Satisfying Real-Time Requirements in Multi-Label Text Classification of Traveler Feedbacks with Transformer Models

Dennis Imhof

Darmstadt University of Applied Sciences - Faculties of Mathematics and Natural Sciences & Computer Science

Motivation

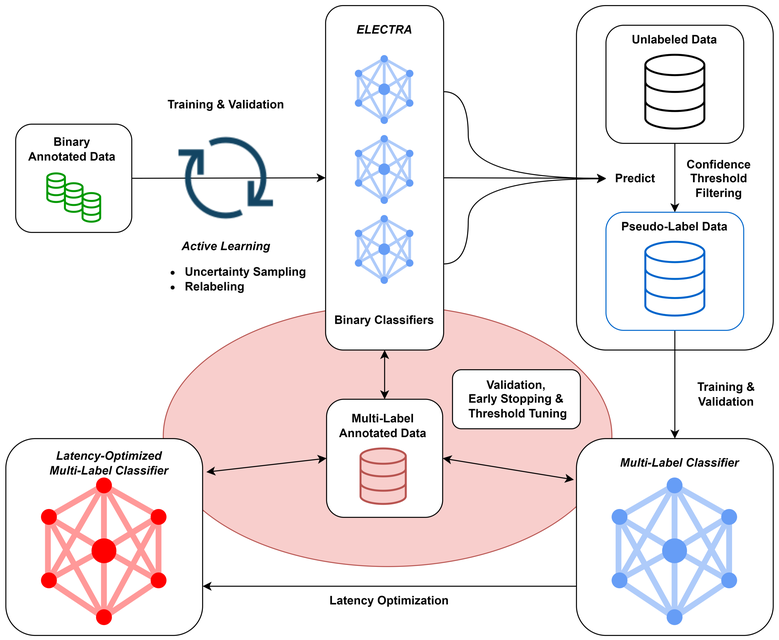
Anonymized textual customer feedbacks submitted in context of personal rail transport potentially cover a broad set of services. These services range from transport, booking and connection scheduling to information distribution and travel services. To derive immediate action from such customer feedbacks, for example by instantly responding to the customer with relevant information or by notifying service staff, a multi-class or multi-label detection system with high predictive quality and at the same time real-time latency is required. Additionally, for the training of a statistical model potentially satisfying these constraints, significant amounts of labeled data are needed. Creating multi-labeled data through human annotation, given constantly changing topic distributions, requires high amounts of human and financial resources. By using *large language model* binary classifiers to generate intermediary multi-label *Pseudo-Label* data, annotations can be performed on a per-topic basis, making the process more modular and scalable. The intermediary *Pseudo-Label* data is used to perform *Knowledge Distillation* in distilling the ensemble of binary *large language models* into a single multi-label model. Motivated by this, we investigate the feasibility of this process in collaboration with *DB Fernverkehr AG*.

Research Questions

1. Is it possible to train a performant multi-label model without expert-annotated multi-label data?
2. Can a multi-label language model outperform multiple binary language models and thus, can the multi-label model implicitly use label correlations to its advantage?
3. Can a high capacity multi-label language model reach an inference latency of sub 100ms on CPU hardware?

Methodology

We present a multi-step training and optimization scheme for real-time *multi-label classification* with *transformer*-based models. Due to the sparse availability of multi-label annotated data, we utilize an ensemble of *large language model* binary classifiers, specifically *ELECTRA* [1] and *DistilBERT* [4], trained on expert-annotated data to generate multi-label *pseudo-labels* using a large unlabeled dataset. We optimize the binary classifiers by iteratively applying *Pool-Based Uncertainty Sampling* [3] and thereby incrementally curating viable training datasets by means of *Active Learning*. We train student multi-label classifiers on the *Pseudo-Label* data that are further latency-optimized by application of neural network graph-optimization and quantization techniques. We show, that we can distill the knowledge of an ensemble teacher model into a latency-optimized student model with only a marginal loss of predictive power in a multi-label classification problem incorporating eleven classes. The conceptual steps of the proposed workflow are shown in figure 1. In a successive scalability experiment, we extend the selection of categories to 82 categories and investigate the performance as well as the category-specific performance degradation of the distilled multi-label large language model compared to the teacher ensemble.



Results

We find that the distillation of an ensemble of eleven binary *large language model* classifiers into a student model is possible with minor performance loss. For a subset of categories we even find an improvement in predictive performance, which we trace back to label correlations that the individual binary models of the teacher ensemble have no access to.

Category	Ensemble ELECTRA	Multi-Label ELECTRA	ML 8-Bit ELECTRA	Multi-Label DistilBERT	Multi-Label LSTM	Multi-Label XGBoost	SVM
Macro	**0.902**	0.899	0.888	0.891	0.860	0.842	0.629
Micro	**0.907**	0.905	0.894	0.901	0.874	0.853	0.681
Face Mask	**0.955**	0.943	0.920	0.927	0.917	0.945	0.644
Loudness	**0.943**	0.933	0.930	0.924	0.921	0.888	0.767
Luggage	0.822	0.812	0.797	**0.837**	0.724	0.775	0.376
Passenger Rights	0.792	**0.830**	0.805	0.789	0.774	0.667	0.305
Punctuality	0.906	0.934	0.939	**0.940**	0.924	0.920	0.805
Seat Availability	**0.897**	0.868	0.838	0.873	0.833	0.788	0.632
Seating Comfort	0.863	**0.870**	0.856	0.810	0.756	0.711	0.539
Temperature	0.936	**0.944**	0.937	0.942	0.938	0.912	0.745
Train Cancellation	**0.916**	0.879	0.877	0.857	0.825	0.807	0.571
Train Service	**0.906**	0.891	0.876	0.900	0.882	0.869	0.641
WLAN / Internet	0.990	0.987	0.990	**0.997**	0.961	0.984	0.892

Table 1 shows the F1-score evaluation results of different model architectures on a holdout test dataset. Most importantly, the left column displays the results of the ensemble of binary *ELECTRA* classifiers trained on expert-labeled data, whereas the remaining columns display results corresponding to models trained on pseudo-label data. In the successive latency optimization experiments we compare per-sample latencies of the model architectures on CPU. In a subsequent scalability experiment extending the category selection to 82 categories, we find that the compressed multi-label model shows competetive predictive performance on 75% of the included classes without further optimization of the employed *Pseudo-Label* dataset. In Figure 1 the latency measurements for unoptimized models (*PyTorch*), graph-optimized *ONNX*[a] models and 8-bit quantized *ONNX* models with an input sequence length of 250 tokens are shown.

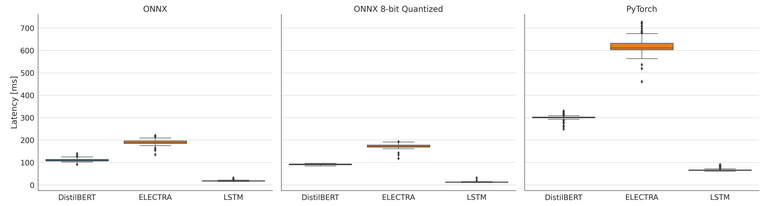


Figure 1. Latency

Combining optimized model graphs with the CPU-optimizations performed by the *ONNX Runtime* execution engine[b] results in latency reductions of up to 72% compared to the unoptimized *PyTorch* models. 8-bit quantization can then additionally reduce the latency by up to 31%, leading to a sub 100ms latency in case of *DistilBERT* and *LSTM* models.

Conclusion

The experimental results of this work show that even without expert-annotated multi-label data, the creation of performant multi-label models is possible through *Knowledge Distillation* with intermediary *Pseudo-Labels*. The resulting multi-label models achieve competitive predictive performance compared to the teacher ensemble model. Through the application of different latency optimization techniques like neural network graph-optimization and quantization, we can also show that a compressed and optimized multi-label model can reach real-time inference latencies below 100ms per-sample. An optimized multi-label model based on the findings of this work is currently live and serving a potential 200 million customers of *Deutsche Bahn*.

References

[1] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning.
Electra: Pre-training text encoders as discriminators rather than generators, 2020.

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean.
Distilling the knowledge in a neural network, 2015.

[3] David D. Lewis and William A. Gale.
A sequential algorithm for training text classifiers, 1994.

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf.
Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019.

[a] `https://onnx.ai` (last visited on 26.03.2023)
[b] `https://onnxruntime.ai/` (last visited on 26.03.2023)

dennis.imhof@stud.h-da.de