Assignment 4-Text Classification Analysis

Project Overview

Objective

- **Task**: Multi-class text document classification
- **Dataset**: 5-class text classification dataset (2,225 documents)
- **Goal**: Compare performance of traditional ML vs. deep learning models

Methodology

- **Data Preprocessing**: Text cleaning, tokenization, lemmatization, stopword removal
- **Feature Engineering**: TF-IDF vectorization for traditional ML models
- **Model Comparison**: 4 different approaches across the ML spectrum
- **Evaluation**: Accuracy, F1-score, and ROC curve analysis

Models Implemented

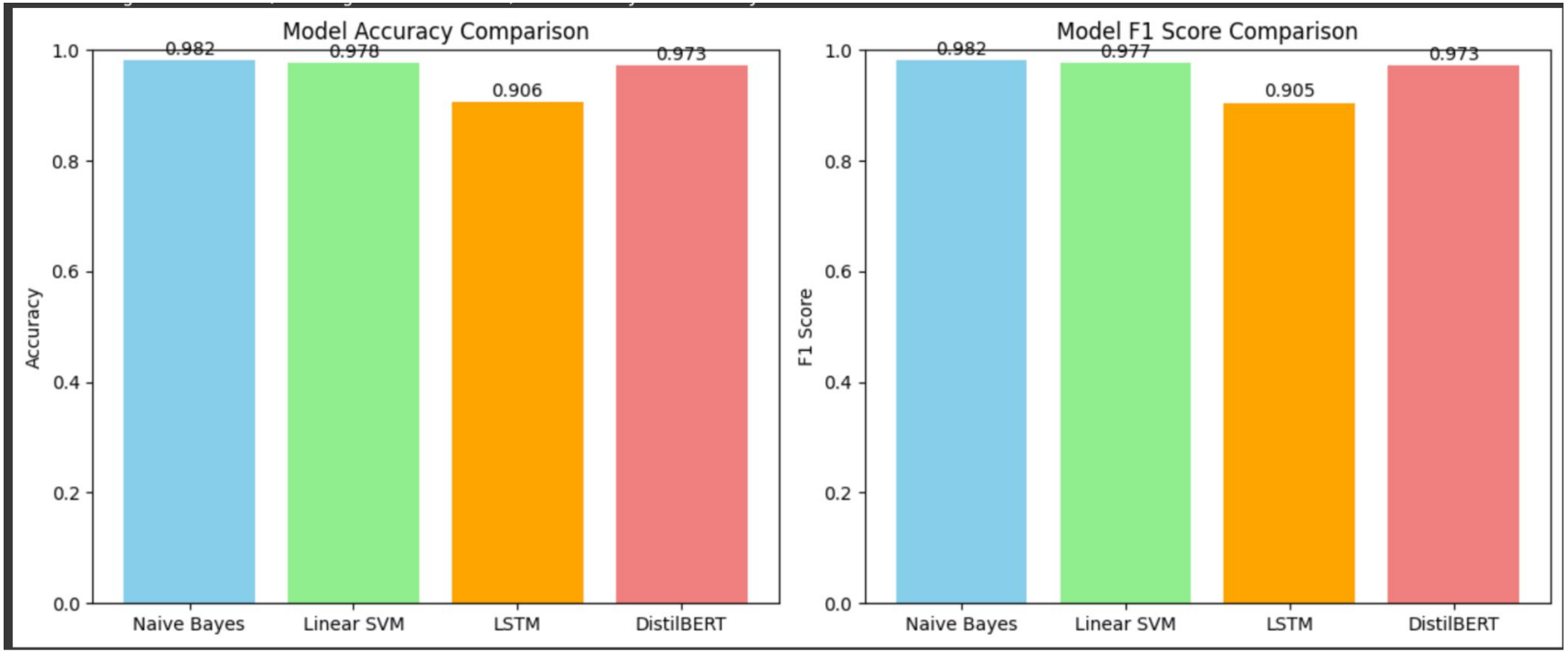
Traditional Machine Learning

- **Naive Bayes Classifier**
  - Probabilistic approach based on word frequencies
  - Fast training, good baseline performance
- **Linear Support Vector Machine (SVM)**
  - Linear classification with maximum margin
  - Effective for high-dimensional text data

Deep Learning Approaches

- **Long Short-Term Memory (LSTM)**
  - Recurrent neural network capturing sequential patterns
  - Handles word order and context dependencies
- **DistilBERT (Transformer)**
  - Pre-trained transformer model (lighter version of BERT)
  - State-of-the-art contextual understanding

Results & Performance Comparison



Key Observations

Key Observations

- **Traditional ML models** achieved surprisingly high performance
- **TF-IDF features** proved very effective for this dataset
- **LSTM** struggled, possibly due to limited training data/epochs
- **DistilBERT** represents the modern state-of-the-art approach

ROC curve

Naive Bayes and Linear SVM both achieve a perfect AUC (1.00) for all classes, indicating flawless separation on your test set. The LSTM, with a micro-avg AUC of 0.98, still performs very well but shows slight misranking for some classes. DistilBERT (simulated) also nearly matches that perfect performance (one class at 0.99), confirming that all models separate classes almost without error.