datasnipper

Dear Candidate,

Congratulations on progressing to the next step - the assessment!

Task: Build a Named Entity Recognition Classifier

Dataset: FiNER

Model: DistilBERT

Instructions:

1. Conduct a basic data analysis of the FiNER dataset to understand the entity distribution.
2. Select any four labels from the FiNER dataset. Then, compile a new dataset that includes only the necessary data for fine-tuning a model to detect these selected labels.
3. Preprocess the data, which includes tokenizing, padding, and encoding for both sentences and labels.
4. Fine-tune the DistilBERT model using your subset of the FiNER dataset.
5. Evaluate the fine-tuned model on the test set. Report the precision, recall, F1 score, and provide a confusion matrix.
6. Convert the NER model to ONNX format for interoperability.
7. Compare the ONNX model's results with those of the original DistilBERT-based model.
8. Document your project by writing a README file that describes the model, dataset, evaluation results, and usage for inference.

Expected Outputs:

- A Jupyter notebook or Python script with the code for the NER task, including comments and explanations.

- A trained model uploaded to the Hugging Face Hub, accompanied by a well-documented README file.
- A report summarizing the evaluation results, with insights and analysis comparing the ONNX model and the original model's performance.

Bonus Task:

1. Utilize the ML.NET API to load the exported model and create a data pipeline for pre-processing and inference.
2. Write a C# program that reads a text file, tokenizes it, and passes the data through the pipeline. The program should output the predicted named entities along with their corresponding labels for each sentence in the text file.

Expected Output:

- A C# Console application for data pre-processing and inference.

Good luck! And let us know if you have any questions!