Zero-Shot and Few-Shot Natural Language Inference Models for Judgment Prediction

Boris Mottet and Tunahan Öszoy - Supervisor : Joel Niklaus and Janis G

UNIVERSITÄT BERN

Introduction

Court ruling prediction allows to speed up pre-trial processes and can be use to test the anonymization practice of Swiss courts.
Formulating text classification as a Natural Language Inference (NLI) task [2] enables strong Zero-Shot and Few-Shot performance in text classification tasks. The aim of this project is to apply this method to the multilingual Swiss Judgment Prediction benchmark [3].

What is Zero-Shot Inference ?

The aim of Zero-Shot models is to perform a task without training examples. This means that the In the case of Zero-Shot Inference, the model must classify on unseen classes. This is often used in sentiment analysis for example.

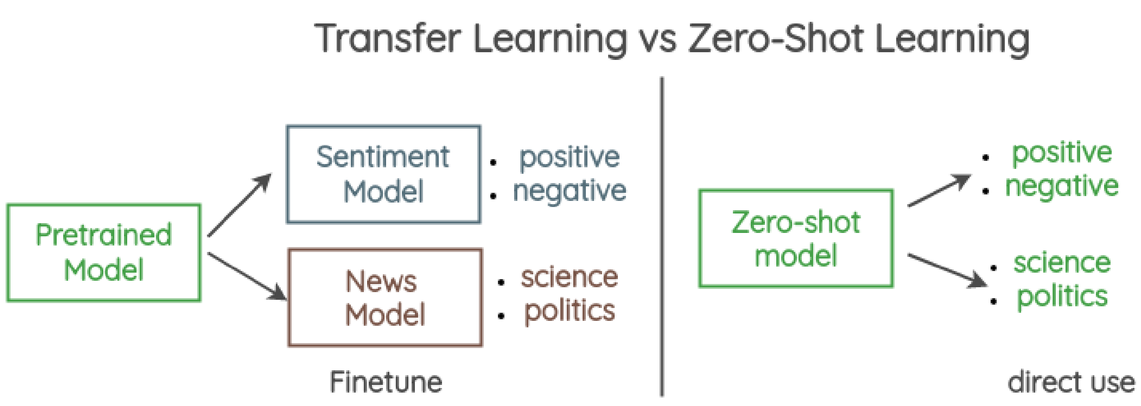


Fig. 1: Traditional transfer learning vs Zero-shot learning

How to apply the Zero-Shot Inference to the Swiss Judgement Prediction ?

Zero-shot models ask for a premise and a hypothesis to infer an entailment. An example with `bart-large-mnli` is

Premise : I have a problem with my iphone that needs to be resolved asap!!
Hypothesis : This example is {label}
Possible labels : urgent, not urgent, phone, tablet, computer
Entailment :

urgent	0.999
phone	0.995
computer	0.135
not urgent	0.001
tablet	0.000

For the Swiss Judgement Prediction, the premise are the facts of a case and the hypothesis is `"This case is {label}"` with `"approved"` and `"dismissed"` as possible labels. The model will output an entailment probability for both labels. The prediction is the label with the higher entailment.

Zero-Shot Models Comparison

We experimented with 3 different hypotheses on the Bart, Mdeberta and Roberta model and evaluated them on their Accuracy, their F1 macro scores and their Matthews correlation coefficient. Testing with different hypotheses and languages the maximum F1 score is at 0.5187 on the Roberta model with the french subset using the second hypothesis. So it is not significantly better than a random model.

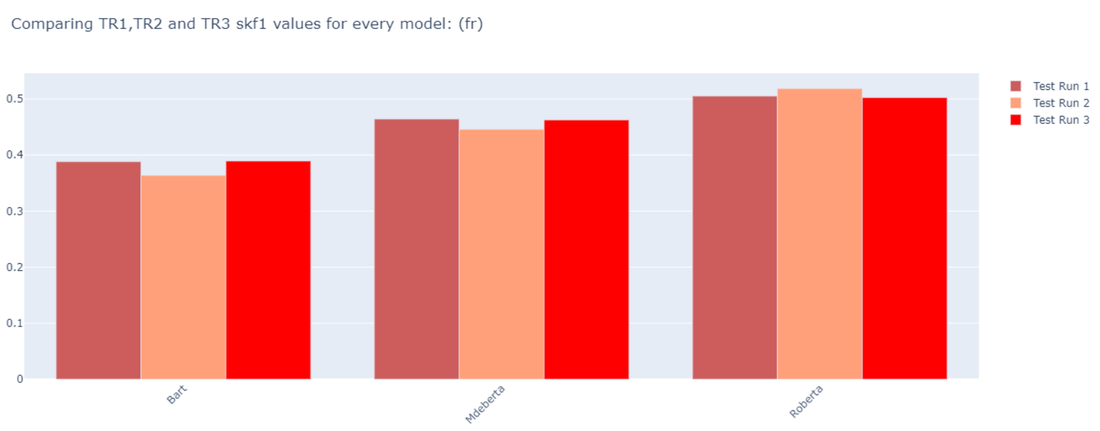


Fig. 2: Comparing the F1 macro score of 3 hypotheses with 3 models on the french subset

Conclusion

The Zero-Shot models tested did perform particularly well on the dataset. The length of the description of each case might have been too long and the lexical field of court ruling might have been too different from the one on which the models were trained.

Results by Legal Areas and Cantons

The legal area that achieved the most accuracy was social law. And it is very clear in this picture that the Roberta model performed the best. Furthermore for the german dataset, we found out that the german-speaking cantons perform best.

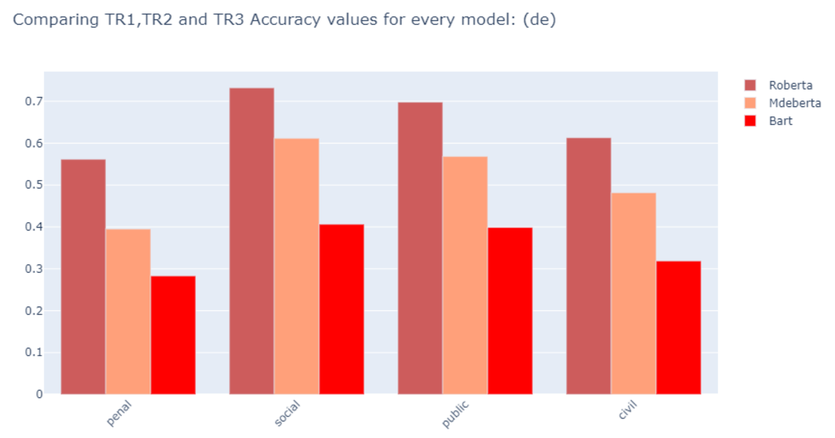


Fig. 3: Accuracy scores of all legal areas for all 3 models on the german dataset

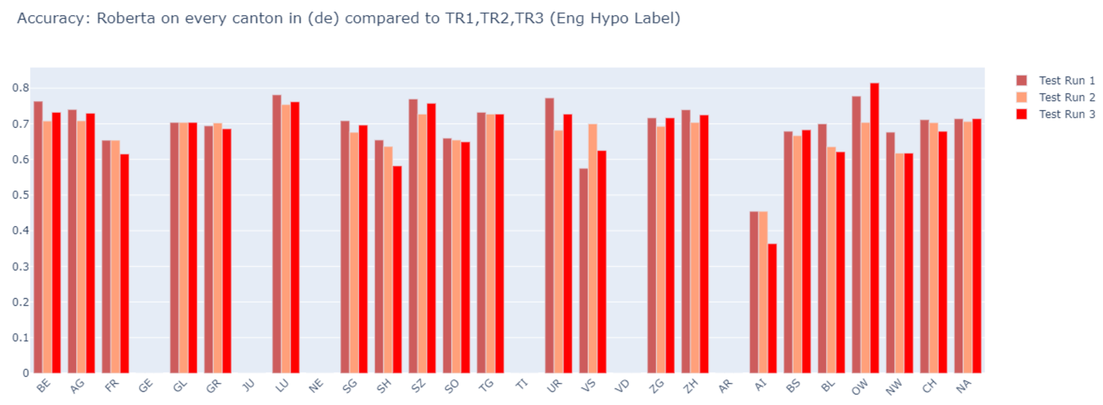
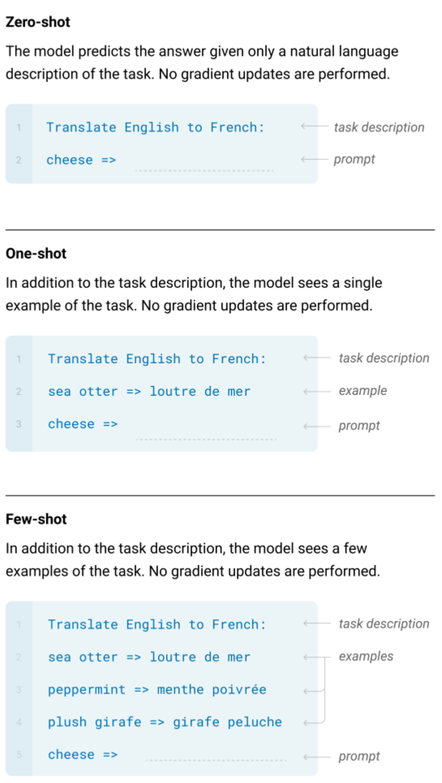


Fig. 4: Accuracy scores of all cantons for all 3 Hypotheses on the german dataset

What's next : Few-Shot Inference



Few-shot inference is similar to zero-shot but we allow the model to train on a few examples. This allows the models to be stronger. With the swiss-judgement-prediction dataset, the chosen model is fine-tuned using K examples with K varying from 10 to 200.
To produce examples, K cases are randomly chosen as training set. With those cases, an hypothesis and a label is associated. If the hypothesis correspond to the cases (i.e. if the case has been accepted, the hypothesis is "This case is accepted" and similarly for rejections) the label is "entailment", else it is "contradiction". With those examples, it is then possible to fine-tune a model.

Fig. 5: Zero-shot and Few-shot learning

References

[1] Vokinger, K.N., Muühlematter, U.J., 2019. Re-Identifikation von Gerichtsurteilen durch "Linkage" von Daten(banken). Jusletter 27.
[2] Yin, W., Hay, J., & Roth, D. (2019). Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. ArXiv, abs/1909.00161.
[3] Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In Proceedings of the Natural Legal Language Processing Workshop 2021, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.