Fine-tuning multi-lingual XLM for Commerce Integrity domain

Bikash Mishra
bikashmishra@fb.com

Vahid Jalali
vjalali@fb.com

Yang Xie
yxie@fb.com

Talha Baig
tbaig6@fb.com

Leyton Ho
leytonho@fb.com

1 Background and Motivation

Commerce Integrity is important to have a safe and trustworthy ecosystem for buyers and sellers to conduct business with peace of mind in any e-commerce platform. Product text including title and description is critical source of information to detect whether a given product is safe or poses risks. Building expressive text representations in form of embeddings can provide a flexible and computationally cheap way of training machine learning models to detect different types of violations. Transformer based pre-trained language models (PLM) like BERT (Devlin et al., 2018) and XLM (Lample and Conneau, 2019) have had significant breakthroughs in many NLP tasks, the later especially in cross-lingual classification and machine translation. However these models are trained on generic datasets, and it has been shown that in-domain fine-tuning can improve model performance (Gururangan et al., 2020; Rietzler et al., 2019; Sun et al., 2019; Edwards et al., 2020) in most cases. In other cases, plain off-the-shelf PLM's are able to provide good performance (Sushil et al., 2021). The impact of domain knowledge and fine tuning of text embedding generator models remain a question for commerce integrity. In this work we address these questions and shed light onto them by introducing our work on building commerce text embeddings used for training integrity violation detection models for a C2C e-commerce platform.

2 Method

We develop a XLM based pre-trained language model for multi-lingual e-commerce integrity violation purposes which can be used for multiple downstream tasks (binary classification, similarity detection) on e-commerce listings. We compile a data corpus of over 10M listings, in 15+ languages, comprising of disjoint data points over all the classes from the binary classification tasks. We do not consider data points which might be shared between the class datasets. To fine-tune our XLM model over the e-commerce integrity violation dataset, we use a single dataset, single task approach. We train the model end-to-end on an exclusive multi-class classification task with a cross entropy loss function. We use SentencePiece tokenization with uncased tokens with a maximum sequence length of 256 tokens. Similar to fine-tuning strategies used before (Devlin et al., 2018; Lample and Conneau, 2019) we use a single linear layer on top of the first hidden state (the CLS token) as our classification layer. After training was completed the encoder weights were frozen, and the CLS embedding is used for downstream tasks.

3 Results

3.1 Comparing baseline XLM vs fine-tuned XLM

We compare an off the shelf pre-trained multilingual language model, against the e-commerce integrity specific model developed as described in the previous section, over downstream binary classification tasks, as well as for similarity detection tasks.

3.1.1 Binary Classification

The respective encoder weights are frozen for both the models to generate the CLS embedding, which is then passed into a single classification layer. Across the multiple binary classification tasks, we observe a 5-10% relative increase in PR AUC between the off the shelf model vs our e-commerce specific model. This shows that language model trained over generic multi-lingual data vary considerably from e-commerce integrity based domain language, and hence can be further improved by fine-tuning.

3.1.2 Similarity detection

We also used our pre-trained embeddings for a KNN based similarity detection task, and compared 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. This again illustrates the importance of domain specific language models for e-commerce integrity domain, as well as the versatility of the domain specific LM over different tasks.

3.2 Comparing fine-tuned XLM vs end-to-end trained XLM

We also compare the performance of our frozen embeddings (passed through a single classification layer) against end-to-end trained XLM model (unfreezing the encoder weights, and learning them together with the classification layer) for each binary classification task. We observed that both models had very similar performance (comparing PR-AUC) across most tasks. Experimental results show that end-to-end models suffer from overfitting compared to using our pre-trained model.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Aleksandra Edwards, Jose Camacho-Collados, Hélène De Ribaupierre, and Alun Preece. 2020. Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5522–5529, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *CoRR*, abs/2004.10964.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. *CoRR*, abs/1908.11860.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

Madhumita Sushil, Simon Suster, and Walter Daelemans. 2021. Are we there yet? exploring clinical domain knowledge of BERT models. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 41–53, Online. Association for Computational Linguistics.