

Members	Savannah Baron and Varsha Kishore		
Title	Predicting Political Ideology		
Hours	Savannah (8), Varsha (8)		
Predicting	Our goal is to classify the political ideology of sentences. In particular, a binary classification problem to start (conservative/liberal) and a three class problem later (conservative/liberal/neutral).		
Data	We will be using data from the Ideological Books Corpus (IBC). This data set contains sentences from authors with known political standings. All sentences are labeled as either liberal, conservative or neutral. There are 4062 sentences total, with 2025 of them being liberal, 1701 conservative, and 600 neutral.		
Features	Currently we are using bag of words for our features, so our feature space is dependent upon the number of words we choose to use in our vocabulary. We've been experimenting with using between 2000 and 5000 words, and have been experimenting with how to form and select these words in a variety of ways. Some examples of things we have tried so far include removal of stop words, n-grams, using stem words, and Tf-Idf (Term Frequency-Inverse Document Frequency).		
Models	So far, we are focusing on a textual classification problem. So, we are using the ever popular SVM and Logistic Regression models to begin with because these models have been used for other similar textual classification problems. Using Logistic regression also gives us a direct probabilistic interpretation and this might be useful in our analysis. An advantage we have with SVM's is that we can use kernels to encode additional relationships between words.		
Results	Model	Best Params	Average Accuracy
	SVM	b	c
	LogReg	b	c
Problems	The lone professional paper that uses this dataset achieves a maximum accuracy 69.3% (using recursive neural nets (RNN)). This is not that great, so whether we'll be able to improve from where we are currently, or beyond that far enough to get interesting results is an open question. Additionally, we don't have a lot of data, which may limit the amount that our model will generalize.		
Future	We are planning to look into better feature extraction techniques. This includes looking for negation in sentences, n-grams and specifically politically biased n-grams, <code>nltk</code> chunking which groups parts of sentences. We also plan to pursue multiclass classification to include neutral examples using random forests. Finally, LDA seems as though it may be interesting in conjunction with our supervised models if we used the results as inputs, however tuning the number of topics in LDA is difficult and subjective, so we plan to time box ourselves to 2 hours initially, and evaluate the results.		
Specific Questions	We feel like we don't have great intuition on how to improve the accuracy of our models, or what sorts of features will perform best. This is perhaps more of a "how to do machine learning" question than a specific one though. Are there other classifiers that we haven't talked about that could give us better results with text classification.		