REPORT

The success of fine-tuning Named Entity Recognition (NER) models is contingent on the quality and representativeness of the training data, hyperparameter tuning, and the choice of pre-trained model. In that regard, the MultiNERD Named Entity Recognition dataset is a high-quality dataset with more NER tags. The selection of a pre-trained model is pivotal in the NER task, where different architectures, including Bert-case, DistilBERT-cased, and XLNet-cased, play a crucial role. This choice is integral to identifying the most suitable base model for the task. In the context of NER, case information is crucial, and "cased" models, such as those provided by Bert and DistilBERT, are most effective.

Despite the promise of the XLNet architecture, which incorporates segment-level recurrence within the Transformer architecture by leveraging the Transformer-XL design—particularly adept at handling long-distance dependencies—the experiments demonstrated suboptimal metrics on the NER task. This anomaly could be attributed to the fact that named entities typically exhibit short spans of dependencies. In such cases, a vanilla Transformer architecture, as used by BERT, emerges as the most adept at accurately representing the intricacies of the NER task. Therefore, while advancements in transformer architectures like XLNet offer notable features, the task-specific nature of NER calls for a nuanced consideration of the architecture that aligns with the inherent characteristics of named entities and their dependencies.

During the fine-tuning of NER models, significant enhancements were observed in the accuracy of entity recognition, particularly when utilizing the DistilBERT-cased pre-trained model. This model demonstrated improved Precision and Recall, ultimately contributing to an overall advancement in the F1 score. Larger pre-trained models consistently correlated with superior performance; however, the need for prudent consideration of resource-constrained pre-trained models became evident. The DistilBERT-cased pre-trained model emerged as a standout performer during the fine-tuning process, showcasing its efficacy in capturing intricate patterns and nuances associated with named entities.

However, fine-tuning also presented challenges. Careful consideration of hyperparameters, including learning rates and batch sizes, was crucial to strike the right balance between model expressiveness and generalization. Unfortunately, due to constraints in computational resources and time limitations, exploring the impact of adjusting diverse hyperparameters during the fine-tuning of pre-trained Language Models (LLMs) could not be performed.

Evaluation Results

Experiment A:
Bert-cased: Precision: 0.902361, Recall: 0.921377, F1: 0.911770, Accuracy: 0.978552
XLNet-cased: Precision: 0.874651, Recall: 0.854772, F1: 0.864597, Accuracy: 0.974132
DIstilbert-cased: Precision: 0.902804, Recall: 0.920523, F1: 0.911577, Accuracy: 0.978789

Experiment B:
Bert-cased: Precision: 0.932510, Recall: 0.939178, F1: 0.935832, Accuracy: 0.985488
XLNet-cased: Precision: 0.839669, Recall: 0.756603, F1: 0.795975, Accuracy: 0.958987
DIstilbert-cased: Precision: 0.932882, Recall: 0.938054, F1: 0.935461, Accuracy: 0.985722