# Advancing Named Entity Recognition with Open-Source Data.

**Introduction**

Named Entity Recognition (NER) is a fundamental task in natural language processing (NLP) that involves identifying and classifying entities such as names, locations, dates, and organizations within text. NER has numerous applications, including information retrieval, question answering, and knowledge graph construction.

**Objective**

The primary objective of this project is to develop a high-performance NER system that leverages both traditional machine learning and modern deep learning approaches. The key goals include:

1. Extracting and preprocessing entity-rich text data from an open-source dataset.
2. Implementing and comparing multiple NER models.
3. Evaluating the impact of different feature engineering and model architectures on NER performance.
4. Providing insights into the challenges and advancements in NER.

**Dataset**

The project will utilize a widely recognized open-source dataset, such as:

* **CoNLL-2003:** Annotated for entities like Person, Organization, Location, and Miscellaneous.
* **WikiANN:** Multilingual NER dataset extracted from Wikipedia.

These datasets offer a diverse range of text types and entity classes, making them suitable for evaluating the generalizability of NER systems.

**Methodology**

1. **Data Preparation**
   * Annotating and pre-processing data as per NER requirements.
   * Splitting the dataset into training, validation, and test sets.
2. **Feature Engineering**
   * Utilizing word-level and character-level features.
   * Experimenting with embeddings like Word2Vec and contextualized embeddings (e.g., BERT, Flair).
3. **Model Development**

* More details on the method will be added as we progress through the course

1. **Evaluation**
   * Metrics: Precision, Recall, F1-Score.
   * Error analysis to identify frequent misclassifications.
2. **Visualization**
   * Highlighting extracted entities within the text.
   * Comparing model performance across different entity types.

**Expected Outcomes**

1. An end-to-end NER pipeline capable of extracting entities with high accuracy.
2. A comparative analysis of traditional and modern approaches to NER.
3. Insights into the advantages and limitations of contextualized embeddings for NER tasks.
4. Recommendations for deploying NER systems in real-world applications.

**Applications**

* Automating document processing in legal, healthcare, and finance industries.
* Enhancing search engines and chatbots with entity-aware features.
* Building knowledge graphs for semantic analysis.

**Tools and Technologies**

* Programming Language: Python
* Libraries and Frameworks: NLTK, TensorFlow, PyTorch, Hugging Face Transformers
* Data Visualization: Matplotlib, Seaborn

**References**

* Available Open-Source Datasets: [CoNLL-2003](https://www.clips.uantwerpen.be/conll2003/ner/), [WikiANN](https://github.com/afshinrahimi/mmner" \o "https://github.com/afshinrahimi/mmner).