ASSIGNMENT 1: SENTIMENT MINING

Deepanshu Jindal
2016CS10312

Text Preprocessing

Preprocessing steps combination used for final model:

- convert to lowercase
- remove stopwords: stopword corpus curated for the given data
- mark words following negation with suffix NEG uptil the first punctuation mark
- remove punctuation marks
- identifying capital letter blocks

Tried stemming, replacement of numerals with <NUM> and searching for regex like “[\d] star(s)” but it did not give good results.

Data vectorization

- TFIDF vectorization
- Bigram features
- Limit to most frequent 200k bigrams
- Minimum document frequency of 3

Higher number of features beyond a limit seemed to hurt the model performance.

Model

Trained sklearn’s LogisticRegression model. The biggest trick here was to use not the model’s prediction but the model’s expected prediction. LogisticRegression model also gives the probability of each class.

Using these probabilities, I computed the prediction value as $\Sigma i * p_i$ where i takes value from {1, 2, 3, 4, 5}, instead of giving a class label as output.

Another model I trained with good scores was an ensemble of LogisticRegression, LinearSVC and Naive Bayes, where the first two were the main predictors and Naive Bayes was for tie breaking. I also tried upweighting class priors to better suit for skewed data, it gave rise in accuracy but no improvement in MSE.
However, this model was outperformed with the above one.