Text Classification using Custom Multinomial Naive Bayes Classifier

1. Introduction:

In the world of computers understanding human language, one important task is sorting text into categories. Think of it as teaching a computer to distinguish between emails about sports, science, or religion. We used a technique called Multinomial Naive Bayes, which helps in this sorting process. It's like the computer making educated guesses based on words it knows.

2. Implementation:

We built a special tool called CustomMultinomialNB that learns from a test set of texts to make these guesses. It figures out the chances of seeing certain words in different categories. For example, it might learn that words like "goal" and "score" are often related to sports. The tool was trained on a set of texts to get good at guessing.
The training process involves calculating class probabilities and word probabilities, incorporating Laplace smoothing to handle words absent in the training set.

The implementation integrates with scikit-learn's CountVectorizer to convert text data into a bag-of-words representation. Training involves learning from the frequency of words in each category, enabling the classifier to make informed predictions.

3. Performance Evaluation Result Using Test Dataset:

The classifier's performance was assessed on the 20 Newsgroups dataset. Predictions were made on a separate test set, and the accuracy score, a widely-used metric, was calculated. The accuracy score represents the percentage of correctly predicted labels compared to the true labels in the test set.
Accuracy: 85.15%