

Towards Explainability and Fairness in Swiss Judgement Prediction: Benchmarking on a Multilingual Dataset

Santosh T.Y.S.S [†], Nina Baumgartner [‡], Matthias Stürmer ^{‡§}, Matthias Grabmair [†], Joel Niklaus ^{‡§¶}

[†] School of Computation, Information, and Technology; Technical University of Munich, Germany

[‡] Institute for Computer Science, University of Bern, Switzerland [¶] Stanford University

[§] Bern University of Applied Sciences, Switzerland

1. Need for Explainable LJP

- LJP: Determine case's outcome from the facts description
- Deep learning methods predict solely based on case facts, bypassing the interpretable legal reasoning process.
- Significant risk, when they rely on factors that may be predictive but lack legal relevance or involve sensitive attributes
- Such reliance lead to unjust and biased outcomes undermines the principles of fairness and equal treatment within the legal system.
- Need to be analyzed from an explainability standpoint to enhance the trust

2. Contribution

- Explainability
 - Rationales for 108 Swiss cases at fine-grained sub-sentence level
 - Labels: support/oppose Judgment and neutral
 - Perturbation-based Occlusion :Remove the rationales from the fact and measure the change in the prediction confidence
- Bias
 - Supreme Court of Switzerland handles the cases arose from lower court
 - Test bed to assess how much model rely on these lower court names and measure this bias through lower court insertion (LCI)
 - Insert other lower court names into each case and measure the changes in prediction confidence scores.

3. SJP Dataset [4]

- 85,000 cases from the Federal Supreme Court of Switzerland (FSCS)
- Spanning from 2000-2020, chronologically split into training (2000-14), validation (2015-16) and test (2017-20)
- Written in three languages: German (50K), French (31k), Italian (4K)

4. Explanation Rationale & Lower court Mentions

- Annotations for 108 cases from validation and test set
- Equal proportion of three languages, legal areas
- 3 legal experts - 2 law students, 1 lawyer
- Annotation Task: Annotate sentences or sub-sentences in the facts that "support" or "oppose" the final outcome
- Additional label for "oppose", unlike previous works, this considers perspectivism in legal reasoning
- Annotate "Neutral" sentences, not a label, rest of the other sentences to assist in segmenting legal text into sentences
- Annotate the lower court mentions in the fact

5. IAA

IAA score between the annotators in the first cycle for German subset

IAA metric	A1&A2	A1&A3	A2&A3
Rouge-1	0.78	0.69	0.87
Rouge-2	0.74	0.64	0.85
Rouge-L	0.77	0.68	0.87
BLEU	0.75	0.69	0.85
METEOR	0.77	0.71	0.88
Jaccard Sim.	0.73	0.64	0.82
Overlap Max.	0.68	0.61	0.74
Overlap Min.	0.83	0.73	0.81
BERTScore	0.91	0.86	0.93

6. Occlusion LCI Dataset

- Occlusion dataset
 - Instances by occluding 1,2,3,4 number of sentences with same label (opposes/supports/neutral)
 - Pair with baseline actual text to measure difference of prediction probability between them
- LCI dataset
 - Derive counterfactual based test set by replacing actual lower court mention with other lower court names
 - Pair with actual baseline to measure change between them

Statistics of the Occlusion and LCI test sets across the three languages.

	Occlusion				LCI
	Opposes	Neutral	Supports	Total	Total
DE	201	11124	1243	12568	351
FR	64	3811	2467	6342	391
IT	63	9155	203	9421	312

7. Metrics

- Explainability using occlusion
 - Difference in temperature scaled baseline and occluded instance
 - negative - oppose
 - positive - support
 - do not change - neutral
 - Report F1-score for each label
- LCI Bias
 - change in explainability score and report average of positive and negative values to measure
 - positive - pro dismissal
 - negative - pro-approval
 - Flip ratio of predicted label with inserted lower court name from baseline

8. Models

- Hierarchical model to deal with longer input
- Monolingual model [4]: GermanBert (German) [1], CamemBert (French) [3], Umberto (Italian) [6]
- Multilingual: XLM Roberta [2]
- Mono/Multilingual with DA [5]: easyNMT2 to get translated data
- Joint training with all languages [5]

9. Result- Occlusion

- Better scores for supports than neutral and opposes
- French better in support judgement but does not do well for other labels
- Multilingual models/Joint training improve the scores for supports mainly, neutral in some cases
- DA helped in multi/monolingual to improve explainability, but it did not improve in case of Joint training

Model	German		
	Opposes	Neutral	Supports
MonoLingual	3.02	16.78	15.1
MultiLingual	2.04	11.90	17.46
MonoLingual + DA	3.21	16.26	18.08
MultiLingual + DA	3.64	19.06	20.83
Joint Training	2.62	15.72	26.97
Joint Training + DA	3.75	14.54	21.95

10. Result- LCI

- Change in 5% of the confidence score in both directions
- These can bring label flips
- With DA component the bias further increased
- Joint training model improved prediction performance on Italian but it came at a cost of increasing bias scores of Italian, with higher flip rate indicating representational bias of dataset

Model	German			
	+ MES	- MES	Flip 1→0	Flip 0→1
MonoLingual	3.48 _{5,12}	-2.3 _{2,82}	2.28	0.43
MultiLingual	3.39 _{5,43}	-2.77 _{3,64}	1.71	0.32
MonoLingual + DA	3.09 _{5,15}	-2.77 _{5,42}	2.56	0.22
MultiLingual + DA	5.32 _{8,27}	-3.35 _{5,24}	4.56	2.56
Joint Training	3.32 _{6,18}	-1.86 _{2,89}	3.13	1.99
Joint Training + DA	3.23 _{4,46}	-1.84 _{2,45}	2.85	1.99

11. Conclusion

- Rationale dataset of 108 trilingual cases at fine grained level for supporting and opposing factors for SJP
- Perturbation based occlusion dataset to assess explainability
- Lower explainability scores across models indicate the current models do not predict right for the right reasons
- Bias of lower court using LCI test
- Average of 7 token has potential to flip predictions

References

- [1] Branden Chan, Timo Möller, Malte Pietsch, Tanay Soni, and Chin Man Yeung. Deepset-open sourcing german bert, 2019.
- [2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.
- [3] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villenote de La Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*, 2020.
- [4] Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Language Processing Workshop 2021*, pages 19-35, 2021.
- [5] Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. An empirical study on cross-x transfer for legal judgment prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 32-46, 2022.
- [6] Loreto Parisi, Simone Francia, and Paolo Magnani. Umberto: an italian language model trained with whole word masking. *Original-date*, 55:31Z, 2020.