Assignment 4

Name: Adhija Satish Bachhav

Roll No: 381053

PRN: 22311656

Problem Statement

To build a Named Entity Recognition (NER) system capable of extracting named entities from real-world textual data such as news articles or social media content, and to evaluate the system using standard performance metrics including accuracy, precision, recall, and F1-score.

Objective

The objective of this assignment is to understand how Named Entity Recognition works in Natural Language Processing and to implement an end-to-end NER pipeline using a pre-trained model. The assignment also focuses on evaluating the performance of the NER system using appropriate metrics.

Terminologies Used

1. Named Entity Recognition (NER)

Named Entity Recognition is an NLP technique used to identify and classify key information (entities) in text into predefined categories such as Person, Organization, Location, Date, etc.

2. Entity Types

- PERSON: Names of individuals
- ORG: Names of organizations
- GPE: Geopolitical entities such as countries or cities
- DATE: Dates or time-related expressions

3. Pre-trained Model

A pre-trained model is a machine learning model trained on large datasets beforehand. In this assignment, a spaCy English NER model is used to extract entities without training from scratch.

4. Gold Labels

Gold labels refer to manually annotated, ground-truth entity labels used for evaluating the performance of the NER system.

Evaluation Metrics

1. Accuracy

Accuracy measures the proportion of correctly predicted labels out of all predictions.

2. Precision

Precision indicates how many of the predicted entities are actually correct.

3. Recall

Recall measures how many of the actual entities were correctly identified by the model.

4. F1-Score

F1-score is the harmonic mean of precision and recall, providing a balanced measure of performance.

Methodology

The following steps were performed:

1. Real-world, news-style text data was collected.
2. A pre-trained spaCy NER model was loaded.
3. Named entities were extracted from the text.
4. Manually annotated gold labels were prepared.
5. Predicted labels were compared with gold labels.
6. Accuracy, precision, recall, and F1-score were calculated.

Output Description

- Extracted Entities: Named entities identified from each input sentence along with their entity types.
- Evaluation Scores: Numerical values representing accuracy, precision, recall, and F1-score of the NER system.

Applications

- Information extraction from news articles
- Social media analysis
- Resume parsing

- Question answering systems
- Search and recommendation systems

Conclusion

The Named Entity Recognition system successfully extracted meaningful entities from real-world text using a pre-trained NLP model. The evaluation metrics demonstrate the effectiveness of the system and highlight the importance of proper entity extraction in downstream NLP tasks.

Result

A functional Named Entity Recognition system was implemented, and its performance was evaluated using standard classification metrics. The objectives of the assignment were achieved successfully.