Project: HP Printer-Product Reviews Classification

Problem Statement:

The goal was to classify HP printer reviews into Positive, Neutral, and Negative categories to provide insights for product improvement, marketing strategies, and customer service.

Approach:

Data Preparation:

- Loaded HP printer reviews, cleaned column names, and removed duplicates.
- Preprocessed text by removing punctuation, tokenizing, removing stopwords, and lemmatizing with POS tagging.
- Augmented the original 42 records to 5,000 for better model training.

Exploratory Analysis:

- Analyzed word frequency to identify common terms.
- Visualized review lengths to understand typical review structure.

Initial Rule-Based Sentiment Analysis:

- Implemented a basic rule-based approach using positive/negative word lists and negation handling.
- Provided a baseline and labeled the dataset for model training.

FastText & RNN Models:

- Generated FastText embeddings and trained LSTM, GRU, and BiLSTM models.
- Applied Borderline-SMOTE to address class imbalance.
- Performance was reasonable but struggled with subtle sentiment distinctions.

BERT-Based Sentiment Classification:

- Fine-tuned a pre-trained BERT model for sentiment classification.
- Used stratified splits, custom PyTorch DataLoaders, and class-weighted loss to handle imbalance.

- Achieved significantly higher accuracy and robustness than FastText models, evaluated with classification reports and AUC-ROC.

Insights:

- Rule-based analysis provided a rough baseline but lacked accuracy.
- FastText with RNNs improved performance but had limitations in capturing nuanced sentiments.
- BERT captured subtle sentiment patterns effectively, showing strong precision, recall, F1-scores, and AUC-ROC.
- Challenges:
- Small original dataset required data augmentation.
- Class imbalance needed careful handling with Borderline-SMOTE and class-weighted loss.

Challenges:

- Tiny original dataset required augmentation.
- Rule-based methods couldn’t handle complex language or sarcasm.
- Managing class imbalance was important for training effective models.

Conclusion:

The project successfully built a robust BERT-based sentiment analysis model for HP printer reviews, accurately classifying customer feedback and providing actionable insights for product and service improvements.