Arabic Name Entity Recognition project

Introduction to the model:

Purpose: identifying names of entities in text.

Example: "الملك سلمان في الرياض" →-per"الملك",per"سلمان",o"فى",loc"الرياض"

Dataset information:

-**Dataset name**: ANERCORP (Arabic Named Entity Recognition Corpus)

-**Purpose**: Named Entity Recognition (NER) — identifying entities like names of persons, locations, and organizations in Arabic text.

-**Training samples** ~ 4000(80%)

-**Test samples** ~ 1000(20%)

-**Total Sentences**: ~5000 sentences

-**Total Words**: ~150,000 words

```
print("the total number of sentences: ",len(sentences))
[7] ✓ 0.0s
... the total number of sentences:  4876
```

```
data.info()
✓ 0.0s
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 148563 entries, 0 to 148562
Data columns (total 2 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   Word    148556 non-null  object
 1   Tag     148563 non-null  object
dtypes: object(2)
memory usage: 2.3+ MB
```

```
x_train, x_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

print("training samples count: ", len(x_train))
print("testing samples count: ",len(x_test))
✓ 0.0s
training samples count:  3900
testing samples count:  976
```

-**classes**:

1) PER: Person
2) LOC: Location
3) ORG: Organization
4) O: Miscellaneous

-**Annotation Format**:

-Each word is labeled with a corresponding tag using BIO tagging (e.g., B-PER, I-LOC, O)

-One word per line, with its tag

-Empty lines separate sentences

-**Link**: https://github.com/EmnamoR/Arabic-named-entity-recognition

---

Model parameter:

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 50, 64)	2,124,992
spatial_dropout1d (SpatialDropout1D)	(None, 50, 64)	0
bidirectional (Bidirectional)	(None, 50, 200)	132,000
time_distributed (TimeDistributed)	(None, 50, 9)	1,809

Training details:

-Optimizer: Adam
-Activation function: softmax
-Loss function: sparse_categorical_crossentropy
-Batch= 32
-Epoch= 10
-Accuracy= 97.3%

Model limitation:

Limitation	Description
Limited Entity Types	The model is restricted to only 3 entity types (PER, LOC, ORG).
Poor Generalization to Informal Texts	Performance drops on informal, dialectal, or noisy Arabic (e.g., tweets).
Vocabulary Dependency	Struggles with out-of-vocabulary (OOV) or unseen words during training.
No Handling of Nested Entities	Cannot recognize entities that are embedded within other entities.
Data Imbalance	Unequal distribution of entity types may bias the model's predictions.
Lack of Context Beyond Sentence	Model does not consider document-level or paragraph-level context.
Confusion Between Similar Entities	Mistakes often occur between similar tags (e.g., ORG vs. LOC).
Domain-Specific Bias	Performance is tuned to news domain (ANERCorp); may not generalize elsewhere.