1. In this task, the raw text data from both the training and testing datasets was pre-processed to convert it into a numerical format suitable for machine learning models. The following steps were applied:
    a. Lowercasing: All text was converted to lowercase to ensure that the model treats "Word" and "word" as the same.
    b. Punctuation Removal: All punctuation marks were removed from the text to avoid unnecessary tokens.
    c. Tokenization: The text was split into individual tokens (words) for further processing.
    d. Stopword Removal: Commonly used words such as "the", "and", "is", which do not contribute much meaning, were removed using NLTK's stopword list.
    e. Stemming: Words were reduced to their base or root form (e.g., "running" becomes "run") using the Porter Stemmer to reduce variations of the same word.
    f. TF-IDF Vectorization: The processed text was converted into numerical features using Term Frequency-Inverse Document Frequency (TF-IDF). This method assigns importance to words based on their frequency in a document relative to their frequency in the entire corpus. The TfidfVectorizer was used with a maximum of 5000 features to transform the text into feature vectors.

This process ensured that the raw text data was transformed into a format that could be used as input for machine learning models.

2. For this task, multiple classification models were considered, including:
    a. Decision Tree
    b. Random Forest
    c. AdaBoost
    d. Gradient Boosting

Among these models, Random Forest was chosen as the optimal model after performing cross-validation. A Random Forest is a robust and ensemble-based model that trains multiple decision trees on different parts of the data and averages the predictions, leading to reduced overfitting and improved accuracy.

The following parameters were tuned using GridSearchCV for the Random Forest model:

n_estimators: This parameter determines the number of trees in the forest. We searched over values such as 100, 200, and 300.

max_depth: This parameter limits the depth of the trees to prevent overfitting. Values such as None, 10, 20, and 30 were explored.

min_samples_leaf: This parameter specifies the minimum number of samples required at a leaf node. It was tested with values of 2, 5, and 10.

GridSearchCV with 5-fold cross-validation was used to find the best combination of these parameters based on the training data.

3. After tuning the parameters of the Random Forest model, the best combination of hyperparameters was selected:

- n_estimators = 300

- max_depth = 20
- min_samples_leaf = 5

The model was trained on the entire training dataset with these parameters, and its performance on the training data was evaluated using several metrics:

- Training Accuracy: The model achieved a high accuracy on the training set, indicating that it was able to fit the data well.
- Precision, Recall, and F1-score: These metrics were calculated for each class (category), and the weighted averages were reported. The model performed well across different categories, balancing precision and recall effectively.
- Confusion Matrix: The confusion matrix showed that most categories were correctly classified, with few misclassifications.

Overall, the Random Forest model with the chosen parameters showed strong performance on the training data, achieving both high accuracy and balanced performance across different categories.

4. In this section, we used three main models: **Decision Tree**, **Random Forest**, and **Gradient Boosting**. Each model was trained and evaluated, but **Random Forest** was found to provide the best performance based on cross-validation results.

Conclusion:

- **Random Forest** was selected as the final model after tuning parameters and comparing multiple models.
- The model's strong performance on the training data indicates that it is well-suited for this classification task.
- The preprocessing steps, including stemming and TF-IDF vectorization, played a crucial role in improving the model's ability to interpret and learn from the textual data.