LREC-COLING 2024

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

Santosh T.Y.S.S[1], Nina Baumgartner[2,] Matthias Stürmer[2,3],
Matthias Grabmair[1], Joel Niklaus[2,3,4]

[1]Technical University of Munich, [2]University of Bern,
[3]Bern University of Applied Sciences, [4]Stanford University

# Need for Explainable 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

# Our Contributions: Explainability & Bias

- 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.

# SJP (Niklaus et al., 2021)

- 85,000 cases from the Federal Supreme Court of Switzerland (FSCS)
- 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)

# Explanation Rationale & Lower Court

- 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

# IAA

Agreement within the categories "Lower Court" and "Supports Judgment" is notably high in comparison to "Opposes Judgment".

Can be attributed to difficulty in identifying them

| 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 |

# Occlusion & LCI Dataset

- Occlusion dataset
  - Instances by occluding 1,2,3,4 number of sentences with same label (opposes/supports/neural)
  - 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

| | 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 |

# 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 fairness:
  - 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

# Models

- Hierarchical model to deal with longer input
  - Monolingual model
    - GermanBert (German),CamemBert (French), Umberto (Italian)
  - Multilingual
    - XLM Roberta
  - Mono/Multilingual with DA
    - easyNMT2 to get translated data
  - Joint training with all languages

# Occlusion Explainability

- 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, neural 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 |

# LCI bias

- 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 $1 \rightarrow 0$ | Flip $0 \rightarrow 1$ |
| 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 |

# 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