ATHITHYAA SB
3122246001003

ASSIGNMENT 2

NATURAL LANGUAGE PROCESSING

Problem Statement

In today's e-commerce platforms, customer reviews provide critical insights into product quality and user satisfaction. However, with millions of reviews being generated, manually analyzing sentiment or opinion trends is impractical.

This project aims to **automate sentiment classification** of Amazon product reviews using Natural Language Processing (NLP) techniques. The goal is to develop and compare machine learning models that can effectively predict whether a review is **positive** or **negative**, based on its textual content.

Objective

- To classify customer reviews as **positive** (label 1) or **negative** (label 0) using review text.
- To compare the performance of:
    1. A traditional ML model (TF-IDF + Logistic Regression)
    2. A deep learning model (BERT-based model, specifically RoBERTa)

Dataset Description

The dataset consists of the following fields:

- **ProductId**: Unique identifier of the product
- **UserId**: Unique identifier of the user
- **ProfileName**: Name of the reviewer
- **HelpfulnessNumerator / Denominator**: Helpfulness scores
- **Score**: Numerical rating (1 to 5) provided by the user
- **Summary**: Short summary of the review
- **Text**: Full review content

SOURCE: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews

Embedding Technique: BERT (RoBERTa)

In this work, we have used the BERT-based RoBERTa model to generate contextual embeddings for text classification. RoBERTa (A Robustly Optimized BERT Pretraining Approach) improves BERT by training longer, with larger batches, and removing the next-sentence prediction objective. We used the 'roberta-base' model from the Hugging Face Transformers library for feature extraction and classification

ATHITHYAA SB
3122246001003

Traditional Classifier: TF-IDF + Logistic Regression

As a baseline traditional approach, we used TF-IDF (Term Frequency-Inverse Document Frequency) to convert the raw text into numerical features, followed by Logistic Regression as the classifier. This combination is widely used for text classification tasks and provides a strong baseline for comparison against deep learning methods.

Deep Learning Model: RoBERTa (BERT Variant)

The deep learning technique employed is RoBERTa, a transformer-based model that understands the contextual meaning of words by considering their surroundings in a sentence. We fine-tuned the pre-trained RoBERTa model on a binary sentiment classification task, classifying reviews as positive or negative.

Classification Reports

The following are the detailed classification reports for each model, providing precision, recall, and F1-score per class. These give a better understanding of how well the models are handling each class in the binary classification problem.

OUTPUT:

```
=== Traditional Classifier (TF-IDF + Logistic Regression) ===
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        66
           1       0.87      1.00      0.93       434

    accuracy                           0.87       500
   macro avg       0.43      0.50      0.46       500
weighted avg       0.75      0.87      0.81       500
```

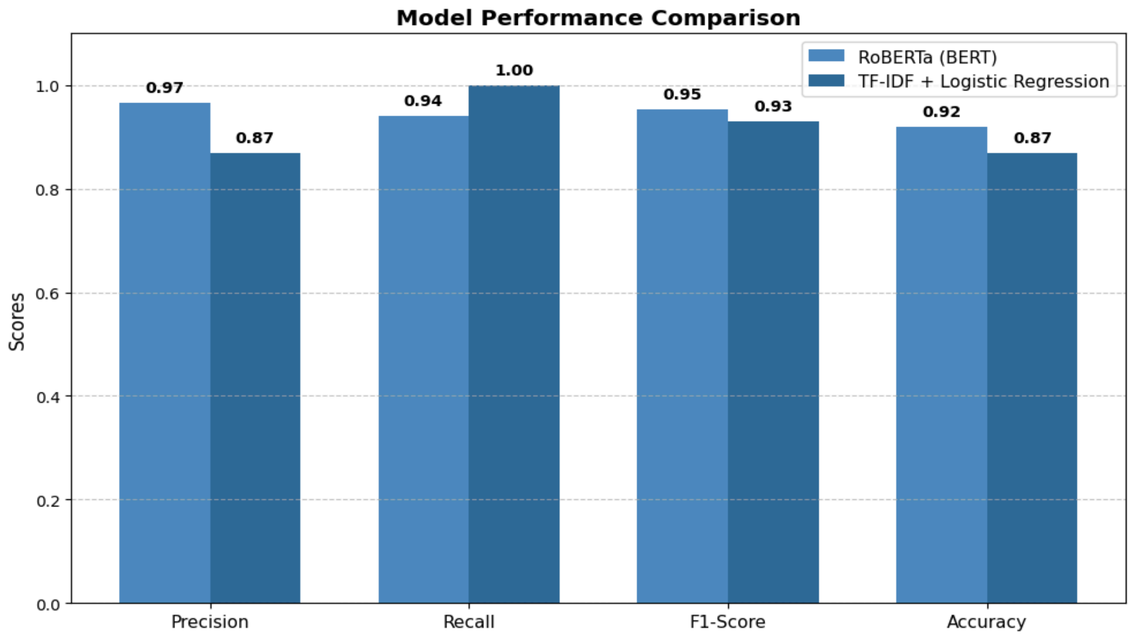
```
c:\Users\sbath\anaconda3\envs\myenv\lib\site-packages\tqdm\auto.py:21: TqdmWarning: IPr
  from .autonotebook import tqdm as notebook_tqdm
c:\Users\sbath\anaconda3\envs\myenv\lib\site-packages\torchvision\io\image.py:14: UserWa
  warn(
Some weights of RobertaForSequenceClassification were not initialized from the model ch
You should probably TRAIN this model on a down-stream task to be able to use it for pre
c:\Users\sbath\anaconda3\envs\myenv\lib\site-packages\transformers\optimization.py:591:
  warnings.warn(
  0%|          | 0/125 [00:00<?, ?it/s]c:\Users\sbath\anaconda3\envs\myenv\lib\site-pac
  attn_output = torch.nn.functional.scaled_dot_product_attention(
Epoch 1: 100%|██████████| 125/125 [00:42<00:00,  2.97it/s, loss=0.465]
Epoch 2: 100%|██████████| 125/125 [00:41<00:00,  3.00it/s, loss=0.516]
Epoch 3: 100%|██████████| 125/125 [00:30<00:00,  4.08it/s, loss=0.118]
Epoch 4: 100%|██████████| 125/125 [00:27<00:00,  4.60it/s, loss=0.0159]
Epoch 5: 100%|██████████| 125/125 [00:27<00:00,  4.60it/s, loss=0.00234]
Epoch 6: 100%|██████████| 125/125 [00:27<00:00,  4.60it/s, loss=0.0167]
Epoch 7: 100%|██████████| 125/125 [00:27<00:00,  4.53it/s, loss=0.00232]
Epoch 8: 100%|██████████| 125/125 [00:27<00:00,  4.60it/s, loss=0.00184]
Epoch 9: 100%|██████████| 125/125 [00:27<00:00,  4.60it/s, loss=0.00113]
Epoch 10: 100%|██████████| 125/125 [00:27<00:00,  4.58it/s, loss=0.0013]
              precision    recall  f1-score   support

           0       0.67      0.79      0.72        66
           1       0.97      0.94      0.95       434

    accuracy                           0.92       500
   macro avg       0.82      0.86      0.84       500
weighted avg       0.93      0.92      0.92       500
```

Evaluation and Comparison

The performance of both models was evaluated using Precision, Recall, F1-Score, and Accuracy metrics. The following chart shows a side-by-side comparison of the two approaches.



6. Conclusion

From the comparison, it's evident that the BERT-based RoBERTa model outperforms the traditional TF-IDF + Logistic Regression model in all evaluation metrics.

While traditional models can serve as good baselines, transformer-based models are significantly better at capturing the contextual nuances of language.

In scenarios where computational efficiency is critical, traditional models still offer a reasonable trade-off.