NAÏVE BAYES TEXT CLASSIFICATION

Use Naive Bayes algorithm for text classification.

The dataset for this problem is a subset of the IMDB movie review dataset. Given a movie review, task is to predict the rating given by the reviewer. Separate training and test files containing 25,000 reviews samples each are given. A review comes from one of the eight categories (label). Here, label represents rating given by the user along with the review.

You are provided four files i) Train text ii) Train labels iii) Test text iv) Test labels. Text files contain one review in each line and label files contain the corresponding rating.

(a) Implement the Naive Bayes algorithm to classify each of the articles into one of the given categories. Use the Laplace smoothing for Naive Bayes to avoid any zero probabilities. Use c = 1.

(b) Test accuracy that you would obtain by randomly guessing one of the categories as the target class for each of the articles (random prediction). What accuracy would you obtain if you simply predicted the class which occurs most of the times in the training data (majority prediction)?

(c) Draw the confusion matrix for (a). What other observations can you draw from the confusion matrix?

(d) Remove the stop words and perform stemming and then create model.

(e) Come up with at least two alternative features and learn a new model based on those features. Examples: bi-grams, tri-grams etc.