Project Report: Intent Classification Using ML and BERT on CLINC150 Dataset

Objective

The primary goal of this project is to build an intent classification model using the **CLINC150 dataset**, which contains user utterances labeled with intent classes. The project explores both traditional machine learning (Random Forest, SVM) and a deep learning-based approach (BERT) for performance comparison.

Dataset Description

The dataset is a JSON structure that includes:

- train, val, and test splits for in-scope (IS) data.
- oos_train, oos_val, and oos_test splits for out-of-scope (OOS) data.
- Each data point contains:
    - text: the user query.
    - intent: the corresponding intent label.

The data is loaded into Pandas DataFrames and combined into comprehensive train, validation, and test sets.

Data Preprocessing

1. **Label Encoding**:
    - Intent labels are encoded using LabelEncoder to convert them into numerical classes.
2. **Vectorization**:
    - TF-IDF vectorization with unigrams and bigrams is applied to transform the text into numerical feature vectors.

Machine Learning Models

Random Forest Classifier

- Utilized with 200 estimators.

- Parameters include balanced class weights, min_samples_split=5.
- Achieved 72% accuracy and classification report results on the test set.

Support Vector Machine (SVM)

- Used as a second traditional model.
- Applied on the same TF-IDF vectorized features.
- Performance is also evaluated via accuracy (81% achieved) and classification metrics.

Deep Learning Model

BERT-based Classification

- Transformers library is used to fine-tune a BERT model.
- Steps include:
    - Tokenization using BERT tokenizer.
    - Dataset preparation with attention masks and proper formatting.
    - Training with AdamW optimizer and learning rate scheduling.
    - Evaluation using classification metrics and confusion matrix.

Evaluation Metrics

- Accuracy Score 9 (84% achieved)
- Classification Report (Precision, Recall, F1-score)
- Confusion Matrix (visualized with Seaborn)

Qualitative Error Analysis

- Misclassified samples are analyzed to understand model weaknesses.
- Helps identify areas for potential improvement like better context handling, longer training, or data augmentation.

Conclusion

- Both traditional ML models and BERT showed strong performance on intent classification.
- BERT generally outperformed SVM and Random Forest due to its ability to capture deeper contextual information.
- This project demonstrates the viability of using both approaches, but highlights BERT's superiority for complex language understanding tasks.
