Jenna Gatley

Dr. Wang

INLS 690-270

3/17/2024

Kaggle Competition: Methodical Accuracy

My baseline accuracy through the provided source code was 0.87741. My first step was to walk through the steps of the code. To tweak the current code to its best capabilities, I made the decision to stick with the logistic regression method for this assignment. When I saw that this code used unigrams only, I immediately changed the vectorizer to include unigrams and bigrams. I also queried an open ai source to see common suggestions for improving logistic regressions. From this, I saw suggestions such as incorporating cross validation, removing stop words, and trying different data cleaning. I immediately incorporated code for cross validation alongside the inclusion of unigrams and bigrams which increased my score to 0.90439. This was my highest score achieved but I kept tweaking different sections to try and understand what was the most impactful. I included stop words (manually as I could not figure out how to import them). This slightly lowered my accuracy to 0.89906 and I found that it may be the specific words I was choosing to remove. I looked over some of the training reviews and chose to keep ‘and’ and ‘in’ (so removing them from my stop word list) and my accuracy jumped back up to 0.90173. I felt like I had reached the farthest progress on the three main sections I wanted to adjust (n-grams, cross-validation, and stop words) so I decided to leave my code where it was, especially as several of my exploratory adjustments dropped the accuracy and I struggled to understand why.

Display name: Jenna Gatley