What information does an algorithmic legal judgment prediction give?

Henry Prakken

JURIX 2024

Brno

12 December 2024

With thanks to Floris Bex

Some press on AI & Law (2016)

”**Artificially Intelligent ‘judge’ developed which can predict court verdicts with 79% accuracy**” (…)
“Computer scientists … developed an algorithm which can not only weigh up legal evidence, but also moral considerations.” (Daily Telegraph 24 Oct 2016)

N. Aletras, D. Tsarapatsanis, D. Pietro-Preotiuc & V. Lampos (2017). Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. *PeerJ. CompSci* 2e:93, DOI 10.7717/peerj-cs.93

The ECHR 'predictor'

- Trained on full text of decisions concerning 3 articles in the European Convention on Human Rights.
- **TASK**: did the court rule that an article was violated?
- **Results:** system's answer correct in 79% of the cases.
- **But:**
- Prediction not explainable on legal grounds
- The system does not predict outcomes
  - It needs most of the decision to be predicted

A survey

- Often claimed to be practically useful for judges
- But Medvedeva & McBride (2024):
  - 159 of 171 papers (93%) claiming to model legal judgment prediction need the decision-to-be-predicted
    - Remaining 7% has < 80% accuracy

M. Medvedeva & P. McBride, Legal Judgment Prediction: If You Are Going to Do It, Do It Right. *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 73–84.

Remainder of talk

- What about the 7% that does predict?
  - Which information does a prediction give to judges, lawyers, citizens?
  - Does the use of predictive algorithms promote consistency and predictability?

Prediction is not decision-making

- Judges don't predict but decide
  - justifying their decisions
  - not with statistical correlations but on legal grounds

An important distinction

- Algorithmic judgment predictors: do not do the same as judges
  - so performance cannot be compared
- Algorithmic experts: do the same as judges
  - E.g. recidivism prediction
  - So performance can be compared

What information does a judgment prediction give to judges (or citizens)?

F.J. Bex & H. Prakken (2021a). On the relevance of algorithmic decision predictors for judicial decision making. *Proceedings ICAIL 2021*, 175-179.

A Dutch judge in 2018:

- ‘Soon judges will have to explain why they deviate from an algorithmic decision prediction’
  - ‘If they deviate too often, they will have a problem’

- My question: does this make sense?

Underlying assumptions

- A decision predicted by a 'good' algorithm is the normal case decision
  - the decision an arbitrary competent judge would probably take
- So a judge can only deviate from the prediction if s/he can point at special circumstances
- My claim: the usual performance metrics do not imply a decision probability

'Decision probability'

From test set performance to decision probabilities (example)

- Suppose: an algorithm predicts that plaintiff will win, and:
  - 85% of the predictions for test cases were correct
  - The training and test cases are representative and their decisions correct and not outdated
  - The learned model is not 'overfitted'
- Is the probability that plaintiff will win 85%? No!

Analogy (1)

- 98% of Italians eat pasta at least once a week, Claudia is Italian
  - What is the probability that Claudia eats pasta at least once a week? 98%?

Analogy (2)

- 98% of Italians eat pasta at least once a week, Claudia is Italian
  - What is the probability that Claudia eats pasta at least once a week? 98%?
  - Claudia has a pasta allergy. So 0%!

Problem of the reference class

- The step from frequency to individual probability is a relevance judgement
  - Relying on a prediction = 'only the prediction is relevant'
  - But the judge always knows more about the case!
    - So the frequency does not apply to it
- But what if we have statistics about classes of cases?
  - Either too specific, so not enough data
  - Or too coarse, so reference class problem

Informing litigants about chance of success

- If domain can be modelled in terms of stable set and categories of features
- And enough data
- And system can engage in dialogue
- Then maybe useful
- But this requires KR!

Conclusions so far

- Claims that current research on legal judgment 'prediction' is useful for judges:
  - ignore that 93% does not predict
  - confuse predicting with taking decisions
  - overlook the reference-class problem
- My claim: LJP does not give any useful information to judges

Can legal judgement prediction improve the predictability and consistency of judicial decision-making?

F.J. Bex & H. Prakken (2021b). Can predictive justice improve the predictability and consistency of judicial decision-making? *Proceedings JURIX 2021, 207-214.*

Questions

- What do 'predictability' and 'consistency' of judicial decision-making mean?
  - Deciding the same case the same way
  - Deciding similar cases the same way

- How can algorithmic judgment predictors improve predictability & consistency?

Deciding the same case the same way

- Predictability & consistency promoted if all judges have to follow the same algorithm at all times
  - But what about incorrect, dubious or controversial predictions?
  - And if algorithm is not blindly followed, then predictions don't give useful information

Deciding similar cases the same way

- Predictions + numerical quality measure don't say much about similar cases
  - Algorithm might treat similar cases differently and vice versa
  - NB: textual similarity is not the same as legally relevant similarity!

D. Shu, H. Zhao, X. Liu, D. Demeter, M. Du & Y. Zhang, LawLLM: Law Large Language Model for the US Legal System. *33rd ACM International Conference on Information and Knowledge Management, October 2024*

Different contexts

- Banks want to reduce losses on loans in the long run
- Gamblers want to maximise gains in the long run
- …
- But judges want to optimize individual justice

Conclusions (2)

- Supporting judicial decision-making by data-driven judgment predictors can at best promote predictability & consistency in undesirable ways

Why is there so much research on legal judgment prediction?

Evaluation: basic concepts

- Evaluation = verification + validation
  - Building the system right vs building the right system
- Performance v. usefulness
- Laboratory studies v. field studies

Evaluating GOFAI

- MYCIN (1970s)
  - Lab, performance
  - 8 experts + MYCIN diagnosed and 'treated' infections
  - 3 senior experts rated quality
  - MYCIN performed best

Buchanan, B. G.; Shortliffe, E. H. (1984). *Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, Massachusetts: Addison-Wesley.
Online at https://www.shortliffe.net

Evaluating GOFAI&Law

- Tessec (Nieuwenhuis 1989):
  - Lab, real workers, supported or not
  - Usefulness
  - Better decisions with support
  -
- Tessec (De Bakker & Wassink 1991):
  - Before intro Tessec: 34 of 50 cases had errors
  - After intro Tessec: 18 of 50 cases had errors

J. Svensson (2002), The use of legal expert systems in administrative decision making.
In A. Gronlund (ed.): *Electronic Government: Design, Applications and Management*. London etc. Idea Group Publishing, pp. 151-169.

Evaluation of CATO

- Field test (Legal writing class)
- Usefulness
- Comparing groups instructed with resp. CATO and human instructor
- Pre- and post-test written argument exams, graded by instructor
- Both groups improved significantly and equally

Confusion matrix

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Key Metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
https://www.youtube.com/watch?v=RYFViaaJxE8

Evaluation in current research on LJP (and LLMs?)

- Focusses on performance, not on usefulness
- Hardly compares with human performance
- This should change