Milestone 1 Report

BankBot AI Chatbot for Banking FAQs

Name: Yuvaraj Gowda D K

Batch: 5

Objective:

- **Intent Classification**: Classifying user queries into predefined intents.
- **Entity Extraction**: Extracting key entities from the queries (e.g, account numbers, dates).
- **Response Generation**: Providing a predefined response based on the predicted intent.

Tasks Completed:

1. **Dataset Loading and Preprocessing**:

    - Loaded a CSV file containing columns: text (query), intent (predicted intent), response (response for the intent), and entities (optional).
    - Checked for missing required columns (text, intent, response) and ensured the dataset was complete.

2. **Entity Parsing**:

    - Implemented a function (safe_parse_entities) to handle the entities column, ensuring safe parsing of different formats. This function handles missing or malformed entities gracefully.

3. **Intent Classification**:

    - Prepared the data for training by splitting it into X (input text) and y (intent label).
    - Used **TfidfVectorizer** for converting text into numerical features (tf-idf scores).
    - Applied a **Logistic Regression** model to classify the intent based on the vectorized text.
    - The classifier was trained using 80% of the dataset (train-test split) and tested on the remaining 20% to evaluate its performance.

4. **Model Evaluation**:

    - Evaluated the classifier's performance using the **classification report**, which provides metrics such as precision, recall, and F1-score for each intent class.

```
=== Classification Report ===
                precision    recall  f1-score   support

    block_card       0.00      0.00      0.00         1
branch_locator       0.00      0.00      0.00         1
  card_inquiry       0.00      0.00      0.00         1
 check_balance       1.00      1.00      1.00         2
       goodbye       1.00      1.00      1.00         1
  loan_inquiry       1.00      1.00      1.00         2
  out_of_scope       0.00      0.00      0.00         1
transfer_money       0.20      1.00      0.33         1

      accuracy                           0.60        10
     macro avg       0.40      0.50      0.42        10
  weighted avg       0.52      0.60      0.53        10
```

5. **Entity Extraction**:

    - Integrated **spaCy** for Named Entity Recognition (NER) to extract entities from the text.
    - A custom rule was added to extract numeric tokens as account_number, assuming account numbers are typically numeric.
    - Extracted entities are stored in a dictionary where the key is the entity type (e.g., account_number, ORG, GPE), and the value is the entity's text.

6. **Response Retrieval**:

    - A function (get_response) was implemented to retrieve the corresponding response from the dataset based on the predicted intent.
    - The system returns a predefined response when a valid intent is predicted; if no response is found for the predicted intent, a default message is returned.

Tools and Libraries Used

- **Pandas**: For handling the dataset and performing data cleaning.
- **spaCy**: For Named Entity Recognition (NER) and custom entity extraction.
- **Scikit-learn**: For the machine learning pipeline, including:

    - **TfidfVectorizer** for text vectorization.
    - **Logistic Regression** for intent classification.
    - **train_test_split** for splitting data into training and testing sets.
    - **classification_report** for model evaluation.

Conclusion:

Milestone 1 provided a solid foundation for building a conversational AI system that classifies user intents, extracts relevant entities, and generates appropriate responses. Key skills learned include data preprocessing, feature extraction using tf-idf, intent classification with Logistic Regression, and named entity extraction with spaCy. The system is functional but can be improved by tuning the model and expanding entity recognition capabilities.