Predicting Political Affiliation

TA Mentor:
Subba Reddy Oota

Submitted By:
Mehak Agarwal(201405517)
Prerna Chauhan(201405544)
Gunjit Bansal(201405568)

ABSTRACT:

We present the analysis of the political speeches made by members of the Democratic and Republican parties in the United States. Here we throw light upon to learn which features best differentiate speeches made by the two parties, and present a comparative analysis of various models used to classify speeches as either Democrat or Republican.

INTRODUCTION:

Nowadays, division among the political parties in the United States has become an increasingly large problem. When members of one party try to bridge the divide and provide support in times of trouble, they are met with criticism from their own party.Polls and polarization research show that partisan divisions drive the debate amongst those who are responsible for solutions [1].

This paper uses a variety of supervised techniques to filter out these divisions, based on certain features extracted from the content and rhetoric of political speeches.

DATA COLLECTION AND HANDLING:

The dataset consists of **344 speeches** (171 Republican /173 Democrat) by American politicians delivered during or after the presidency of Franklin Roosevelt. All of the data was collected by scraping online sources for text.

The data is heavily biased towards presidents, but we have also included speeches by Congressional

politicians, governors, and other major political figures to help generalize our model for the future. Preprocessing of the data was done, to remove all the non ASCII characters and other stop words.

Finally the data left was clean and rich in content words.

APPROACH:

We defined a set of **277 features** based on various categories of features such as Stylometric,Lexical and Function words and Semantic.
These features values were then calculated on 344 documents and our feature vector was created and uploaded for CROSS-VALIDATION to divide it into training and testing data.

FEATURES:

- Top 20 words differentiating democratic speeches from republican speeches.
- Top 20 words differentiating republican speeches from democratic speeches.
- Sentence length for each document.
- POS features

Initially we used **SVM, LDA, Naive Bayes** for training and discovered that SVM was outperforming all the others.Since our dataset is very small, the best results were given by **3-fold Cross-Validation**.

EXPERIMENTS AND RESULTS:

The following results are based on 3-fold Cross_validation of trained models over dataset of 344 documents.

==========LDA==========

Experiment Set 1 :
Features : only a few Document Statistics and features with feature extraction though Pearson Correlation + few Stylistic (**PoS**) features.
Accuracy : 64.0 %

Experiment Set 2 :
Features : Added more Content Words,along with all previous features.

Accuracy : 65.3 %

Iteration	Accuracy
1	**60.115942029**
2	**61.2173913043**
3	**65.3623188406**
4	**70.0144927536**
5	**66.8115942029**

==========SVM==========

Experiment Set 1 :

Features : only a few Document Statistics and features with feature extraction though Pearson Correlation + few Stylistic (**PoS**) features.

Accuracy : 64.0 %

Iteration	Accuracy
1	**68.115942029**
2	**65.2173913043**
3	**75.3623188406**
4	**71.0144927536**
5	**76.8115942029**

Experiment Set 2 :

Features : Added more Content Words,along with all previous features.

Accuracy : 65.3 %

==========NAIVE BAYES==========

Experiment Set 1 :

Features : applied Bayes' theorem with the "naive" assumption of independence between every pair of features.

Accuracy : 71.1764705882 %

Iteration	Accuracy
1	**77.4509803922**
2	**67.6470588235**
3	**71.568627451**
4	**80.3921568627**
5	**58.8235294118**

==========**Logistic Regression**==========

Experiment Set 1 :

Features : only a few Document Statistics and features with feature extraction though Pearson Correlation + few Stylistic (**PoS**) features.

Accuracy : 64.0 %

Experiment Set 2 :

Features : Added more Content Words,along with all previous features.

Accuracy : 68.6956521739 %

Iteration	Accuracy
1	**69.5652173913**
2	**57.9710144928**
3	**71.0144927536**
4	**73.9130434783**
5	**71.0144927536**

CONCLUSION AND FUTURE WORK:

We observe that Syntactic and Lexical features can only help in the task upto a limit then we need more contextual and semantic information of the text create features out of that information. One such information that we captured was POS n-grams and it increased our accuracy significantly.

As future work ideas, I believe more semantic and grammatical information of the text should be extracted, an example could be depth of dependency trees, depth of noun phrases, adjectival phrases, etc.

REFERENCES:

[1]:“History & Politics Out Loud: Famous Speeches”.<http://www.wyzant.com/resources/>
[2]:“American Rhetoric Speech Bank”. <http://www.americanrhetoric.com/>
[3]:“Presidential Rhetoric”. <http://www.presidentialrhetoric.com>