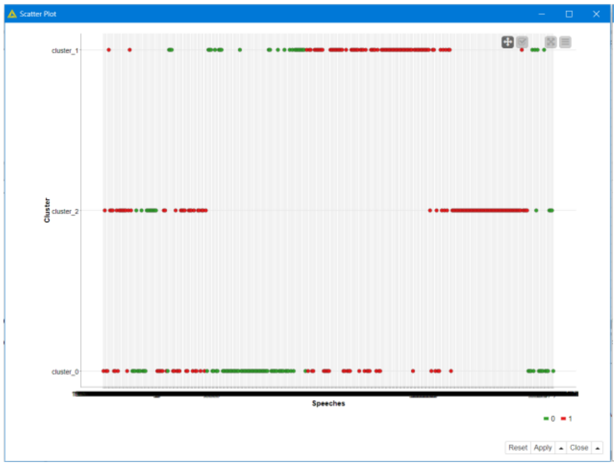


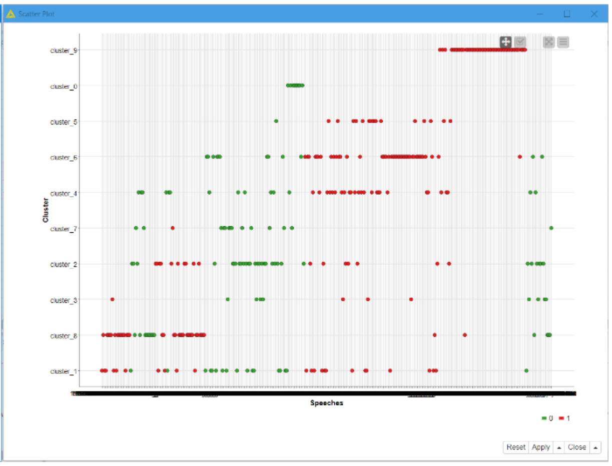


5. Clustering

To better understand the data and determine how hard predictions from this data set were going to be, three different types of clustering were used: K-means, DBSCAN and PCA.

K-means quickly identified that this data set would be difficult to make accurate predictions from. I chose to separate the data into 3(left) and 10(right) clusters to see the difference in a scatter plot (green dots are losses, and red dots are wins). There was no recognizable divide until the 8th and 9th clusters of the 10-means model, where there were two small clusters separated by wins and losses. This indicates that larger, more complex models would probably be more helpful when creating predictions.





DBSCAN clustering was used to see if there were cluster closer together than others were, and hopefully use it to group the features with the most variance and impact on the data set. Only one cluster was found when it was attempted regardless of how low or high the values were (Figure 7). This is probably because many of the points are too close to each other, so it is hard to cluster them into different groups, which is another clue that making predictions from these values would be difficult.

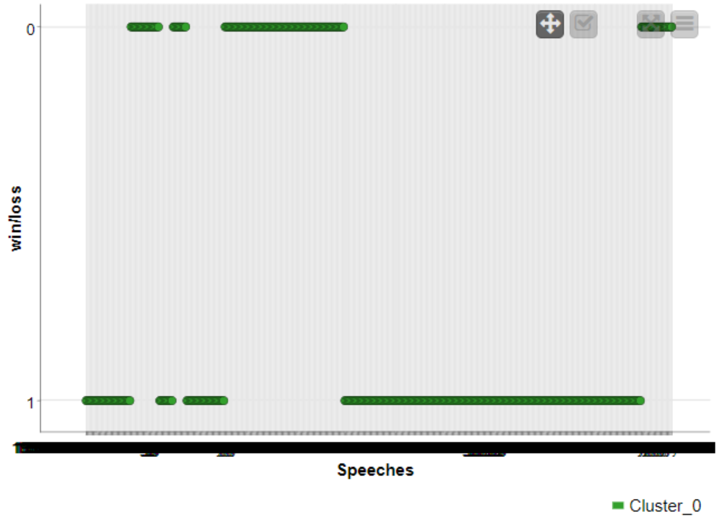


Figure 6: DBSCAN on 1000 most frequent words scatter plot

Since the K-means and DBSCAN weren't the most useful for understanding the data set, additional research was used to find PCA (or Principal Component Analysis). PCA reduces the dimensionality of large data sets by transforming many variables into a smaller group. This way, the data becomes simpler to visualize and understand without losing a lot of information. It mainly works for data that is strongly correlated. PCA proved to be the most useful clustering method in terms of visualization (Figure 8). A "Speeches vs PCA dimension 1" scatter plot was chosen for this report to point out how, while there are no well-defined clusters. There are 7 "shapes" in the data, somewhat separated by colour and trend direction (rough, manual divisions are included in Figure 8 to show thought process). The PCA can also point out a few outliers that don't follow the general direction of the data; while most speeches follow similar patterns, some speeches can vary significantly and influence the data set's pull.

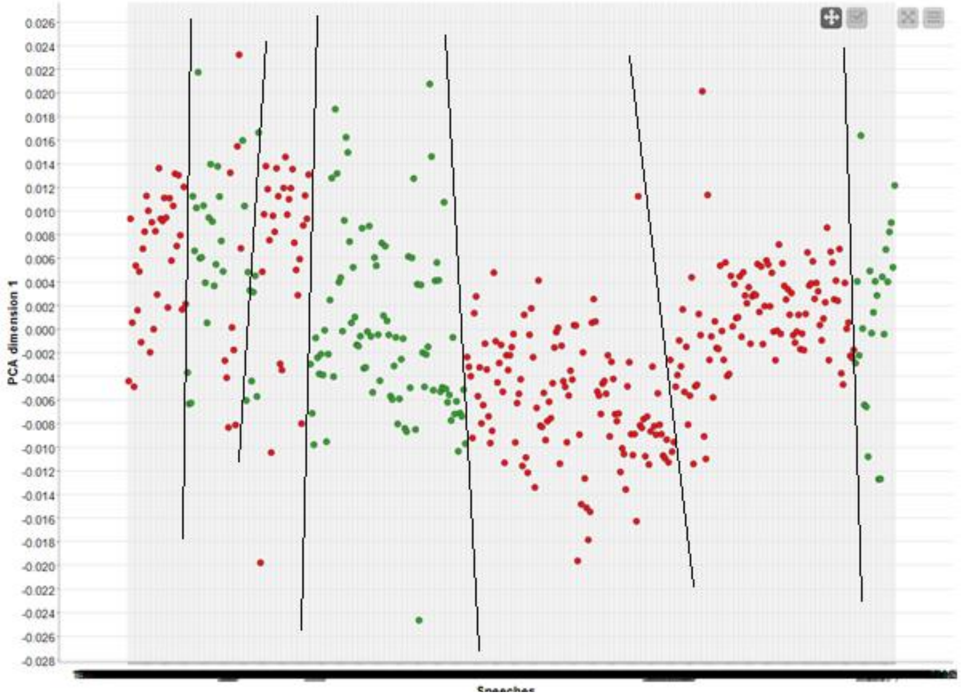


Figure 5: PCA Scatter Plot: Speeches vs PCA dimension 1

Once clustering has helped gain a better understanding of the data, predictions can be made with these discoveries in mind.

6. Predictions

The primary process to find what influences a presidential win based on language is creating prediction models, finding the most influential data, and exploring the most significant influences.

The models used were: (A) Decision Trees, (B) Random Forests, (C) Neural Networks, (D) SVM, (E) Boosted Decision Trees. There is a summary table (Table 1) highlighting the original prediction accuracy and values used, compared to the highest prediction accuracy found. Boosted Decision Tree's resulted in the highest prediction accuracy of 96.6%. Future explorations of this data set will use Random Forests and Boosted Decision trees to understand the data better. Following Table 1, there are the reasons why each model was performed.

Table 1: Summary table of Prediction Models

Model	Original Prediction Accuracy & Values used in Calculations	Highest Prediction Accuracy & Values used in Calculations
Decision Trees	81.609% Quality Measure: Gini Index Pruning: MDL Threads = 8	81.609% (No Change)
Random Forests	83.908% Split Criterion: Information Gain Ratio Tree depth = 10 Models = 100	94.253% Split Criterion: Gini Index Tree depth = 10 Models = 100
Neural Networks (RProp MLP Learner)	92.111% Iterations = 100 Hidden Layers = 1 Hidden neurons per layer = 10	92.343% Iterations = 100 Hidden Layers = 4 Hidden neurons per layer = 13
SVM	59% Overlapping penalty = 1 Polynomial kernel: Power = 1	95.402% Overlapping penalty = 101.05 RBF kernel: sigma = 0.7.
Boosted Decision Trees	96.552% Tree Depth = 4 Models = 100 Learning rate = 0.1	96.6% Tree Depth = 4 Models = 100 Learning rate = 1

*note the above table does not include the average of each

A. Decision Trees

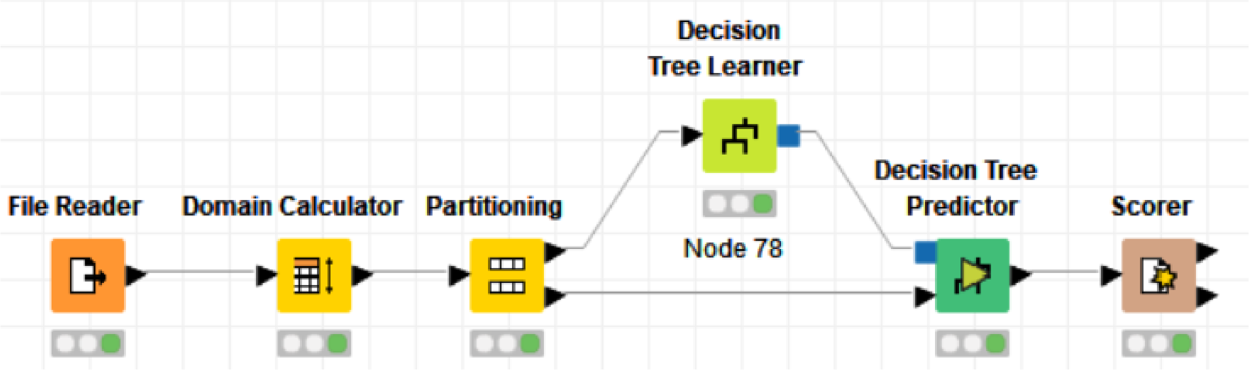


Figure 7: Decision Tree KNIME Workflow

Decision trees are easy to understand, and each branch and leaf's reasoning is clearly indicated in the nodes. They are a good stepping stones for bigger data sets to get a rough estimate of what a good base accuracy is and allow other models to build off. It is the foundation for creating more cohesive models. Decision trees are beneficial for comparing with random forests and seeing where they are similar and different.

B. Random Forests

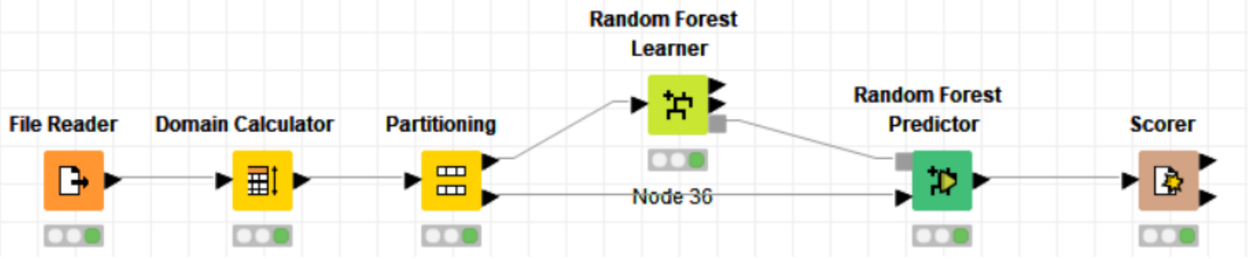


Figure 8: Random Forests KNIME Workflow

After decision trees, random forests are an excellent model to be compared to. They can work with many attributes, many classes (win vs loss) and have low variance and bias in their modelling. With the many iterations and their foundation in decision trees, there's no need for cross-validation, yet it still can produce high accuracy. Random forests can be formatted, so they don't overfit the data. It's advantageous to do attribute selection because it's possible to determine why the produced model is built that way due to the decision tree branches. It is an excellent tool further to explore the data set and find the most impactful words.

C. Neural Networks RPROP MLP)

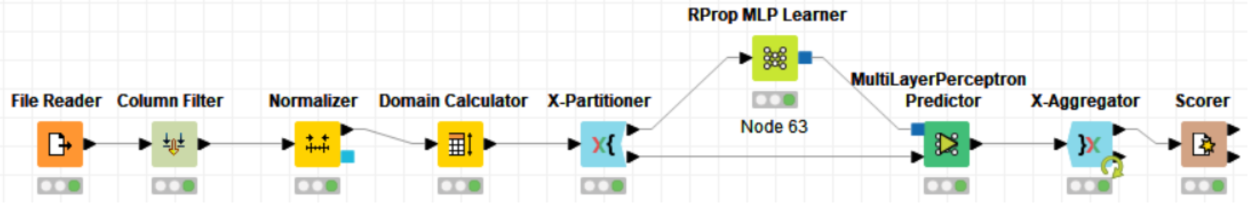


Figure 9: Neural Network KNIME Workflow

Neural networks are a good step up in complexity, especially if a given data set allows you to work with many attributes, with subtle interactions between them. As seen from the DBSCAN clustering done earlier, the features' values are relatively small and are clustered together. This indicates that they would be dependent on one another. Neural networks are also suitable for non-linear functions despite the messy dependencies like the data provided has.

D. SVM

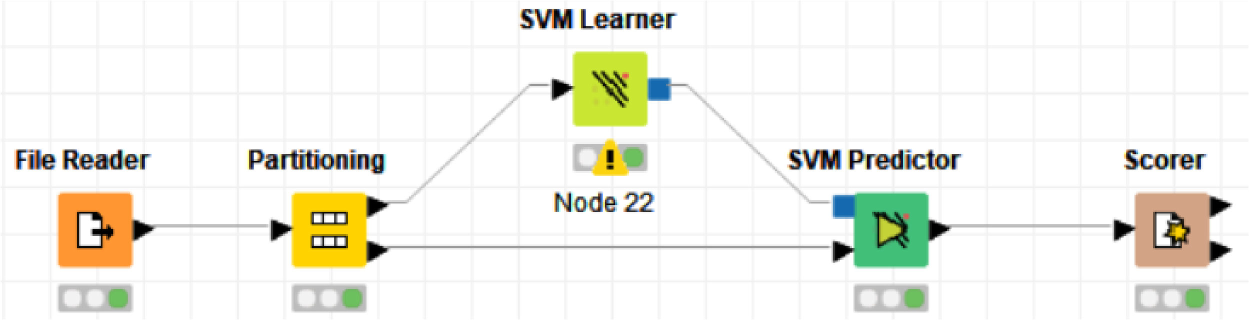


Figure 10: SVM KNIME Workflow

SVM's are another appropriate model for two classes (win/loss) and data that is not easily linearly separable. The PCA clustering showed that the data could overlap (in the clusters), so it would be appropriate to account for that. SVM can also separate data in higher dimensions than just linear, and it is easily scalable for larger data sets.

E. Boosted Decision Trees

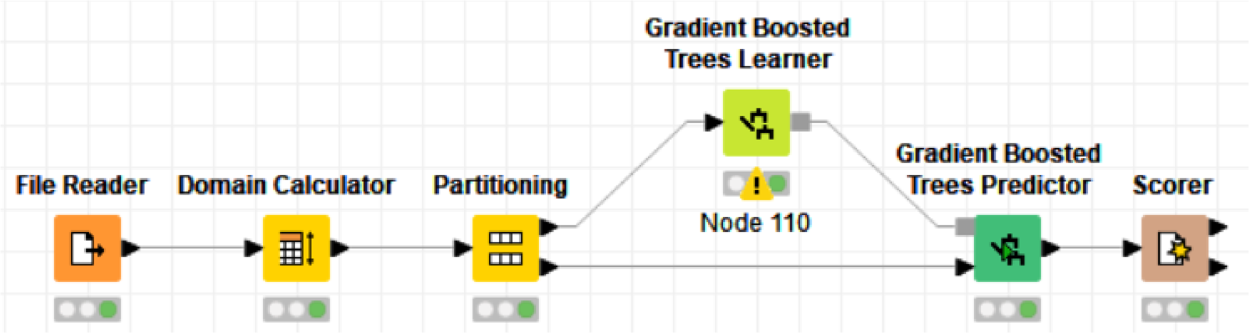

Gradient Boosted Trees Learner
File Reader
Domain Calculator
Partitioning
Node 110
Gradient Boosted Trees Predictor
Scorer

Figure 11: Boosted Decision Trees KNIME Workflow

Finally, boosted decision trees were chosen since they result in higher accuracy for decision trees and accounts for the variance by cancelling it. During the procedure, random forests had high accuracy. One way to increase it was to use boosting since boosting provided valuable information about which objects are the hardest to classify.

Conclusions:

The highest prediction accuracy found was Boosted Decision Tree's with 96.6%; however, its average accuracy was 89.68% - a statistic that will be useful to judge attribute selection. Random forests can be used for attribute selection since the reasoning is provided in the KNIME node. Future explorations of this data set will use Random Forests and Boosted Decision trees to understand the data better and find important patterns and words.

7. Attribute Selection and Improved Data Sets

To find which words have a more substantial predictive power, I used the feature selection nodes on a Random Forest model to sort through the words that ended at the top of the multiple decision tree models created by the Random forest learner. Each improved data set was created KNIME using column and row filter nodes. The process of making the new, improved data sets are included below.

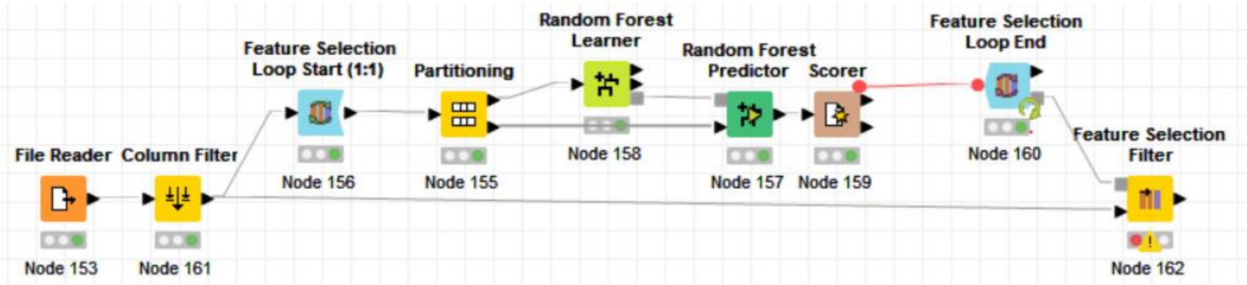


Figure 12: Random Forests Feature Selection Workflow

I found that the words most often at the top of the highest percentages were “the” or “at,” but I felt that wasn’t an appropriate measure that helps you with the election. Using the Feature Selection Filter, I noticed that many attributes near the top were nouns. I wanted to explore this discovery, so I used the column filter to send different POS tags into the feature selection loop and see their accuracies. The following table shows their accuracies:

Table 2: Comparing POS tag accuracies

Base Accuracy with all the POS tags:					89.655%
POS Tags	Nouns (“NIL", "NN", "NN$", "NNS", "NP", "NP$", "NPS", "NR")	Verbs ("VB", "VBD", "VBG", "VBN", "VBZ")	Adjectives ("JJ", "JJR", "JJS", "JJT")	Questions ("WDT.", "WP$", "WPS", "WQL", "WRB")	Numbers ("CD, CS")
Percentages	90.8 %	85.4 %	83.6 %	87.3 %	75 %
Noun		94.2 %	88.5 %	88.5%	89.7 %
Noun + Verbs			88.5 %	90.8%	89.7%
Nouns + Verbs + Question			88.5%		89.7%
Nouns + Verbs + Questions +Adjectives			89.7%		

At first, I chose the POS tags that included words of “content,” meaning words that aren’t used simply for grammar (at, the, etc.). And I chose the POS tag that had the highest accuracy at each level and compared that to a group of POS tags with the previous highest level. For instance, since nouns were the highest accuracy in the first row (90.8%), I took nouns and compared it to the other pos tags, and found that nouns and verbs had the highest group accuracy (94.2%), then I repeated the same process for the largest accuracy for each row. The highlighted green boxes are the highest accuracy for each row, and the darkest green is the highest accuracy for the table. This information tells me nouns and verbs hold the most predictive power over the data set – which is a given since most speeches are content heavy. However, this also shows that posing (and answering) questions and talking about numbers are a valuable factor to include in speeches.

Another feature selection workflow can be used to find more specifically which words have more of a predictive power in this data set. This time, a "Low Variance Filter" is added to exclude values that don't have much impact on the data set. This process is repeated twice, once for only nouns and verbs, then a second time for all the data.

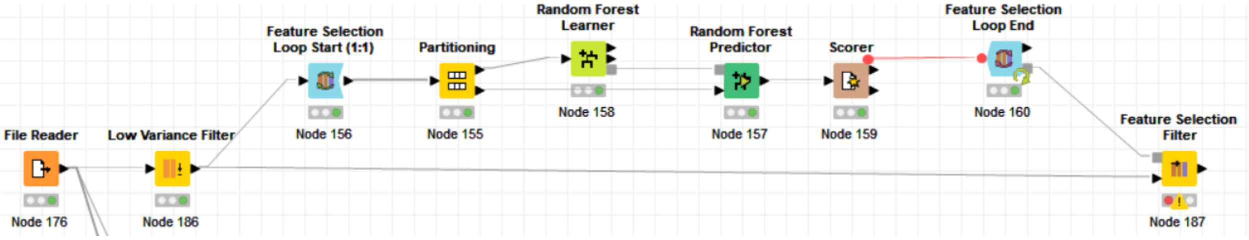


Figure 13: Low Variance and Feature Selection KNIME Workflow

Then, from the Feature Selection Filter node, a model with high accuracy yet a low number of features is chosen. The features are extracted and compared to the other models' features in the node for similarities and the most popular trends. Afterwards, the data set is sent to a Boosted decision tree

Low Variance and Feature Selection for Nouns and Verb POS Tags

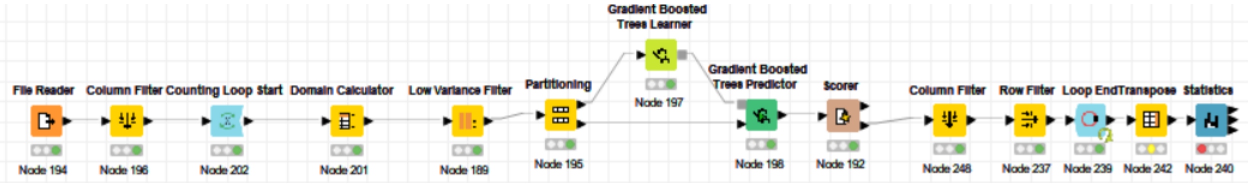
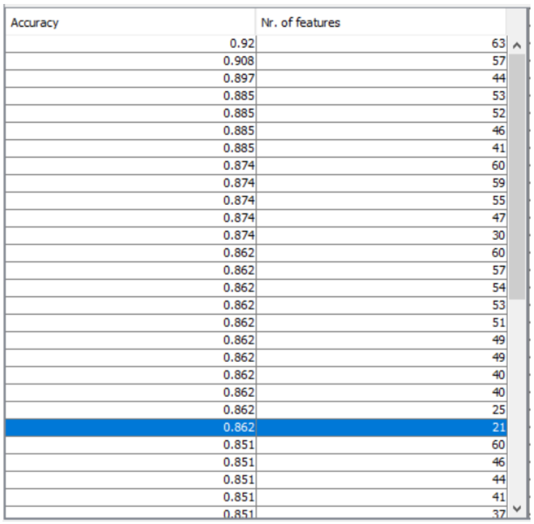


Figure 14: Low Variance and Feature Selection KNIME Workflow

Feature Selection Filter Data: The model chosen to explore was a model with 86.2% accuracy and 21 features included. (Figure 15)

Features: America, want, time, need, tax, got, states, families, say, energy, security, senator, workers, change, businesses, campaign, women, street, school, McCain,'

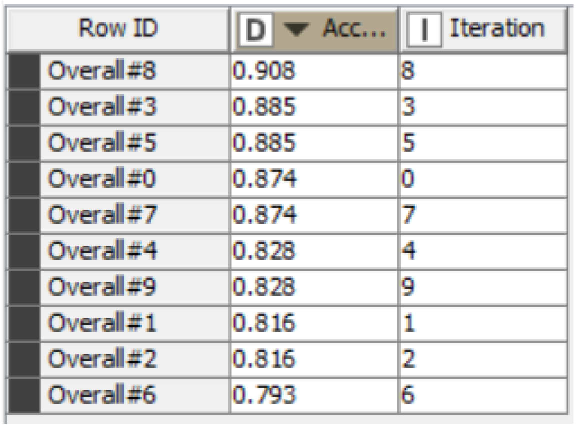
Boosted decision tree average accuracy: 85.07% (Figure 16)



Accuracy	Nr. of features
0.92	63
0.908	57
0.897	44
0.885	53
0.885	52
0.885	46
0.885	41
0.874	60
0.874	59
0.874	55
0.874	47
0.874	30
0.862	60
0.862	57
0.862	54
0.862	53
0.862	51
0.862	49
0.862	49
0.862	40
0.862	40
0.862	25
0.862	21
0.851	60
0.851	46
0.851	44
0.851	41
0.851	37

Figure 15: Accuracy Statistics for Features Selection

Conclusions: While nouns and verbs are the best POS tags to use, and it's important to have content included in speeches, from this data, it isn't the best indication of whether or not a candidate is going to win the election, but instead outlines the type of words to say in the content. The words included above (except tax, McCain and') lean towards the positive spectrum of speech. These are all "positive language" potentially to make promises about what candidates intend to do. When compared to other models, they also include high amounts of positive languages. As a result, presidential candidates should lean towards using more positive language, make promises and portraying themselves as a proactive force.



Row ID	D Acc...	I Iteration
Overall#8	0.908	8
Overall#3	0.885	3
Overall#5	0.885	5
Overall#0	0.874	0
Overall#7	0.874	7
Overall#4	0.828	4
Overall#9	0.828	9
Overall#1	0.816	1
Overall#2	0.816	2
Overall#6	0.793	6

Figure 16: Accuracy Statistics for Boosted Decision Trees

Low Variance and Feature Selection for All Words

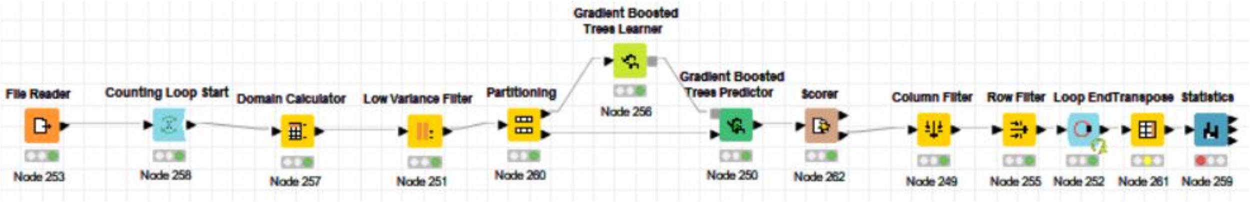
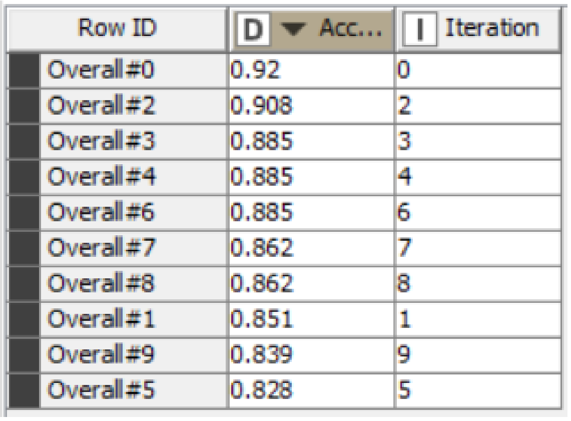


Figure 18: Low Variance and Feature Selection KNIME Workflow

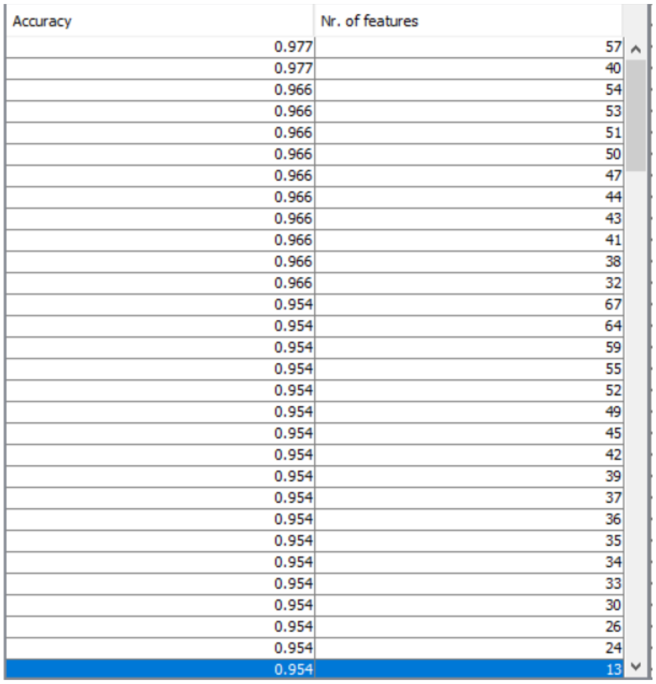
Feature Selection Filter Data: The model chosen to explore was a 95.4% accuracy model and 13 features included. (Figure 17)

Features: for, you, they, it, America, has, country, I'm, Americans, senator, street, Obama, public

Boosted decision tree average accuracy: 87.27% (Figure 19)



Row ID	D Acc...	I Iteration
Overall#0	0.92	0
Overall#2	0.908	2
Overall#3	0.885	3
Overall#4	0.885	4
Overall#6	0.885	6
Overall#7	0.862	7
Overall#8	0.862	8
Overall#1	0.851	1
Overall#9	0.839	9
Overall#5	0.828	5



Accuracy	Nr. of features
0.977	57
0.977	40
0.966	54
0.966	53
0.966	51
0.966	50
0.966	47
0.966	44
0.966	43
0.966	41
0.966	38
0.966	32
0.954	67
0.954	64
0.954	59
0.954	55
0.954	52
0.954	49
0.954	45
0.954	42
0.954	39
0.954	37
0.954	36
0.954	35
0.954	34
0.954	33
0.954	30
0.954	26
0.954	24
0.954	13

Figure 17: Accuracy Statistics for Features Selection

Figure 19: Accuracy Statistics for Boosted Decision Trees

Conclusions: The general feature selection model was more accurate than the Noun and Verb specific one and used fewer features too, which is the ideal model. However, these words aren't quite as useful. It supports the previous statement to use more positive language, but these words seem more ambiguous. Regardless, candidates should use positive language as much as possible. When these words are considered, there is a lot of reference to "America" and the "people," showing a united front and a sense of togetherness. When checking other models, there are a substantial amount of words that support this theme. This is an indication to include words that promote unity in presidential speeches.

8. Deception Words Effects

Deception words are words that make the candidate seem more positive than they actually are. It is a theory that more deceptive candidates do better in elections. This section aims to support that claim.

The first thought was to try to create predictions with the deception words only. Unfortunately, that wasn't successful, and there were very small accuracies in each of the models, regardless of the changes in parameters. Next, a dataset of the 1000 words missing the deception words was attempted to be created. However, only a handful of the words were in the 1000 word dataset, and when it was used in prediction models, it was comparable (with slightly less accuracy) to the 1000 word models. Both traditional methods of predictions were dead ends.

As a result, data exploration started from the ground up, and a scatterplot on excel was made comparing the means of each of the deception words for the winners and the losers (Figure 20). When plotted on the same graph, it's recognizable that the winners use deception words much more often than losers (as seen with the trend lines).

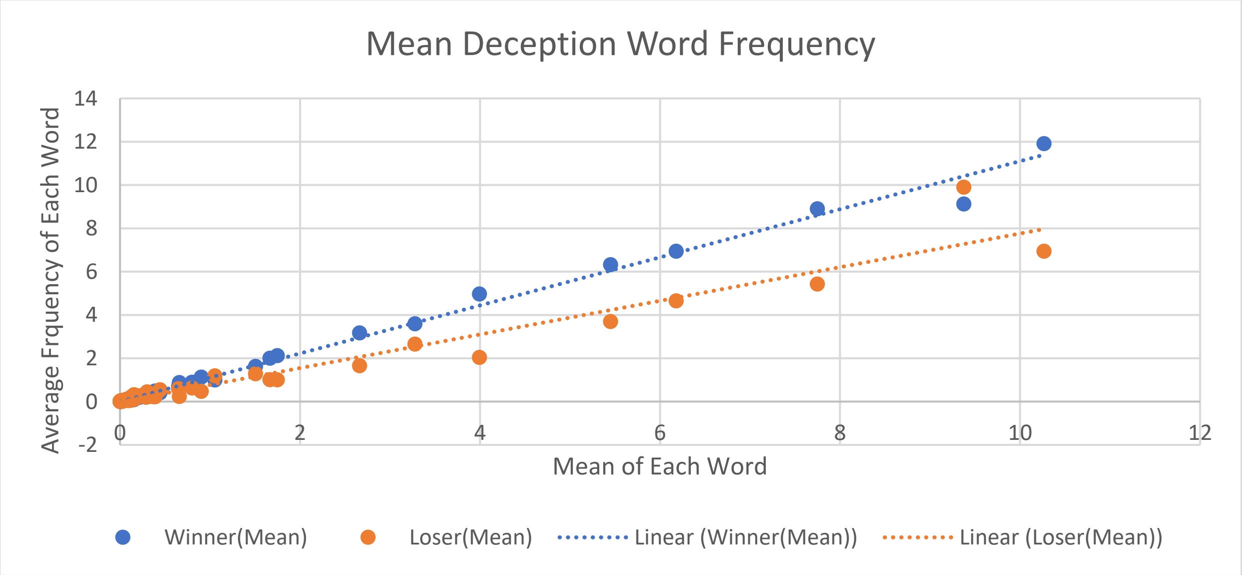


Figure 20: Mean Deception Word Frequency

Furthermore, more research was used to find a linear correlation, a method of visualization that compares one variable to another to see if they are associated with each other. A matrix of the deception words was created that compared the frequency of each word's appearance. The darker the square, the more often it's used. Two linear correlations were made, one for the winners and one for the losers. The more colourful the matrix means that the words are used more often.

When comparing the winner versus the loser's matrix, it's noticeable that the winner's matrix is more colourful than the loser's (Figure 21). This indicates that the winners use more deceptive words on average.

This supports the claim that more deceptive candidates are more likely to win elect

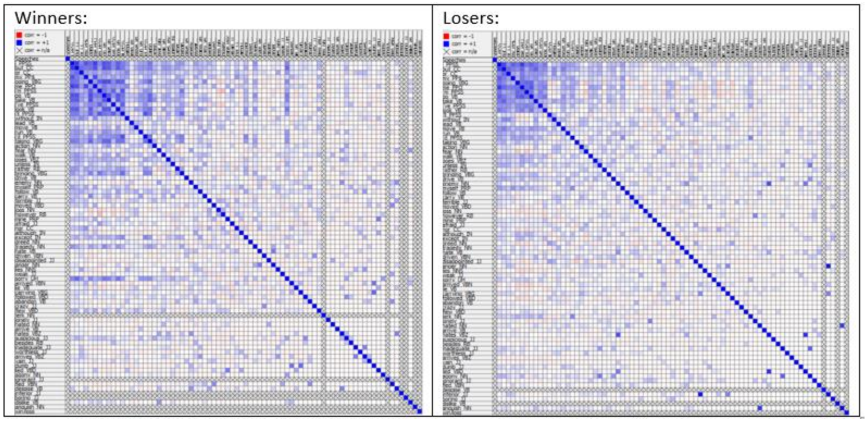


Figure 21: Winner vs Loser Linear Correlation

9. Final Predictions and Conclusion

Predictions came from a few different areas of this exploration. Comparing POS tags to each other, we find that candidates should use more positive language to make promises and present themselves as a good person willing to make changes. To extrapolate, candidates should avoid using negative language to detract from the positive goal they are trying to present. The original 1000 word data set with low variance found that candidates should enforce unity within their audience, referencing the American people and their sense of nationalism. From the analysis of the deception words, candidates should present themselves as better than they are. With all these consideration factors, presidential candidates should use as much positive language as possible, remembering to be deceptive and paint themselves in a positive light. They should use language that promotes unity within the American people and appeals to them. These strategies align with general persuasive strategies and popular trends in American politics. They hint that Americans care more about a candidate's persona, personality and promises rather than their policies and planned actions. It creates a conflicted narrative on the presidential campaign, where it appeals to the mass interest is vague positive reassurances rather than clear plans. The strategies found in this exploration imply that an American political campaign lacks focus on the important things and finds solace in agreeable unproven promises.