Kyle Ashburn (kyle0718)

Professor Yue Wang

Information and Library Science 690.270

5 November 2021

One-page memo: Homework 3

I started this homework working to change my approach to better fit the best approaches from homework 2. I started by changing from count vectorization to TFIDF vectorization and trying to use review length as a feature. I followed up by trying to tune my C parameter before switching to adding in extra data. I found and cleaned/converted extra Amazon review data and Yelp data. Using that, I iteratively added bigrams and trigrams to my feature set.

Unfortunately, the Amazon data had overlap with the test data, so I had to start over. Using the extra data but with the test data removed, I worked on implementing Vader sentiment as a feature along with bigrams and trigrams. I ran into issues when I combined the review heading with the review text, so I dropped the header text from my training data. Ultimately, I worked on tuning my C to squeeze out an improvement in my model.

One consistent trend in the performance was an accuracy improvement of several percentage points from the validation to the test. I suspect it’s due to an increase in coverage by the data when it combines the validation data with the training data.

**Here’s a partial table of my models and results**

|  |  |
| --- | --- |
| Model Description | Model Accuracy on Validation | Test |
| Baseline approach Ray gave | On test: 0.88633 |
| Adding the Yelp data and using Vader sentiment on logistic regression | On test: 0.88500 |
| Unigram, bigram, trigram representation using TFIDF and stopwords with vader sentiment and extra data (not Yelp though) and a C of 1 on my logistic regression | On test: 0.94666 |
| Unigram, bigram, trigram representation using TFIDF and stopwords with vader sentiment as a feature and extra data (Yelp and Amazon) and a C of 40 on my logistic regression. | On validation: .93595  On test: .95000 |