ABSTRACT

In this work we present a multi-step training and optimization scheme for real-time Multi-Label Classification (MLC) in the context of the long-distance personal rail transport industry. We are specifically dealing with anonymized traveler feedback texts submitted through multiple digital channels. To provide customers with context-dependent information, an instant response to their feedback has to be generated since anonymization prohibits the association of feedback and customer at a later point in time. Due to the sparse availability of multi-label annotated data, we utilize an ensemble of Large Language Model (LLM) binary classifiers trained on expert-annotated data to generate multi-label Pseudo Labels (PLs) for a large unlabeled dataset. A student LLM multi-label classifier is trained on the PL data and further latency-optimized through the application of neural network graph-optimization and quantization techniques.

We show, that we can distill the knowledge of an ensemble teacher model into a highly optimized student model with only a marginal loss of predictive power in a Multi-Label Classification problem incorporating eleven classes. We find only marginal performance degradation in the optimized student model, with the teacher model reaching a macro F1-score of 0.902 and the latency-optimized student model reaching a macro F1-score of 0.891 on a multi-label holdout testset. Contrastingly, the per-sample inference time of the student model can be reduced to 92ms on a commodity CPU, whereas the per-sample inference time of the ensemble model teacher depends on the slowest model in the ensemble and the degree of parallelization with around 500-700ms per-sample latency. Additionally, it can be shown that the multi-label student model even outperforms the binary ensemble teacher for some of the classes by utilizing learned label correlations the binary predictors have no access to. In a subsequent scalability experiment we extend the proposed workflow to a category selection of 82 categories and find that the student model reaches competitive performance on over 75% of the selected categories.

Keywords – Multi-Label Text Classification; Pseudo Labels; Latency; Transformer; ELECTRA; DistilBERT.

ZUSAMMENFASSUNG

In dieser Arbeit wird ein mehrstufiges Trainings- und Optimierungskonzept für Multilabel-Klassifikation in Echtzeit im Rahmen des Schienenpersonenfernverkehrs vorgestellt. Als Datengrundlage dienen anonymisierte Reisendenfeedbacks, die über verschiedene Digitalkanäle von KundInnen bereitgestellt werden. Die Klassifikation und die davon abgeleitete Antwort auf das Kundenfeedback muss in Echtzeit erfolgen, da durch die Anonymisierung der Feedbacks eine spätere Zuordnung und Antwort nicht mehr möglich ist. Da multi-label-annotierte Daten kaum verfügbar sind und sich der korrespondierende Annotationsprozess bei hohen Klassenanzahlen als enorm zeit- und ressourcenintensiv herausstellt, greifen wir auf ein Ensemble aus binären LLM Klassifikatoren zurück, die auf Basis von expertenannotierten Daten trainiert werden. Jenes Ensemble dient zur Erzeugung von PLs auf einem großen, ungelabelten Datenbestand. Anschließend wird ein multi-label *student model* mithilfe der PL-Daten trainiert und mithilfe von Netzwerkgraphoptimierung sowie Quantisierung latenzoptimiert.

Die Ergebnisse dieser Arbeit zeigen, dass ein Ensemble aus elf binären LLM Klassifikatoren ohne merklichen prädiktiven Performanzverlust in ein latenzoptimiertes *student model* distilliert werden kann. Hierbei weist das Ensemble ein makro F1-Evaluationsergebnis von 0.902 und das *student model* einen minimal geringeren Wert von 0.891 auf einem vorgehaltenen Testdatensatz auf. Gleichzeitig kann die per-Sample Inferenzzeit des *student models* auf gewöhnlicher Konsumentenhardware von den 500-700ms Inferenzzeit des Ensembles auf 92ms reduziert werden. Weiterhin zeigt sich, dass das multilabel *student model* auf einer Teilmenge der Labels eine bessere prädiktive Performanz erzielt als das Ensemblemodell, was auf die Ausnutzung von Labelkorrelationen zurückzuführen ist, die dem binären Ensemble nicht zur Verfügung stehen. In einem anschließenden Skalierbarkeitsexperiment wird die Kategorieauswahl auf insgesamt 82 Kategorien erweitert und gezeigt, dass durch das *student model* ohne weitere Optimierung der Trainingsdaten eine mit dem Ensemblemodell vergleichbare prädiktive Performanz auf über 75% der Klassen erreicht werden kann.

Stichworte – Multi-label Text Classification; Pseudo Labels; Latency; Transformer; ELECTRA; DistilBERT.