Mason Holcombe
1/17/2026

Homework 1 Writeup

For Homework 1, my first idea was to look into words that uniquely describe positive and negative reviews. Words that are in the list of positive words (or negative words) that are not in the other list. If a review has words that are in one of these positive or negative only lists, then that is likely to be the sentiment. To do this I wrote a simple tokenizer that breaks each review into a list of words and removes punctuation. After tokenizing all of the reviews and building the positive & negative word lists, I saw plenty of times where someone uses a seemingly nice word to negatively describe a movie, or vice versa. Example: "I hate how much I rewatched this movie!"

This led me to incorporate frequency into this classifier and look at how many times these words were used in the positive/negative word lists. After seeing how much each word appears in the positive/negative word lists, I removed words that were seen in both datasets >= 500 times. I'm considering this a hyperparameter of the classifier and is a tool to remove filler words from the sentence and sway the results. For example, the word "movie" was seen in the negative reviews more than the positive reviews but the word "movie" doesn't intrinsically have a positive/negative connotation. There is something to be said about if a review contains the word "movie", then the review might be more elaborate and detailed in the sentiment, but this semantic meaning is lost in this particular classifier. After removing these words from the review, I normalized the remaining to determine what percentage of the times the word was positive/negative. To give a final prediction, I averaged the final positive/negative percentages and argmax to predict the sentiment. In this case, index 0 = Positive, index 1 = Negative.

This method is very similar to naive bayes and looking across words of the review to see the frequency of this word in the various classes. In this scenario the prior can be disregarded as there were an equal number of positive/negative reviews.

This approach achieved **95.1%** accuracy **(10,138 / 10,662)** on the positive/negative reviews provided.