Homework Assignment: Classification of IMDB Movie Reviews

Goal:

In this assignment, students will work on classifying IMDB movie reviews using different embeddings and a Random Forest classifier. You will implement three different text embedding techniques: Word2Vec, fastText, and BERT, to process the text data and observe their effectiveness for text classification.

Tasks

Task 1: Dataset Preparation

1. Download the IMDB dataset:
    - Use a standard library like `nltk` or download the dataset directly from a source like TensorFlow Datasets.
    - Split the dataset into training and testing sets.
2. Preprocess the text data:

- Perform tokenization, lowercasing, and removal of unnecessary characters like punctuation.
- Ensure the text is ready for embedding techniques.

Task 2: Word Embeddings

Implement the following embedding methods:

1. Word2Vec Embeddings

- Train a Word2Vec model on the IMDB dataset or use a pre-trained Word2Vec model (e.g., GloVe or Google News vectors).
- Convert each review into an embedding by averaging the vectors of the words in the review.

2. fastText Embeddings

- Use the `gensim` library to load pre-trained fastText embeddings or train your own fastText embeddings on the IMDB dataset.
- Similar to Word2Vec, create review embeddings by averaging the word vectors.

3. BERT Embeddings

- Use the Hugging Face `transformers` library to obtain embeddings for the reviews using a pre-trained BERT model - all-MiniLM-L6-v2.
- Generate sentence-level embeddings for each review using the `[CLS]` token representation or average pooling of all token embeddings.

Task 3: Classification Using Random Forest

1. Use the Random Forest classifier from `sklearn` for text classification.
2. Train separate Random Forest classifiers for each of the embedding techniques (Word2Vec, fastText, BERT).
3. Evaluate the performance of each model using appropriate metrics such as accuracy, precision, recall, and F1-score on the test set.

Task 4: Compare and Report

1. Compare the performance of the models trained with different embeddings.
2. Discuss the strengths and weaknesses of each embedding technique based on your results.
3. Summarize your findings in a brief report.