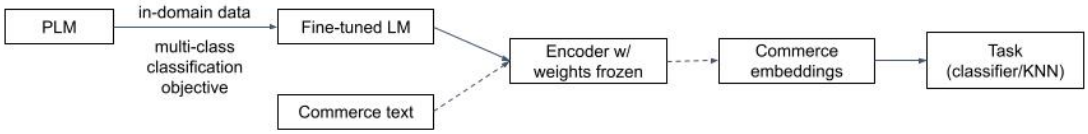
Fine-tuning multilingual XLM for E-Commerce Integrity domain

METHOD

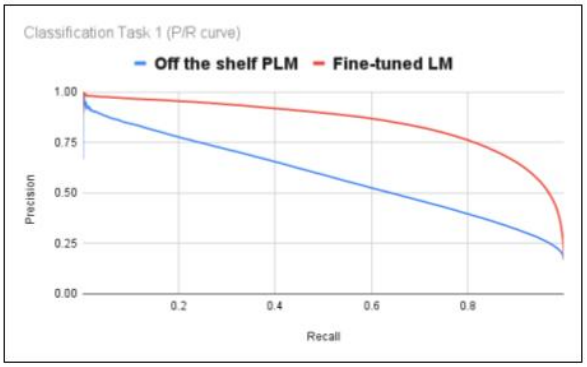
- E-Commerce data corpus of over 10 million documents in over 15 languages
- Pre-trained multilingual XLM model trained over Common Crawl is further fine-tuned
- In-domain (e-commerce integrity) fine-tuning is carried out as an exclusive multi-class classification objective
- [*CLS*] token of fine-tuned LM used for downstream binary classification tasks

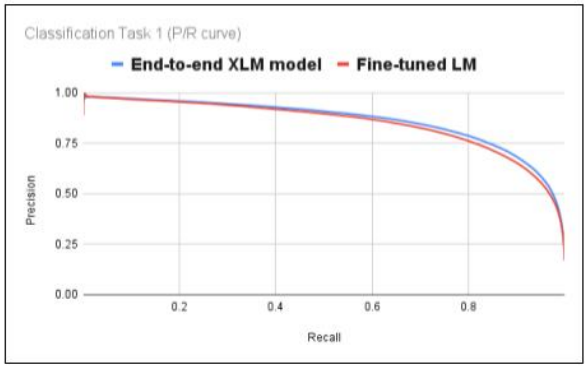


After fine-tuning the pre-trained language model with domain data, the encoder of the fine-tuned model is frozen. The model is used to generate embeddings for our e-commerce text which can then be re-used for multiple downstream tasks at no additional cost.

RESULTS

BINARY CLASSIFICATION TASK





Task/Model	ROC AUC	P/R AUC
Task 1, end-to-end	0.969	0.867
Task 1, fine-tuned	0.964	0.850
Task 2, end-to-end	0.988	0.949
Task 2, fine-tuned	0.982	0.923
Task 3, end-to-end	0.984	0.927
Task 3, fine-tuned	0.971	0.88

Table: Metrics across multiple tasks

COMPARING OFF-THE-SHELF PLM vs FINE-TUNED

A comparison of a pre-trained multilingual XLM language model (PLM) over common crawl data corpus versus the same model fine-tuned over in-domain (i.e. commerce integrity) data corpus shows a large gain in performance, illustrating the advantage of domain specific LM's for commerce. The PR AUC improved from 0.59 (for PLM) to 0.85 (for fine-tuned LM) for the binary classification task

.

COMPARING FINE-TUNED vs END-TO-END

The performance of the fine-tuned LM [CLS] embeddings (passed through a single classification layer) is at par with a multilingual XLM model trained end-to-end (e2e) using the classification task objective (i.e. learning encoder weights from scratch). The training run-time of the classification task using fine-tuned LM however was ~1/4X of training a model end-to-end (over the same architecture and same number of GPUs)

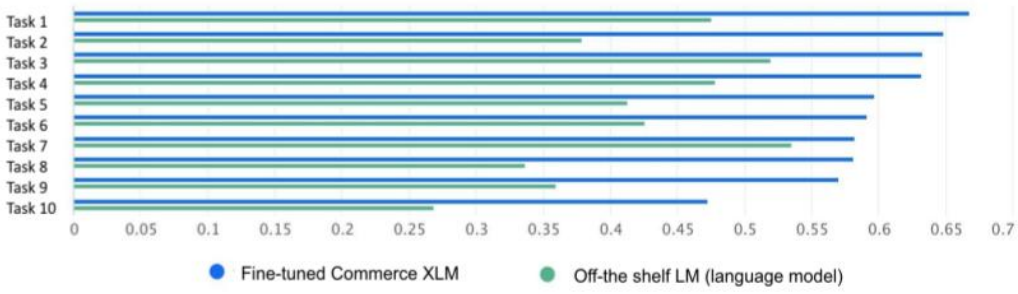
.

In fact as shown by the table, the AUC remains at par for multiple downstream classification tasks using the lightweight model (fine-tuned embeddings + single classification layer) and provides a less expensive modeling strategy for in-domain tasks

SIMILARITY DETECTION TASK

We also used our pre-trained embeddings for KNN based similarity detection tasks, and compare it with another off-the-shelf transformer based PLM.

Over 20 similarity detection tasks, we observed an average of 16% improvement in F1-metric for our model, as compared to the off-the-shelf model.



F1-score Comparison of Top-10 Tasks in Similarity Detection Evaluation

CONCLUSION

Our work in the e-commerce integrity domain shows the significant improvements one can have from using commerce specific multilingual dataset to fine-tune a pre-trained XLM model, and the versatility of the fine-tuned embeddings over different downstream tasks.