COMP 550

Assignment 1 Report

Instructor: Prof. Jackie Chi Kit Cheung

Haochen Liu

260917834

I. Problem Setup

The goal of this assignment is to find out suitable preprocessing decisions for sentence-level sentimental classification. The logistic regression model is used for the analysis. To find out the best-performing decision, I combined them as follows:

Unigram	Bigram
Lemmatization + Unigram	Lemmatization + Bigram
Stemming + Unigram	Stemming + Bigram
Remove Stop Words + Unigram	Remove Stop Words + Bigram
Remove Stop Words + Lemmatization + Unigram	Remove Stop Words + Lemmatization + Bigram
Remove Stop Words + Stemming + Unigram	Remove Stop Words + Stemming + Bigram

II. Experimental Procedure

Data was loaded from files and then cleaned by removing punctuations and numbers, labeled with 1 for positive and 0 for negative, then divided into the training set with 80% of the data and testing set with the rest 20%. The sentiment of data is evenly in both training and testing sets – each has 50% of positive reviews and 50% of negative reviews. After this, the data is tokenized and applied with different preprocessing decisions. The performance of models is evaluated by training accuracy, the average accuracy of 5-fold cross-validation, and the accuracy on the unseen training set.

III. Parameter Settings

To maintain reproducibility, the random state of the logistic regression model was set to 1.

IV. Results and Conclusions

	Unigram	Lemmatization + Unigram	Stemming + Unigram	Remove Stop Words + Unigram	Remove Stop Words + Lemmatization + Unigram	Remove Stop Words + Stemming + Unigram
Training Accuracy	0.97936	0.97385	0.96178	0.97643	0.97221	0.95697
Validation Accuracy	0.75214	0.75320	0.75835	0.74745	0.74417	0.74780
Testing Accuracy	0.76782	0.76360	0.76876	0.76079	0.75891	0.76595
	Bigram	**Lemmatization + Bigram**	**Stemming + Bigram**	**Remove Stop Words + Bigram**	**Remove Stop Words + Lemmatization + Bigram**	**Remove Stop Words + Stemming + Bigram**
Training Accuracy	0.99953	0.99941	0.99953	0.99906	0.99906	0.99906
Validation Accuracy	0.68472	0.68906	0.69176	0.60253	0.61086	0.61461
Testing Accuracy	0.70263	0.70450	0.71341	0.60788	0.61398	0.61726

It could be clearly seen that the models adapted unigram had better performance than those with bigram. And the best performing model was the one adapted stemming and unigram, with a testing accuracy of 0.76876. According to the experimental data, models adapted stemming had slightly better performance than those with other preprocessing decisions, but their influence of stemming, as well as lemmatization, was relatively mild.

It was apparent that for unigram models, removing stop words had only mild influence, while bigram models with stop words removed is more poorly performed than those kept them. I believe this was caused by the mechanism of bigram. For example, “not” and many other words with negative meanings were included in the collection of stop words. By removing them, the meaning of a sentence is highly possible to be completly opposite, and this would result in misclassification.

According to the comparison of training accuracy and testing accuracy, all these preprocessing decisions causes overfitting problem, but bigram models, especially those with stop words removed, were impacted even more. Compared to unigram, the vocabulary of bigram models is richer, but each one has a very low frequency. With a relatively small training sample and lots of token types, the trained model will be more likely to misclassify unseen data, which will result in overfitting.

In conclusion, the best performing model for sentence-level sentimental analysis is stemming with unigram, and unigram models performed better than bigram models. Removing stop words did not result in better performance for bigram models while applying lemmatization and stemming would slightly improve it.

V. Limitation of Study

For this project, due to the lack of linguistic knowledge, I chose a brute-force way to solve the problem, and it is possible that I have missed some combinations of preprocessing decisions that have better performance. Also, this experiment did not include a deeper analysis of the raw data. I believe by sorting, such as tag sentences as questioning, declarative, rhetorical, and so on would help in choosing better preprocessing decisions.