Proposal

Jun Wang

**Introduction**

Sentiment analysis is an important application of natural language processing. For example, people can apply the sentiment analysis to speeches of a politician to predict the political opinion. People can also utilize the sentiment analysis to the movie reviews to predict the performance of box-office of a specific movie. Here, I will utilize the sentiment analysis technique to analyze the product reviews and predict whether people like or dislike a specific product.

**Model**

In my final project, I will implement the naive bayes model and support vector machine to predict the sentiment label for the product reviews. The Multinomial naive bayes model of unigrams without feature selection will be selected as the baseline methods. The Bernoulli naive bayes model, support vector machine with feature selection will be chosen as the improved models for sentiment analysis.

**Dataset**

Since nltk module has two datasets named "product\_reviews\_1" and "product\_reviews\_2", I will simply use these two datasets for training and testing. I will also try to find some way to extract the product reviews from Amazon, then train and test these reviews if possible.

**Implementation**

The implementation process will be divided into two phases: preprocessing and training.

In preprocessing phase, I will extract all the unigrams and bigrams from the training set as the features at first. Considering some of the unigrams and bigrams may appear in both of the positive reviews and negative reviews, these features are harmful for predicting. Thus, I will calculate the information gain for all the features and rank these features according to the information gain. Finally, select the features with more information gain as the features for predicting.

In training phase, I will implement the algorithm of Multinomial naive bayes model, Bernoulli naive bayes model and support vector machine, then use these features with the corresponding labels of the reviews to train the models. Finally, predict the labels of the reviews in the test set.

**Evaluation**

I will build the confusion matrix for all the models in each iteration of the cross validation. Then, use t-test to compare the accuracy, correlation coefficient and the recall with different models after cross validation.