Assignment 1 Report : Sentiment Mining

Shubham Jain
2016CS10317

Step -1 (Pre-processing Text) :

Preprocessing steps combination used for final model:

- remove stopwords from a custom list
- converted to lowercase. Performed no stemming and no lemmatization since they decreased performance.

Step-2(Training Model):

- TFIDF vectorization
- Used Trigram features
- Limit to most frequent 150k 1/2/3 -grams
- Minimum document frequency of 5

Step3 (Post-processing): (Most Important):

The idea is to optimise the cost metric and not actually the accuracy. Since we have a probability distribution and the cost metric is squared error, we know that the mean taken over the 5 classes will give the best result. This in itself reduced the cost by (>20k).

But, the above idea is only true if the probability distribution is already correct. What if it is not? It maybe that the model is less confident over the predictions than it should be, and it could be the other way around too. Thus, I normalised the probabilities by taking power of probabilities. The best result came at p^(1.3). It means that the model was indeed underfitting.

This could also have been corrected by using parameter “C” as Shashank did. Increasing “C” reduces regularisation cost and enables the model to overfit.

This part led to a further decrease in cost by about (3k).