# ABSTRACT

This thesis investigates the detection of label noise in documents from the reinsurance industry. The study focuses on a large collection of internal PDF documents stored in R+V Re's document management system, which are categorized into different document types such as Slip, Wording, or ELDO. To identify miscategorizations, various methods are implemented and compared: the CNDC model, the SENT framework, and an uncertainty metrics-based sorting strategy using a fine-tuned BERT-base model. The first two methods are based on different feature groups extracted from the documents. These include text-based features such as TF-IDF vectors and NER entities, structural features based on document layout, and embedding features.

The results show that the CNDC model using text-based features achieves the highest precision of 73% in detecting noisy instances and delivers consistently good precision values across various document types. The model shows limitations only for document types with high semantic similarity, such as Wording, General and Special Conditions. The modified SENT framework, which uses a two-layer neural network instead of a BERT model, and the uncertainty metrics-based sorting strategy can also successfully identify noisy instances. However, their effectiveness depends more strongly on the respective document type, which results in lower average precision values. Among the investigated uncertainty metrics, the LowGT metric with a cutoff value of 10% proves to be a particularly suitable indicator for label noise.

The thesis demonstrates that automated noise identification methods can be successfully applied to documents in the reinsurance industry. The results suggest that particularly the use of more powerful language models in combination with the investigated approaches offers potential for further improvements. Thus, this work contributes to quality assurance in automated document processing.



**Keywords** - Label Noise Detection; Reinsurance; Feature Engineering; CNDC; SENT; BERT-base.