Project Report 2: Exploring and Comparing Text Embeddings

Objective

The objective of this project is to explore different types of text embeddings, including Word2Vec, GloVe, and BERT, and apply them to a text classification task.We aim to compare the performance of these embeddings using metrics such as accuracy, precision, recall, and F1-score and provide a detailed analysis of their strengths and weaknesses.

Development Process

1. **Text Embeddings Selection**
    - Chose three popular text embeddings: **Word2Vec**, **GloVe**, and **BERT**.
    - Each embedding represents different approaches to word representation:
        - **Word2Vec**: Context-independent, dense vectors.
        - **GloVe**: Pre-trained, context-independent vectors with global co-occurrence statistics.
        - **BERT**: Contextual embeddings using transformers.
2. **Dataset Selection and Preparation**
    - Selected the **IMDb Movie Reviews** dataset for a sentiment analysis task.
    - Preprocessed the dataset: tokenization, stop-word removal, and text normalization.
3. **Applying Text Embeddings**
    - **Word2Vec**: Used pre-trained embeddings from Google News.
    - **GloVe**: Loaded pre-trained GloVe embeddings with 100-dimensional vectors.
    - **BERT**: Utilized the `bert-base-uncased` model from Hugging Face Transformers to obtain contextual embeddings.
4. **Text Classification Task**
    - Trained a Random Forest classifier using the different embeddings.
    - Evaluated performance using a standard train-test split (80/20).
5. **Performance Evaluation Criteria**
    - Measured accuracy, precision, recall, and F1-score for each embedding.

Results

Embedding	Accuracy	Precision	Recall	F1-Score
Word2Vec	0.83	0.82	0.81	0.81
GloVe	0.85	0.84	0.84	0.83
BERT	0.91	0.90	0.90	0.90