# Named Entity Recognition (NER) Model - Documentation

## 1. Overview

This document provides a detailed explanation of the Named Entity Recognition (NER) model built using BERT (bert-base-uncased). The model is trained to identify named entities such as persons, organizations, locations, and dates.

## 2. Dataset Structure

The dataset consists of two CSV files:

1. sentences.csv: Contains sentence descriptions mapped by sentence\_id.

2. tokens.csv: Contains tokens and their corresponding NER labels mapped by sentence\_id.

|  |  |  |
| --- | --- | --- |
| sentence\_id | token | tag |
| 1 | Elon | B-PER |
| 1 | Musk | I-PER |

## 3. Training Pipeline

The model is trained using the BERT tokenizer and classifier. The pipeline follows these steps:

1. Load and preprocess the dataset

2. Tokenize and align labels

3. Train BERT for token classification

4. Save the trained model

## 4. Testing and Evaluation

After training, the model is tested on unseen data. The predictions are compared against actual labels, and a classification report and confusion matrix are generated.

## 5. Sample Output

|  |  |  |
| --- | --- | --- |
| Token | Actual Tag | Predicted Tag |
| Elon | B-PER | B-PER |
| Musk | I-PER | I-PER |
| founded | O | O |

## 6. Conclusion

The NER model successfully identifies named entities and classifies them accurately. The performance is evaluated using precision, recall, and F1-score. The model can be further improved by fine-tuning on a larger dataset.