# Political Ideology Detection Using Text Classification

Dong Wook (Daniel) Chung

## Abstract

*In this paper we seek to measure politicians' ideologies by using text classification to measure how Democratic or Republican their speeches are. Also, based on the median voter theorem, we hypothesize that candidates move towards the center for the general election, in order to win more votes from the voters on the opposite side of the political spectrum as well. We use the speeches from presidential elections to test the hypothesis.*

## 1. Introduction

In order to vote for a right candidate and give support, it is important to know each politician's ideology correctly. Most of the time, the public receives the information about politicians through the media, but the media are often biased and the information can be easily manipulated by them. For example, Fox News Channel is often accused of favoring the conservatives over the liberals. A subtle difference in a word choice by the media can also affect the tone through which the information is delivered. Related to the tax on the transfer of the estate of a deceased person is the matter of whether the tax should be called the "estate tax" or the "death tax". Opponents of the estate tax like to call it the "death tax", in order to give a negative connotation to the term. Similarly, opponents of abortion like to call their movement "pro-life" instead of "anti-abortion", in order to give the impression that the alternative movement is "pro-death" or "anti-life". These subtle differences by the media can alter the meanings of what politicians say. So just by relying on the information delivered by the media, it is difficult to correctly understand the real values that politicians believe in.

In this paper, we use the speeches from politicians to understand their ideologies. We first train a text classifier with a dataset of political speeches. We test naive bayes and linear support vector machine (SVM) to find the classifier and the set of features that classify political speeches the best. To understand a politician's ideology, we then use the classifier to classify each of the politician's speech into Democratic or Republican and see the ratio of the speeches classified as Democratic and Republican.

Also, using the speeches made by Obama, McCain and Romney in the 2008 and 2012 presidential elections and the text classifier, we test the hypothesis that a presidential candidate moves his political ideology towards the center for the general election in order to win votes from the opposite side of the political spectrum as well. The median voter theorem states that "a majority rule voting system will select the outcome most preferred by the median voter." [1]. To win a primary election, a candidate competes against other candidates in the same party and has to win the votes from the voters, whose political spectrums are concentrated and are close to the party's ideology. On the other hand, to win a general election, a candidate has to win the votes from the whole US population, whose political spectrum is much widely spread out. The median political ideology shifts towards the center for the general election, and we expect the candidates to shift their ideologies towards the center as well. Finally, we extract the keywords that the three candidates used and compare their differences between the primary and the general election.

# 2. Related Work

## 2.1. Political Ideology Detection

Riabinin [2] applied support vector machine algorithm on the 36th Canadian Parliament debates and achieved an accuracy as high as 98%. He also tested and found out that using bag-of-words model, which is the simplest model for text classification, produces a satisfactory classification rate.

Iyyer et al. [3] applied a recursive neural network to identify a political ideology from a text. They focused on detecting ideologies on the sentence level. They selected sentences with political bias from the 2005 US Congress debate transcripts and ideological books and used crowdsourcing to label each sentence as liberal or conservative. They achieved 70.2% classification rate by incorporating the RNN framework, adding phrase-level data, and initializing with word2vec.

Thomas et al. [4] tested whether one can determine whether the given speech supports or opposes to the given proposed legislation, using the same US Congress debate transcripts as the dataset. They used the fact that the speeches in the data set occurred as part of a discussion for the given legislation and exploited the relationships between speech segments. It was found that using this information improves the classification rate significantly compared to classifying each speech in isolation.

### 2.2. Testing the Median Voter Theorem

Sim et al. [5] originally proposed the hypothesis that presidential candidates move towards the center for the general election. To test the hypothesis, they trained their domain-informed bayesian hidden markov model using political books and magazines, and they used the speeches from Obama, McCain and Romney to test their hypothesis. They obtained the result that the candidates indeed drew more from the opposite political side in the general election than in the primary election.

## 3. Approach

Sim et al. [5] used books and magazines to train their model, but we believe the words and tone used in books and magazines are fundamentally different from those used by the politicians in their speeches. Therefore, in this paper we use speeches from politicians to train the model. We test naive bayes and linear SVM classifiers with various features to find the model with the highest classification rate. Both naive bayes and linear SVM classifiers are popular models used for text classification.

Naive bayes classifiers are based on the Bayes' theorem and assume that each feature is independent of all other features in the same document. For example, if we only consider the unigrams in the speeches, when using naive bayes classifiers, we assume that each word is independent of all other words in the same speech. In other words, if we let $y$ denote a class variable, which is either Democratic or Republican in this case, and $x_i$ denote each feature, we assume that

$$P(x_i|y,x_1,...,x_{i-1},x_{i+1},...,x_n) = P(x_i|y)$$

Then we use the Bayes' theorem and the assumption to find the value of $y$ such that

$$\hat{y} = \underset{y}{argmax} P(y|x_1,...,x_n) = \underset{y}{argmax} \frac{P(y)P(x_1,...,x_n|y)}{P(x_1,...,x_n)} = \underset{y}{argmax} \prod_{i=1}^{n} P(y)P(x_i|y)$$

Although the assumption that each word is independent of all other words in the same speech is not accurate, naive bayes classifiers are known to work fast and produce a good classification rate.

Support vector machines (SVMs) are widely used for classification, regression and outliers detection [6]. Linear SVMs find the maximum-margin hyperplane that best separates the data into two categories.

For each of the politicians who ran the 2008 and/or 2012 presidential elections, using the best model that we find, we identify their political ideology by testing how many of their presidential election speeches are classified as Democratic or Republican.

To test the hypothesis, we look at not only the classification results but also the classification probabilities for the speeches from Obama, McCain and Romney from their presidential elections. We use the distribution of the classification probabilities of the speeches to see whether there is a shift in ideology between the primary election and the general election for each candidate. In this paper, we assume that the "center" of the political spectrum is the midpoint, where a speech is equally likely to be classified as Democratic and Republican. Finally, we extract the keywords that the three candidates used by calculating term frequency-inverse document frequency and look at the differences between the primary and the general election.

## 4. Data

### 4.1. Training and Testing Text Classifier

To train and test our model, we use the transcripts from the 2005 US Congress, which are the same data that were used by Iyyer et al. [3] and Thomas et al. [4], retrieved from http://www.cs.cornell.edu/home/llee/data/convote.html [7]. The dataset consists of 8,121 speeches, 4,046 and 4,075 of which are from the Democratic Party and the Republican Party respectively. The website also provides a dataset where some data are filtered out. Speeches simply asking to yield the time (e.g. "Mr.speaker, i yield back the balance of my time.") are filtered. Also, speeches that contain the term "amendment" are filtered, as a speaker's opinion on an amendment can be different from his general political ideology. The filtered dataset contains 3,857speeches in total, 1,868 and 1,989 of which are from the Democratic Party and the Republican Party respectively. We use the filtered dataset in this paper.

### 4.2. Political Ideology Detection

Using the text classifier that we train and test, we identify the ideologies of the politicians who ran for the 2008 and/or 2012 presidential elections using their presidential election speeches and a few of their key speeches made before the two primary elections. The dataset consists of 809 speeches made by 18 candidates. We discard the speeches from the candidates who made less than 10 speeches (Joe Biden, Michele Bachmann, Herman Cain, Jon Huntsman, Ron Paul, and Tim Pawlenty), as there are too little data. This leaves us with 780 speeches from 12 candidates. Table 1 shows a breakdown of the candidates and speeches in the dataset. The top four candidates are from the Democratic Party, and the rest of the candidates are from the Republican Party.

| Candidate | Primary | General | Before Primary | Total |
|---|---|---|---|---|
| Hillary Clinton | 105 | 0 | 1 | 106 |
| John Edwards | 27 | 0 | 0 | 27 |
| Barack Obama | 78 | 81 (2008), 99 (2012) | 14 | 272 |
| Bill Richardson | 29 | 0 | 2 | 31 |
| Newt Gingrich | 15 | 0 | 0 | 15 |
| Rudy Giuliani | 36 | 0 | 1 | 37 |
| Mike Huckabee | 14 | 0 | 1 | 15 |
| John McCain | 44 | 128 | 2 | 174 |
| Rick Perry | 10 | 0 | 0 | 10 |
| Mitt Romney | 8 (2008), 20 (2012) | 19 | 14 | 61 |
| Rick Santorum | 16 | 0 | 0 | 16 |
| Fred Thompson | 15 | 0 | 1 | 16 |

Table 1: Breakdown of the 2008 & 2012 presidential election speeches dataset

## 5. Building a Text Classifier Model

We use Python's scikit-learn to build our model. Using Scikit's list of stop words, we ignore the words that occur frequently, such as articles and prepositions. We use tf-idf (term frequency-inverse document frequency) matrices instead of simple word-count matrices, since tf-idf shows better how important each word is to the given speech relative to the whole corpus. We apply 10-fold cross-validation to test our model and use F1 score to compare the performances. Naive bayes and lienar SVM classifiers are tested with various combinations of unigram, bigram, trigram and quadgram. Table 2 and Table 3 show the results for the naive bayes classifiers and the linear SVM classifiers respectively, using the 2005 US Congress debate transcripts.

| Features | Democratic | Republican | Total |
|---|---|---|---|
| Unigram | 0.6901 | 0.7429 | 0.7165 |
| Bigram | 0.6799 | 0.7680 | 0.7240 |
| Trigram | 0.6031 | 0.7371 | 0.6701 |
| Quadgram | 0.5215 | 0.7178 | 0.6196 |
| Unigram+Bigram | 0.6942 | 0.7586 | 0.7264 |
| Unigram+Bigram+Trigram | **0.7036** | **0.7657** | **0.7346** |
| Unigram+Bigram+Trigram+Quadgram | 0.6945 | 0.7640 | 0.7293 |
| Bigram+Trigram+Quadgram | 0.6592 | 0.7642 | 0.7117 |

Table 2: F1 scores for naive bayes classifiers

| Features | Democratic | Republican | Total |
|---|---|---|---|
| Unigram | 0.7244 | 0.7499 | 0.7371 |
| Bigram | 0.7271 | 0.7555 | 0.7413 |
| Trigram | 0.6903 | 0.7137 | 0.7020 |
| Quadgram | 0.6844 | 0.6647 | 0.6746 |
| Unigram+Bigram | 0.7380 | 0.7613 | 0.7497 |
| Unigram+Bigram+Trigram | **0.7402** | 0.7725 | **0.7564** |
| Unigram+Bigram+Trigram+Quadgram | 0.7384 | **0.7734** | 0.7559 |
| Bigram+Trigram+Quadgram | 0.7287 | 0.7538 | 0.7412 |

Table 3: F1 scores for linear SVM classifiers

Among the naive bayes classifiers that we test, the classifier with unigram, bigram and trigram shows the best performance. On the other hand, F1 scores for classifiers with trigrams or quadgrams are significantly lower than the others. We believe this is because the number of occurrences of trigrams and quadgrams is lower than that of unigrams and bigrams, and therefore smoothing by adding one has unnecessarily large effects on the probabilities, which makes classification inaccurate. On the other hand, when unigrams, bigrams and trigrams are tested together, only few important trigrams occur frequently. Other trigrams have much lower occurrences than unigrams and bigrams and therefore do not have a significant effect on the model. Therefore, we believe by adding trigram to unigram and bigram, we only allow the significant trigrams to correctly contribute to the model.

We observe a similar trend for the linear SVM classifiers that we test. The linear SVM classifier with unigrams, bigrams and trigrams shows the best performance. Also, classifiers with trigrams or quadgrams again have the two lowest F1 scores, even though the difference is not as big in this case.

Although the dataset has been initially filtered (removing speeches asking to yield the time and speeches containing the term "amendment"), we observe that there are still some speeches in the dataset that do not reveal any information about politicians' ideologies. For example, looking at the tfidf data, we observe that the list of top keywords includes the words that are used simply to address other people or to talk about the remaining speaking time, such as "mr", "speaker", "time" and "chairman." One example of such sentence is "mr. speaker , i reserve the balance of my time ." These words are not related to political ideologies, so for further testings we discard these words from the speeches. We also observe that there are some words from HTML entity names, such as "amp", "nbsp" and "lt". We discard these words as well. Lastly, we also discard the speech text files that have size less than 100KB, since the speeches are too short to extract much information from them, and also most of these speeches are not related to political ideologies. One example of such speech is "just for a clarification, the use of a campaign cell phone in this building ?" Table 4 and Table 5 show the results with these additional filters.

| Features | Democratic | Republican | Total |
|---|---|---|---|
| Unigram | 0.7299 | 0.7793 | 0.7546 |
| Bigram | 0.7181 | 0.7916 | 0.7548 |
| Trigram | 0.6386 | 0.7583 | 0.6985 |
| Quadgram | 0.4856 | 0.7298 | 0.6077 |
| Unigram+Bigram | **0.7421** | 0.7931 | **0.7676** |
| Unigram+Bigram+Trigram | 0.7356 | 0.7896 | 0.7626 |
| Unigram+Bigram+Trigram+Quadgram | 0.7378 | **0.7933** | 0.7656 |
| Bigram+Trigram+Quadgram | 0.6939 | 0.7857 | 0.7398 |

Table 4: F1 scores for naive bayes classifiers with additional filters

| Features | Democratic | Republican | Total |
|---|---|---|---|
| Unigram | 0.7421 | 0.7750 | 0.7586 |
| Bigram | 0.7433 | 0.7865 | 0.7649 |
| Trigram | 0.6826 | 0.7520 | 0.7173 |
| Quadgram | 0.4477 | 0.7263 | 0.5870 |
| Unigram+Bigram | 0.7700 | 0.7999 | 0.7850 |
| Unigram+Bigram+Trigram | **0.7721** | **0.8028** | **0.7874** |
| Unigram+Bigram+Trigram+Quadgram | 0.7684 | 0.7987 | 0.7835 |
| Bigram+Trigram+Quadgram | 0.7331 | 0.7844 | 0.7588 |

Table 5: F1 scores for linear SVM classifiers with additional filters

Table 4 shows an improvement in the best F1 score by about 3% compared to Table 2. F1 scores generally improve, except for the classifier with quadgrams, which experiences a drop in the score. As we remove more irrelevant information from the filter, it becomes clearer that quadgrams are not useful features to include, and we believe this is because each quadgram occurs too infrequently that it does not provide much useful information to the classifier.

Similarly, Table 5 shows an improvement compared to Table 3, except for the case of quadgrams. Interestingly, the F1 scores for the Republican speeches are higher than those for the Democratic speeches in most cases for both naive bayes and linear SVM classifiers. We believe this is because the Republicans tend to use jargons and specific terms more often than the Democrats do, and therefore it is easier to distinguish the Republican speeches. Linear SVM classifiers perform consistently better than naive bayes classifiers in our experiment. Other features, such as setting the minimum frequency for each word, are tried but do not produce a higher F1 score. Overall, the linear SVM classifier with unigrams, bigrams and trigrams showed the best performance, with the F1 score of 0.7874 on average. Since the number of Democratic and Republican speeches in the dataset are approximately the same, this means our model is able to correctly classify a speech as Democratic or Republican at about 79% of the time. In the next section, we use this model to identify the political ideologies of the 2008 and 2012 presidential election candidates.

## 6. Identifying Political Ideologies Using Linear SVM

As discussed previously, we measure a politician's ideology by using the following ratio:

$$\frac{\text{the number of speeches classified as Democratic}}{\text{the total number of speeches}}$$

100% means all of the speeches are classified as Democratic, and 0% means all of the speeches are classified as Republican. There are some presidential candidates whose speeches are in the 2005 US Congress debate dataset. Since each politician is likely to use similar words and phrases to give speeches, to ensure fairness, we exclude the speeches made by the presidential candidates from the dataset when training the linear SVM text classifier. There are 14 such speeches in the dataset.

In order to know how good our results are, we compare our ratios with the ideologies of the presidential candidates identified by the website http://www.ontheissues.org [8]. OnTheIssues website offers the ideologies of U.S. politicians on various issues as well as their key quotes. It combines a politician's viewpoints on social issues and economic issues to identify the politician's overall ideology. Table 6 shows the results for the Democratic candidates. Column 2 and column 3 show the total number of candidate's speeches classified as Democratic and Republican respectively. Since each run produces a slightly different result, for each candidate, we run our linear SVM classifier 10 times and get the average.

| Name | Democratic | Republican | Total | Ratio | OnTheIssue's classification |
|---|---|---|---|---|---|
| John Edwards | 25.7 | 1.3 | 27 | 95.2% | Populist-leaning liberal |
| Hillary Clinton | 92.1 | 13.9 | 106 | 86.9% | Hard-core liberal |
| Barack Obama | 212.5 | 59.5 | 272 | 78.1% | Hard-core liberal |
| Bill Richardson | 22.5 | 8.5 | 31 | 72.6% | Populist-leaning liberal |

Table 6: Political ideologies of Democratic presidential candidates

Table 6 shows that the linear SVM classifier classifies majority of the Democratic candidates' speeches as Democratic, and the ratios are over 70% for all four candidates. OnTheIssue website classifies Clinton and Obama as hard-core liberals and Edwards and Richardson as populist-leaning liberals. Clinton and Obama indeed have high ratios, so the results from the classifier are consistent with the OnTheIssue's classifications.

| Name | Democratic | Republican | Total | Ratio | OnTheIssue's classification |
|---|---|---|---|---|---|
| Newt Gingrich | 1.8 | 13.2 | 15 | 12.0% | Hard-core conservative |
| Rick Perry | 1.4 | 8.6 | 10 | 14.0% | Hard-core conservative |
| Rick Santorum | 3 | 13 | 16 | 18.8% | Hard-core conservative |
| Rudy Giuliani | 9.7 | 27.3 | 37 | 26.2% | Libertarian conservative |
| Mike Huckabee | 5.2 | 9.8 | 15 | 34.7% | Populist-leaning conservative |
| Mitt Romney | 23.4 | 37.6 | 61 | 38.4% | Populist-leaning conservative |
| Fred Thompson | 6.7 | 9.3 | 16 | 41.9% | Moderate conservative |
| John McCain | 84.3 | 89.7 | 174 | 48.4% | Libertarian conservative |

Table 7: Political ideologies of Republican presidential candidates

Table 7 also shows that our results are consistent with the website's classifications. Gingrich, Perry and Santorum, for whom the ratios are less than 20% from our model, which means most of their speeches are classified as Republican, are indeed classified as hard-core conservatives by the website. Also, the ratios are below 50% for all the Republican candidates, which means majority of the speeches are classified as Republican for each candidate. From Table 6 and Table 7, we see that the linear SVM classifier is able to classify politicians' speeches with high accuracy.

## 7. Testing the Hypothesis

In order to test the hypothesis that the presidential candidates move towards the center of political spectrum for the general election in order to win more votes from the opposite side as well, we look at the classification probabilities for the speeches made by Obama, McCain and Romney. Figure 1 shows a histogram of classification probabilities of the speeches made by Obama for the 2008 primary election. Higher probability means it is more likely for a speech to be classified as Democratic.

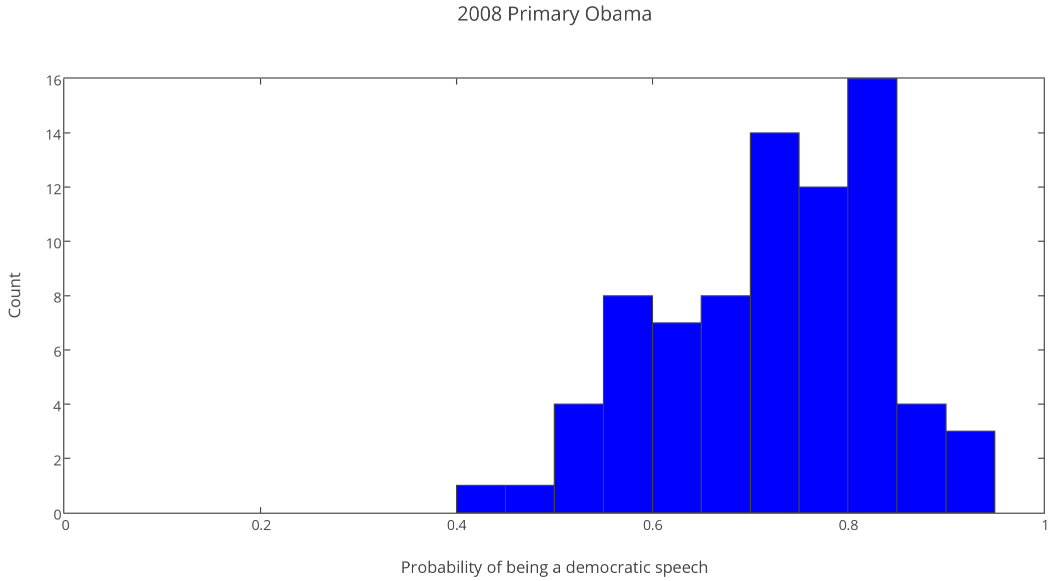

Figure 1: Probability dstribution of Obama's 2008 primary election speeches (average = 0.720)

As Figure 1 and the average classification probability suggests, we see that Obama's speeches were very democratic during the 2008 primary election. Figure 2 and Figure 3 show a histogram of classification probabilities of the speeches made by Obama for the 2008 and 2012 general elections respectively.

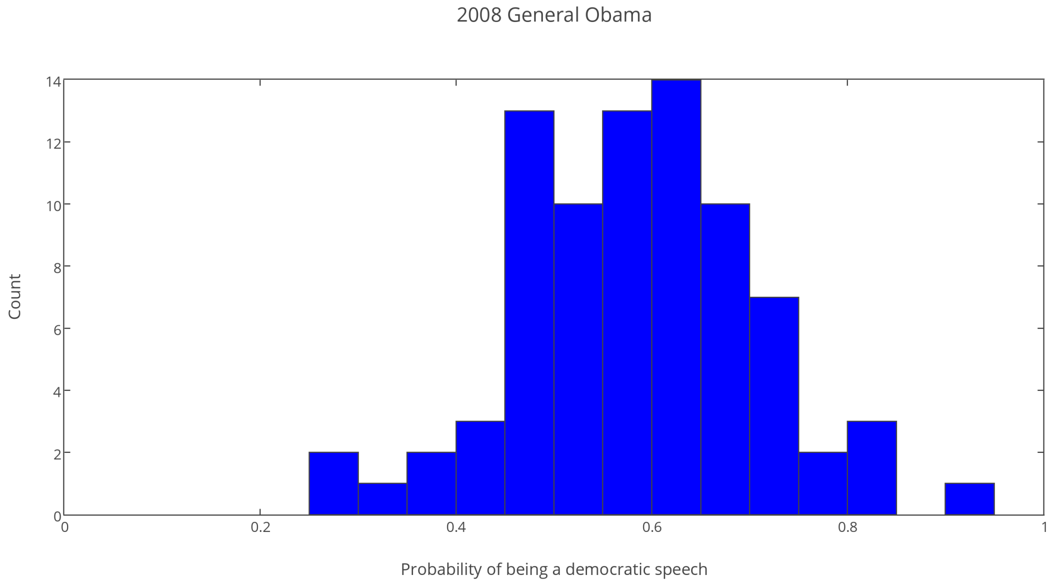

Figure 2: Probability distrubution of Obama's 2008 general election speeches (average = 0.584)

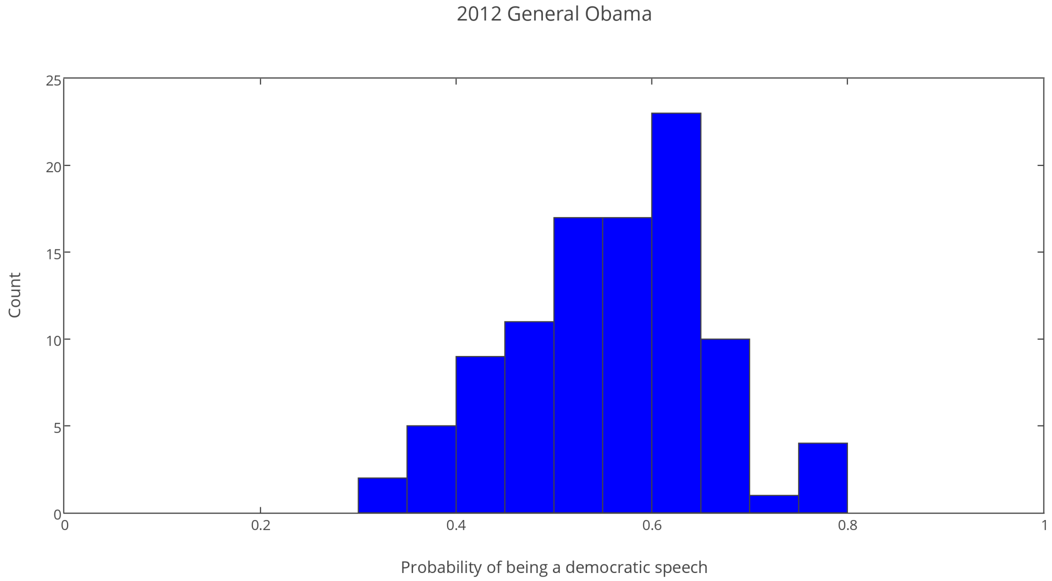

Figure 3: Probability distribution of Obama's 2012 general election speeches (average = 0.561)

In Figure 2 and Figure 3, we see that both probability distributions are shifted towards the center compared to Figure 1. The average probabilities also decrease and are closer to 0.5. This supports the hypothesis that a candidate moves towards the center for the general election. Figure 4 through 7 show the results for McCain and Romney.

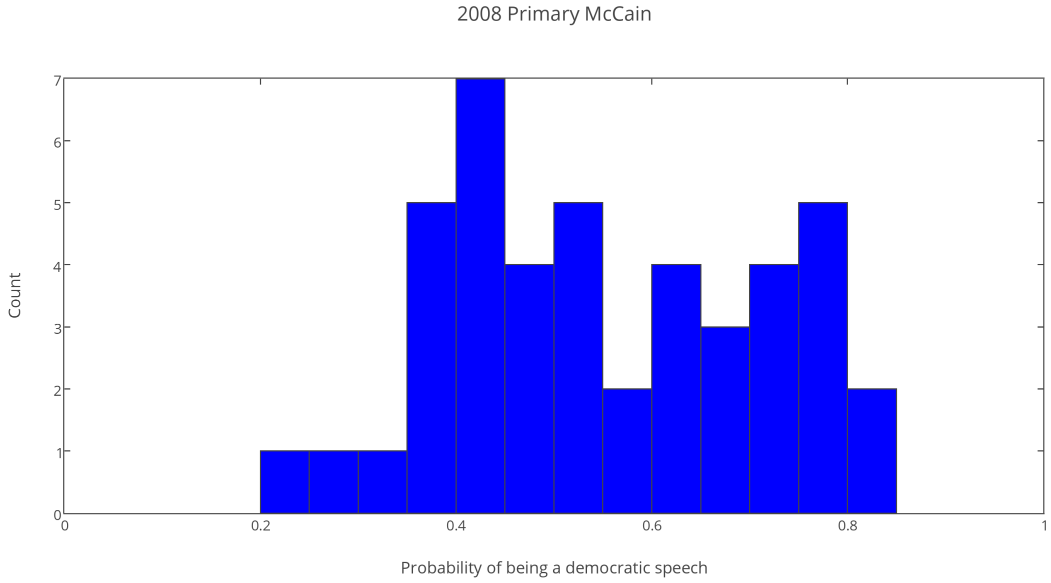

Figure 4: Probability distribution of McCain's 2008 primary election speeches (average = 0.551)

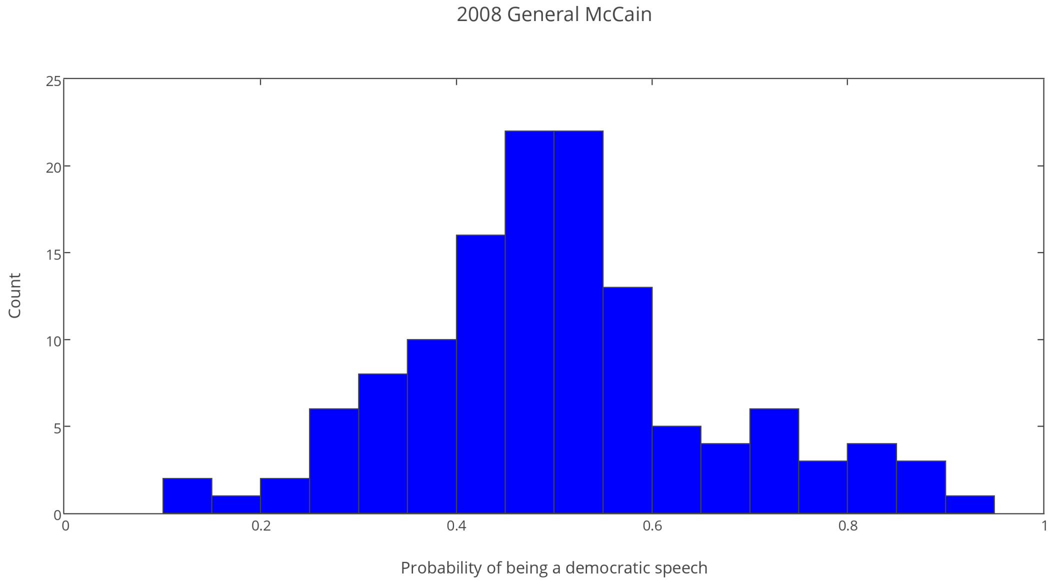

Figure 5: Probability distribution of McCain's 2008 general election speeches (average = 0.500)

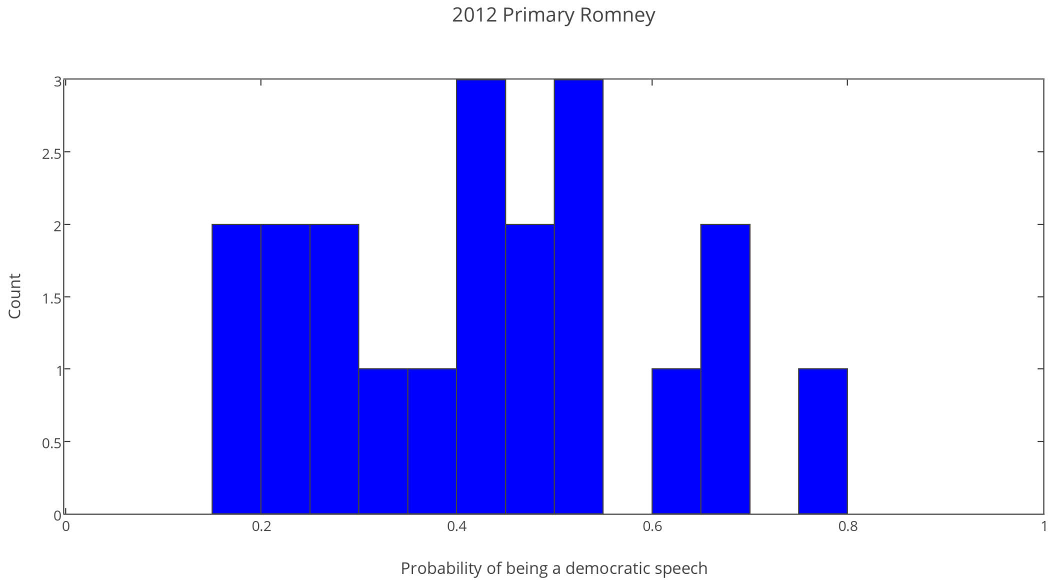

Figure 6: Probability distribution of Romney's 2012 primary election speeches (average = 0.436)

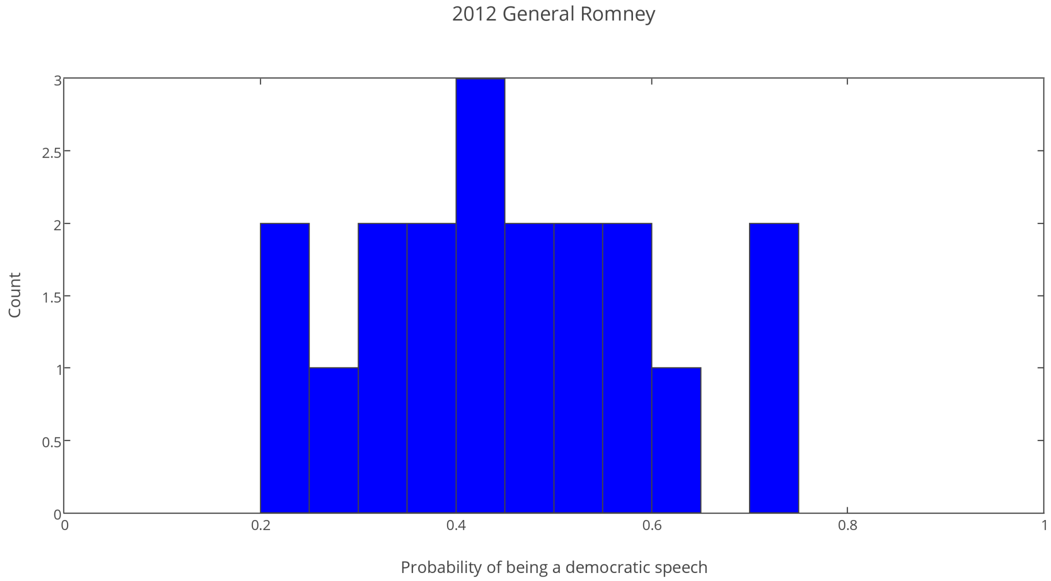

Figure 7: Probability distribution of Romney's 2012 primary election speeches (average = 0.449)

In McCain and Romney's cases, even though the average probabilities for the general elections are closer to the center than the average probabilities for the primary elections, the differences are small. The speeches are distributed around the center for both McCain and Romney for both primary and general elections. So based on the result, it is difficult to conclude that the results from McCain and Romney's speeches support the hypothesis. But at least we observe that it is not the case that the candidates moved away from the center for the general election. If more speeches from the past presidential elections were available, we would be able to test the hypothesis on a larger dataset.

It is interesting that the average probability of McCain's 2008 primary election speeches is 0.551, which means he was being more Democratic than Republican even though he was a Republican candidate. But this result is actually consistent with what happened in the 2008 presidential election, because during the election McCain was often accused by other Republicans and the supporters of the Republican Party for being liberal on many issues [9]. The term "McCain Democrat" was used to denote the Democrats who supported McCain, and according to exit polls, 10% of the Democrats voted for McCain [10].

## 8. Keyword Comparison

Using tf-idf, we also compare the keywords used by Obama, McCain and Romney in primary and general elections. Since a single word does not reveal much information about the relationship between the word and the candidate, we use bigrams and trigrams instead. We discard words or phrases that are too general, such as "United States" and "American people," or those that are duplicates or too similar to the keywords already on the list. Table 8 and Table 9 show the top ten keywords used by the three candidates in each election.

| 2008 Primary | 2008 General | 2012 General |
|---|---|---|
| 1. Health Care | 1. Health Care | 1. Middle Class |
| 2. War Iraq | 2. Wall Street | 2. Tax Cuts |
| 3. End War | 3. Middle Class | 3. Health Care |
| 4. Civil Rights | 4. Tax Breaks | 4. Auto Industry |
| 5. Working Families | 5. Rescue Plan | 5. Reduce Deficit |
| 6. Universal Health Care | 6. Insurance Companies | 6. New Jobs |
| 7. Al Qaeda | 7. Tax Cuts | 7. Insurance Companies |
| 8. Tax Breaks | 8. New Jobs | 8. Grow Economy |
| 9. Climate Change | 9. Small Businesses | 9. End War |
| 10. Dr.King | 10. Oil Companies | 10. War Iraq |

Table 8: Keywords used by Obama

| McCain<br>2008 Primary | McCain<br>2008 General | Romney<br>2012 Primary | Romney<br>2012 General |
|---|---|---|---|
| 1. Al Qaeda | 1. Health Care | 1. Economic Freedom | 1. Free Enterprise |
| 2. Health Care | 2. Small Businesses | 2. Free Enterprise | 2. Middle East |
| 3. War Iraq | 3. Wall Street | 3. Social Security | 3. Middle Class |
| 4. Middle East | 4. Nuclear Power | 4. Smaller Government | 4. Small Business |
| 5. Islamic Extremism | 5. Create Jobs | 5. National Debt | 5. Charter Schools |
| 6. Tax Cuts | 6. Raise Taxes | 6. Create Jobs | 6. National Guard |
| 7. Government Job | 7. Pork Barrel | 7. Private Sector | 7. Teacher Unions |
| 8. National Security | 8. Clean Coal | 8. Balance Budget | 8. Nuclear Weapons |
| 9. Pork Barrel | 9. Social Security | 9. Car Company | 9. Economic Freedom |
| 10. Social Security Medicare | 10. National Security | 10. Raising Taxes | 10. Private Sector |

Table 9: Keywords used by McCain and Romney

In Table 8, we see that Obama talked about various issues during the 2008 primary election, from national security to economics, environment, civil rights and health care. In contrast, the list for the 2008 general election mostly consists of terms related to economics (Wall Street, tax breaks, rescue plan, tax cuts, new jobs and small businesses). It can be argued that this is due to the 2008 financial crisis, but Table 8 suggests that economics was still Obama's main interest during the 2012 general election (tax cuts, reduce deficit, new jobs, grow economy). The website https://www.govtrack.us

[11] suggests that economics is indeed one of Obama's key interests. The website provides data about the bills that senators and representatives have sponsored, and it shows that Economics and Public Finance is the second most category in which Obama sponsored bills (13%) when he was a Senator, after Government Operations and Politics category.

In Table 9, we see that McCain mainly talked about national security during both primary and general elections (Al Qaeda, War Iraq, Middle East, Islamic Extremism, National Security), and in fact, Armed Forces and National Security is a category in which McCain sponsored the most number of bills (24%) during his term as a Senator.

Free enterprise, smaller government, greater role of private sector and economic freedom are all values that the Republican Party supports, and Table 9 shows that Romney used these phrases often during his campaign, from which we can see that Romney's viewpoint on economics is strictly Republican.

So by looking at tf-idf, we can understand not only each candidate's political ideology but also the issues that they are particularly interested in.

## 9. Future Work

It is a difficult task to find a large dataset of political speeches in text. If we could have more dataset, we would be able to build a better text classifier model. We can try other text classification algorithms as well, such as k-nearest neighbor algorithm and decision trees, and see how the performances differ. Also, if we could have more speeches from the presidential elections, the hypothesis could be tested better and a more meaningful conclusion could be made.

It is important not only to look at the classification results and probabilities but also to understand what factors play an important role in determining how Democratic or Republican each speech is. To do so, we will have to find a way to identify and discard sentences that are neutral or do not reveal any information about the speaker's political ideology. One way could be incorporating sentiment analysis.

The goal of this paper is to better understand politicians' ideologies, so the ultimate goal is to make a program that can systematically summarize a politician's ideology on different issues. It will be important to be able to classify each sentence or speech into different issues, possibly by making a list of keywords for each issue. Assigning different weights on the words depending on how close they are to a keyword and incorporating the information about part of speech could be a way to extract important information from speeches.

## 10. Conclusion

In this paper, we identify politicians' ideologies using their speeches and a linear SVM classifier. The result shows that the classifier can identify each politician's ideology accurately. We also hypothesize that candidates move towards the center for the general election in order to win more votes from the opposite side, and we test the hypothesis using the speeches from presidential election. Based on the the result, there is some evidence that supports the hypothesis. We also compare the keywords used by the presidential candidates between primary elections and general elections.

## References

[1] R. G. Holcombe, *Public Sector Economics*. Upper Saddle River: Pearson Prentice Hall, 2006.

[2] Y. Riabinin, "Computational identification of ideology in text: A study of canadian parliamentary debates," Master's thesis, University of Toronto, 2009.

[3] M. Iyyer, P. Enns, J. Boyd-Graber, and P. Resnik, "Political ideology detection using recursive neural networks," *In Proceedings of ACL 2014*, 2014.

[4] M. Thomas, B. Pang, and L. Lee, "Get out the vote: Determining support or opposition from congressional floor-debate transcript," *Proceedings of EMNLP 2006*, 2006.

[5] Y. Sim, B. D. L. Acree, J. H. Gross, and N. A. Smith, "Measuring ideological proportions in political speeches," *In Proc. of EMNLP*, 2013.

[6] Support Vector Machines. [Online]. Available: http://scikit-learn.org/stable/modules/svm.html

[7] Congressional Speech Data. [Online]. Available: http://www.cs.cornell.edu/home/llee/data/convote.html

[8] Ontheissues. [Online]. Available: http://www.ontheissues.org

[9] John McCain: Liberal In Disguise. [Online]. Available: https://www.gunowners.org/mcdisguise.htm

[10] Inside Obama's Sweeping Victory. [Online]. Available: http://www.pewresearch.org/2008/11/05/inside-obamas-sweeping-victory/

[11] Govtrack. [Online]. Available: https://www.govtrack.us