Part 2:

1. Category data
    a. Categories: ['ENTERTAINMENT' 'POLITICS' 'WELLNESS' 'STYLE & BEAUTY' 'TRAVEL']
    b. Instances:
        i. POLITICS 35600
        ii. WELLNESS 17931
        iii. ENTERTAINMENT 17361
        iv. TRAVEL 9883
        v. STYLE & BEAUTY 9813
    c. It is a balanced dataset
2. Results
    a. The three method that were best were the BOW: lemmatized with BOW, snowball stemming with BOW, Porter Stemming with BOW
    b. They both assume that independence and thus are simple to run making it perform better.
    c. TF-IDF vectorization may perform worse with Naive Bayes because it creates continuous variables, and Naive Bayes typically works better with count-based discrete features, like those from BoW. LSA creates synthetic features from original ones, and can also introduce information loss, potentially eliminating important features for classification.
3. Results for steming and lem
    a. The highest improvement is seen with Lemmatized BoW:
        i. Accuracy: Increases from 0.8940 (none_X_BOW) to 0.9018 (lemmatized_X_BOW)
        ii. F1 Macro: Increases from 0.8775 (none_X_BOW) to 0.8867 (lemmatized_X_BOW)
        iii. F1 Micro: Increases from 0.8940 (none_X_BOW) to 0.9018 (lemmatized_X_BOW)

b. Stemming and lemmatization improve results by breaking down words to their simplest form, making the data cleaner and reducing distractions for the model. These methods help focus on the key content of the text. Lemmatization performs best as it understands context and language structures better, giving a more precise simplification of words than stemming.