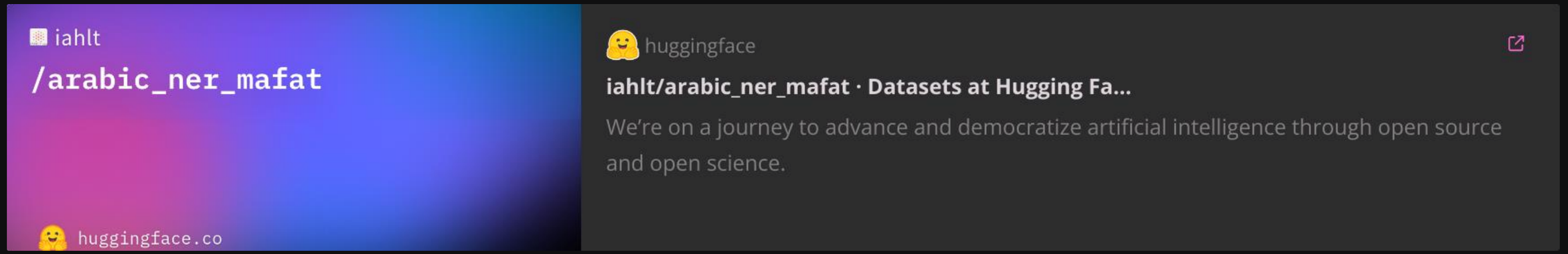
Building a Named Entity Recognition Model for Arabic

This presentation explores our development of a fine-tuned Named Entity Recognition (NER) model tailored for Arabic. Leveraging the Arabic NER MAFAT Dataset, we fine-tuned Asafya BERT base to accurately identify entities in Modern Standard Arabic texts. The model aims to extract valuable information such as person names, locations, and titles from raw Arabic text, driving meaningful insights in various NLP applications.

We will guide you through the dataset, preprocessing steps, model architecture, challenges, and evaluation outcomes to showcase the capabilities and limitations of this specialized NER system.

Dataset Overview: Arabic NER MAFAT

Dataset source:



Dataset Characteristics

The dataset has 40,000 training samples of Arabic text tokens labeled for Named Entity Recognition. It includes seven key columns capturing tokens, entity tags, and metadata.

Samples come from Modern Standard Arabic texts, with annotations for persons, locations, organizations, and more, enabling accurate entity identification.

Data Partitioning

The dataset is split into training (80%, 32,000 samples), validation (10%, 4,000 samples), and test (10%, 4,000 samples) sets.

This partitioning supports fine-tuning, hyperparameter tuning, and unbiased final evaluation.

Data Preprocessing Strategy

1

Column Selection

Kept only essential columns: tokens and their corresponding named entity labels, simplifying input data and reducing noise.

2

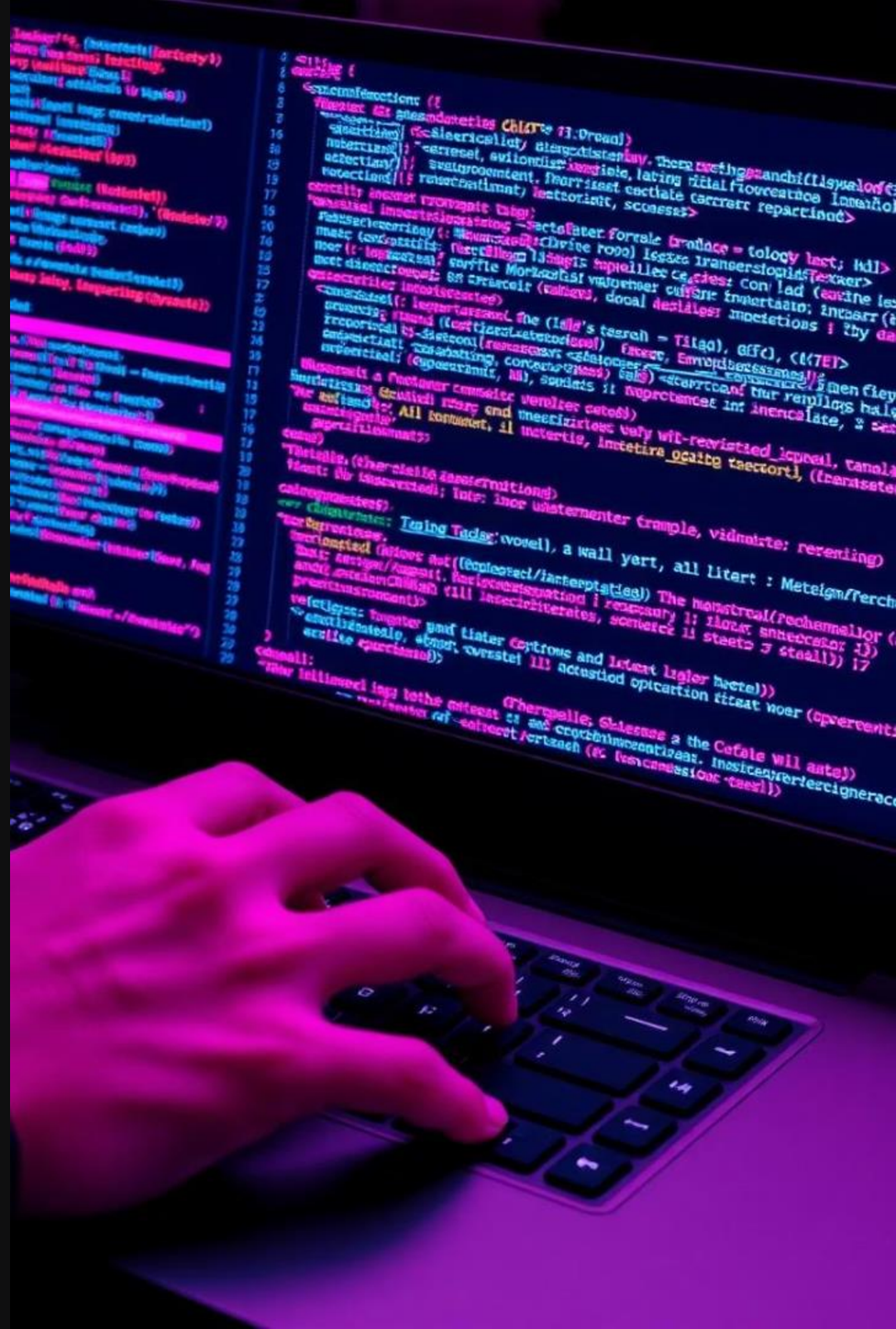
Tag Mapping

Created mappings between original tag names to numerical IDs and back, facilitating efficient model training and inference.

3

Unified Tagging Format

Replaced original textual labels with numerical tags to standardize annotations across the dataset.



Sample Data Visualization

Before Preprocessing

The dataset contains raw tokens alongside verbose tag sequences, multiple columns, and complex metadata.

After Preprocessing

```
{'tokens': ['خليل' ,'محمود' ,'البروفسور'], 'labels': [11, 9, 35]}
```

Data reduced to just tokens paired with numeric labels, streamlining input for training and boosting model efficiency.

Model Architecture: Asafya BERT Base

Asafya BERT Base

A transformer-based model pre-trained on large Arabic corpora, capturing modern standard Arabic linguistic features.

Fine-tuning for NER

Applied domain-specific fine-tuning on the MAFAT dataset, adapting the base model to recognize Arabic named entities with high precision.

Output Layer

Customized classification layer predicts entity tags for each token, enabling token-level entity recognition and sequence labeling.

Model Limitations and Challenges

Language Specificity

Model trained strictly on Modern Standard Arabic, limiting accuracy on diverse dialects and regional variations.

Resource Intensiveness

Fine-tuning requires powerful GPUs and extended computing time, which may limit scalability for very large datasets.

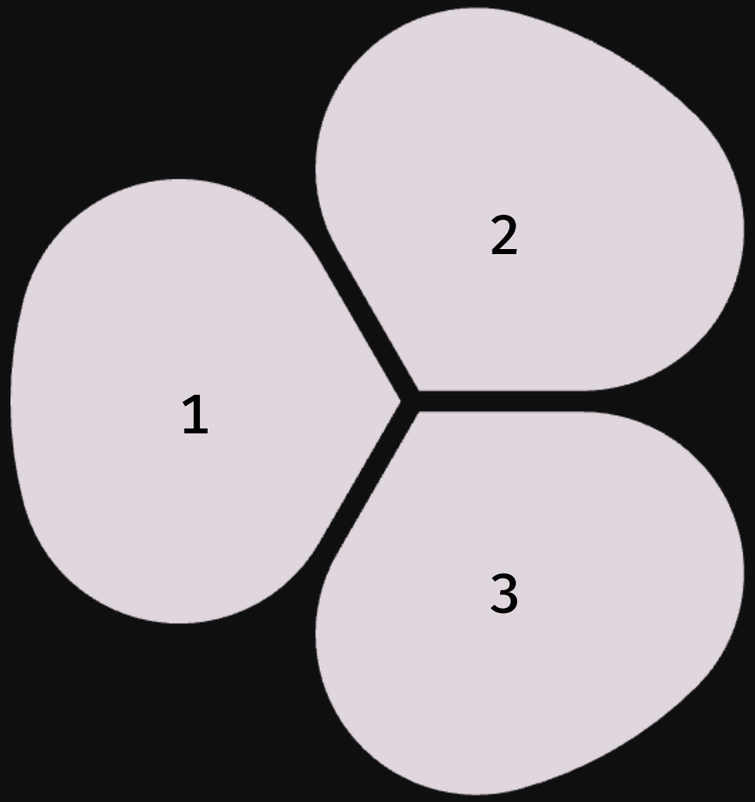
Data Imbalance

Entity class distribution varies, potentially impacting recognition accuracy for less frequent entity types.

Practical Use Case: Token-Level Entity Detection

Input Sentence

ولد العالم إسحاق نيوتن في إنجلترا



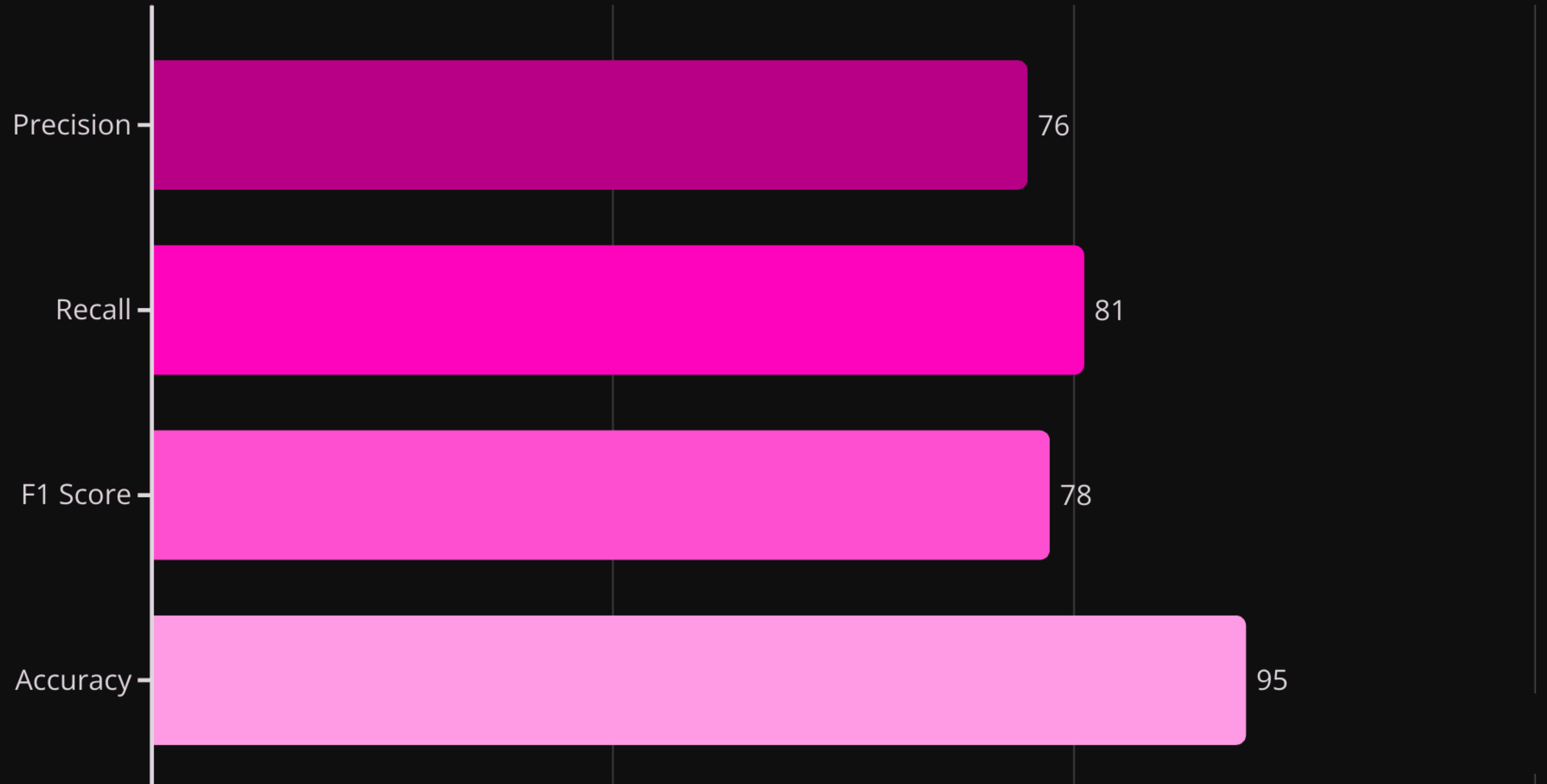
Entity Recognition

- "العالم" tagged as Title (B-TTL)
- "إسحاق نيوتن" tagged as Person entity (B-PER, L-PER)
- Others labeled as Outside (O)

Output Tokens

Tokens aligned with respective entity tags enabling clear entity distinction in text streams.

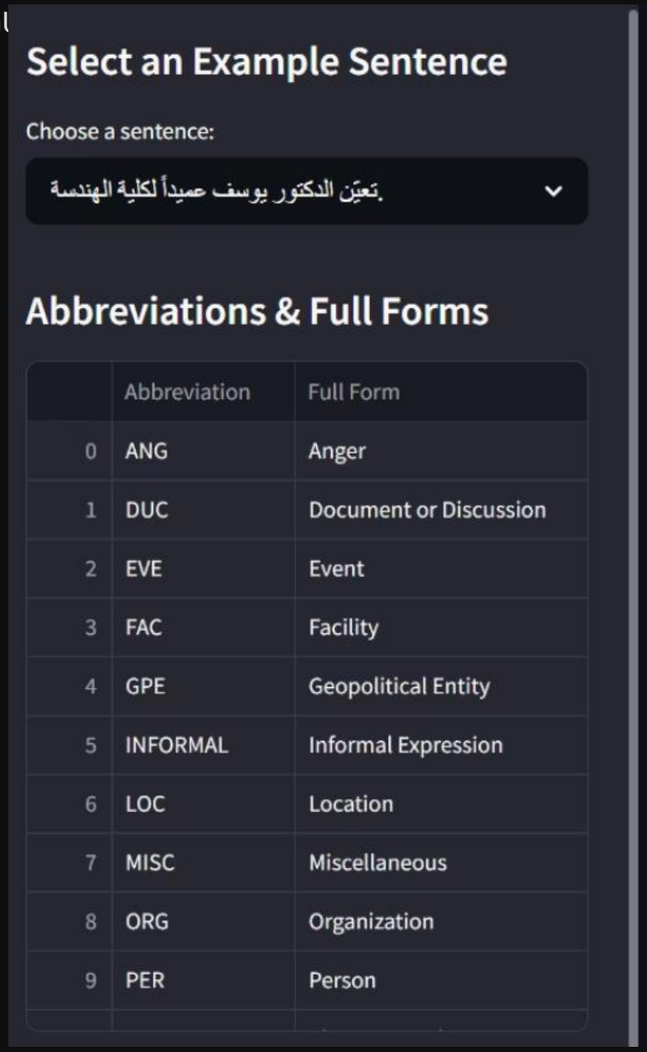
Model Evaluation Metrics



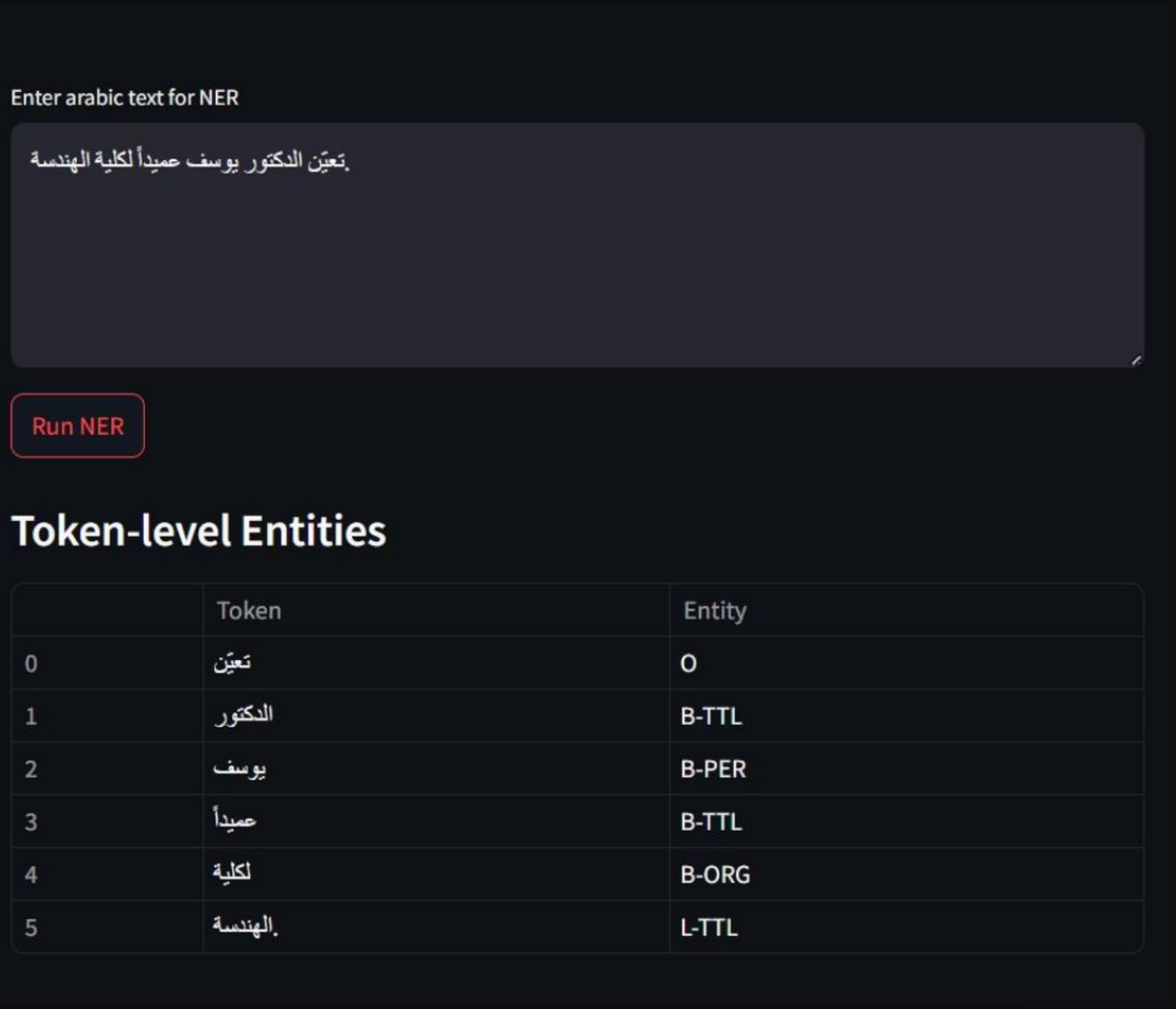
Model GUI Demonstration: Streamlit Interface

Enter a sentence in Arabic and see the detected named entities highlighted immediately.

The GU



	Abbreviation	Full Form
0	ANG	Anger
1	DUC	Document or Discussion
2	EVE	Event
3	FAC	Facility
4	GPE	Geopolitical Entity
5	INFORMAL	Informal Expression
6	LOC	Location
7	MISC	Miscellaneous
8	ORG	Organization
9	PER	Person



	Token	Entity
0	تعيّن	O
1	الدكتور	B-TTL
2	يوسف	B-PER
3	عميداً	B-TTL
4	لكلية	B-ORG
5	الهندسة.	L-TTL