Methodology Report

This report details a refined methodology for building a sentiment analysis system using a Naive Bayes classifier.

Data Preprocessing

1. **Data Collection:**
    - Similar to before, gather a labeled dataset of text data (e.g., movie reviews) with sentiment labels (positive, negative, or neutral).
2. **Text Cleaning:**
    - Perform the same cleaning steps as before (lowercase, remove punctuation/special characters, stop words,html tags etc).
    - Consider additional cleaning techniques:
        - **Normalization:** Convert numbers to text (e.g., "2" becomes "two").
        - **Emoticon handling:** Convert emoticons to sentiment labels (e.g., ":)" to "positive").
        - **Named Entity Recognition (NER):** Identify and potentially remove names (actors, movies) as they might not contribute directly to sentiment.
3. **Feature Extraction:**
    - Implement approach from before:
        - **Bag-of-Words Model** for word frequency.

Model Training

1. **Naive Bayes Classifier:**
    - different Navies Bayes Model are used
    - ComplementNB
    - MultinatinominalNB
    - BernoulliNB

      and see how accuracy each model can be
2. **Training the Model:**
    - Split the preprocessed data into training, validation, and testing sets.
        - The validation set helps fine-tune hyperparameters before final evaluation on the testing set.

3. **Hyperparameter Tuning:**
    - Use the validation set to experiment with different hyperparameters (e.g., smoothing parameter, number of features). Metrics like accuracy or F1-score can be used for evaluation.
    - Common libraries like scikit-learn offer tools for hyperparameter tuning.

4. **Feature Selection:**
    - Analyze feature importance to identify the most impactful words/phrases for sentiment classification.
    - Techniques like Chi-squared test or information gain can be used for selection.

Model Evaluation

1. **Evaluate the final model on the unseen testing set.**
2. **Calculate metrics:**
    - Accuracy: Overall percentage of correctly classified reviews.
    - Precision: Proportion of reviews classified as a specific sentiment that are actually of that sentiment.
    - Recall: Proportion of reviews with a specific sentiment that are correctly classified.
    - Consider F1-score, which combines precision and recall