Lab 4 Report

AYOUB FRIHAOUI G2

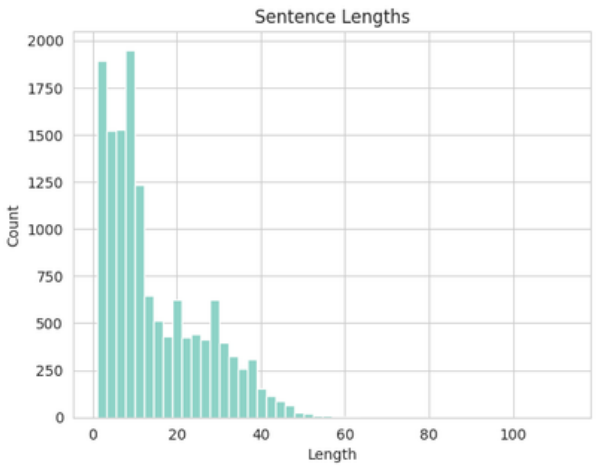
NLP LABS (CLICK) GITHUB

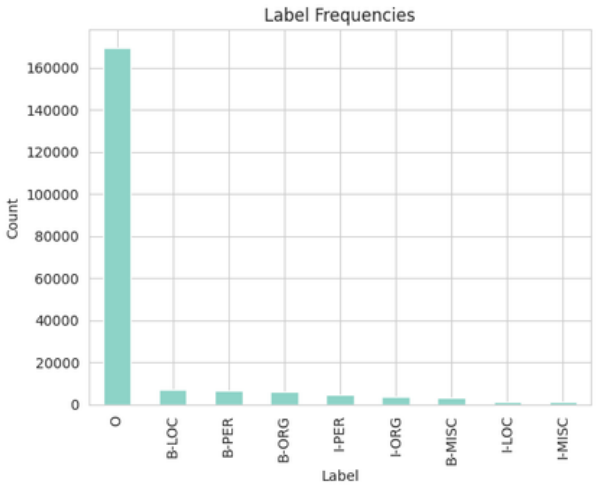
NATURAL LANGUAGE PROCESSING WITH SEQUENCE MODELS

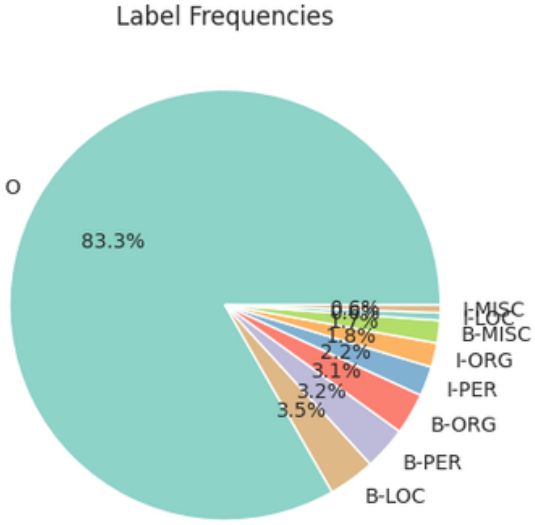
INTRODUCTION:

The data provided is from the CoNLL-2003 dataset, which is used for Named Entity Recognition (NER) tasks. The dataset consists of sentences annotated with named entities such as persons, organizations, locations, and miscellaneous entities. Each line in the dataset represents a token (word or punctuation) along with its Part-of-Speech (POS) tag, chunk tag, and named entity tag.

VISUALISATION:







METHODS:

Tokenize & add some padding to sentences/labels:

For tokenization, we used Tokenizer from Tensorflow Python lab also for the padding we used pad_sequences

Build & Train the Model:

```python
embedding_dim = 256
hidden_units = 100
num_classes = len(label_tokenizer.word_index) + 1
batch_size = 128
num_epochs = 20

# Création du modèle
model2 = Sequential()
model2.add(Embedding(input_dim=vocab_size, output_dim=embedding_dim, input_length=max_len))
model2.add(Bidirectional(LSTM(units=hidden_units, return_sequences=True, dropout=0.2)))
model2.add(BatchNormalization())
model2.add(Bidirectional(LSTM(units=64, return_sequences=True, dropout=0.2)))
model2.add(BatchNormalization())
model2.add(Dense(1024, activation='relu'))
model2.add(Dropout(0.2))
model2.add(Dense(num_classes, activation='softmax'))
```

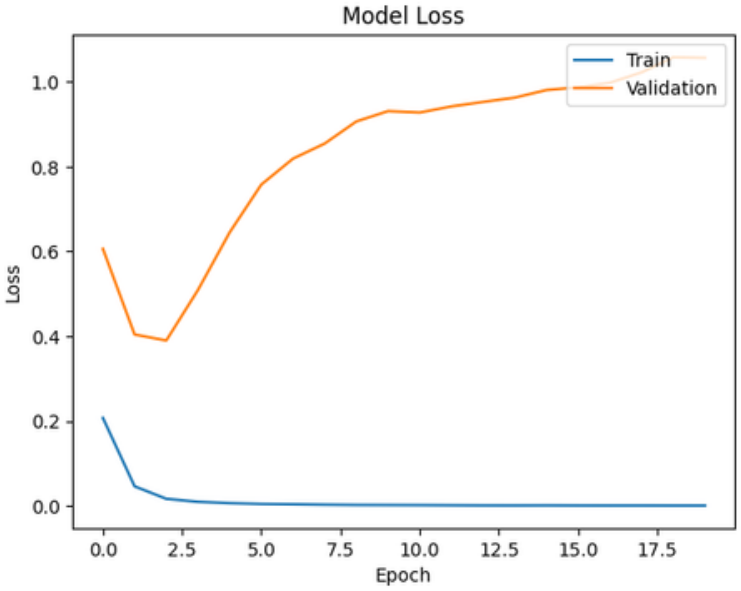
```
Model: "sequential_3"
```

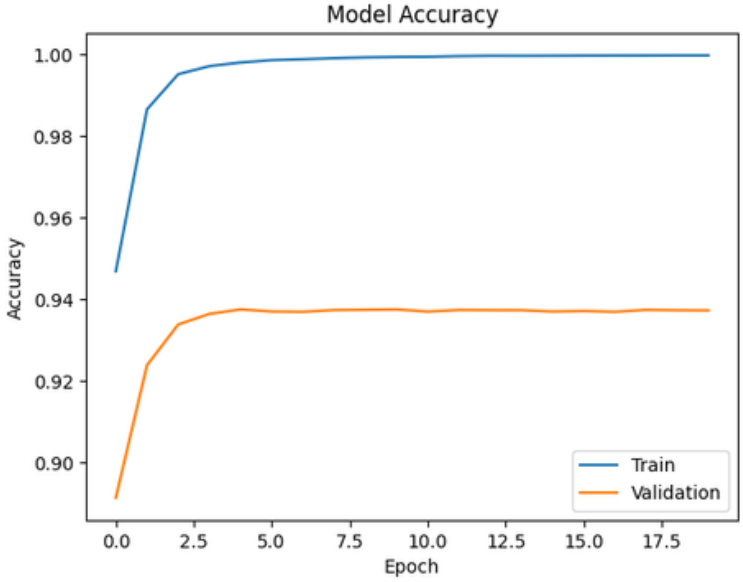
Layer (type)	Output Shape	Param #
embedding_3 (Embedding)	(None, 50, 256)	5,378,560
bidirectional_3 (Bidirectional)	(None, 50, 200)	285,600
batch_normalization (BatchNormalization)	(None, 50, 200)	800
bidirectional_4 (Bidirectional)	(None, 50, 128)	135,680
batch_normalization_1 (BatchNormalization)	(None, 50, 128)	512
dense_2 (Dense)	(None, 50, 1024)	132,096
dropout (Dropout)	(None, 50, 1024)	0
dense_3 (Dense)	(None, 50, 10)	10,250

```
Total params: 17,829,184 (68.01 MB)

Trainable params: 5,942,842 (22.67 MB)

Non-trainable params: 656 (2.56 KB)
```

RESULTS:

Regardless of using deeper model with higher embedding dimensionality, the model failed to preform any better than 93.75% on the validation data





TOKEN-LEVEL METRICS:
ACCURACY: 0.29
PRECISION: 0.33
RECALL: 0.29
F1-SCORE: 0.30