Pretrained Language Models v. Court Ruling Predictions: A Case Study on a Small Dataset of French Court of Appeal Rulings

Olivia Vaudaux,[1,2,3] **Caroline Bazzoli,**[1] **Maximin Coavoux,**[2] **Géraldine Vial**[3] **and Étienne Vergès**[3]
[1]Univ. Grenoble Alpes, CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, 38000 Grenoble, France
[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[3]Univ. Grenoble Alpes, CRJ, EA 1965
olivia.vaudaux@gmail.com, first.last@univ-grenoble-alpes.fr

Abstract

NLP systems are increasingly used in the law domain, either by legal institutions or by the industry. As a result there is a pressing need to characterize their strengths and weaknesses and understand their inner workings. This article presents a case study on the task of judicial decision prediction, on a small dataset from French Courts of Appeal. Specifically, our dataset of around 1000 decisions is about the habitual place of residency of children from divorced parents. The task consists in predicting, from the facts and reasons of the documents, whether the court rules that children should live with their mother or their father. Instead of feeding the whole document to a classifier, we carefully construct the dataset to make sure that the input to the classifier does not contain any 'spoilers' (it is often the case in court rulings that information all along the document mentions the final decision). Our results are mostly negative: even classifiers based on French pretrained language models (Flaubert, JuriBERT) do not classify the decisions with a reasonable accuracy. However, they can extract the decision when it is part of the input. With regards to these results, we argue that there is a strong caveat when constructing legal NLP datasets automatically.

1 Introduction

Natural language processing (NLP) has now many applications in the legal domain, and is used in practice to provide tools for law practitioners, such as lawyers or judges (e.g. for information retrieval, document classification), to quantify judicial risks in specific cases. Judicial decision prediction through NLP tools is also important for law researchers to identify the factors that most influence the final decision of a case, characterize judge's reasoning process, and help data annotation procedures.

In this paper, we present a case study on the prediction of the verdict ruling on the habitual residency of children in appeal cases from French courts. These documents follow a typical structure (facts, judgement of the first court whose ruling was the object of an appeal, decision itself). However, the structure is only implicit (sections are not titled), and judges' writing is not standardised and differs a lot from one court of appeal to another. Finally, it is often the case that information of each type (facts, reasons, decision) are scattered through the document. As a result, the document might contain 'spoilers' for the decision itself at various places in the document.

In order to assess the ability of BERT-style language models to infer a verdict from the legal content itself, rather than the full text of the document, we implemented an annotation campaign aiming at assigning a semantic type to every segment in each document as: (i) facts (ii) information about the first judgment (iii) reasons for the decision (iv) decision itself (containing explicitly the variable to predict) (v) none of these types. Thanks to this annotation procedure, lead by law experts, we compare the usefulness of various types of inputs for several types of classifiers, including bag-of-n-grams classifiers and BERT-based classifiers. In particular, we used 3 French pretrained language models: FlauBERT (Le et al., 2020) and CamemBERT (Martin et al., 2020) that were trained on generic-domain data, and JuriBERT (Douka et al., 2021), that has been specifically trained on French legal texts.

In summary, our contributions are as follows:

- We construct a dataset of 1k judicial decisions annotated at the segment level by law experts (law professors and PhD students);
- A set of empirical results on a judicial decision prediction task on this dataset;
- In regards to our (mostly negative) results, we argue for the need for a careful curation of

Proceedings of the Natural Legal Language Processing Workshop 2023, pages 38–43
December 7, 2023 ©2023 Association for Computational Linguistics

automatically constructed datasets for legal NLP.

2 Related Work

Predicting the outcome of a judicial decisions based on a textual input is rather recent and has first been applied to English decisions from the European Court of Human Rights (ECtHR Aletras et al., 2016). In particular, Aletras et al. (2016) constructed a dataset of ECtHR decisions focusing on cases about a few articles of the European Convention on Human Rights, and evaluated bag-of-n-grams approaches.

Subsequently, several studies were made on the ECHR corpus, consisting of approximately 11 500 rulings from 1959 to our days. Medvedeva et al. (2020) and Liu and Chen (2017) evaluated various linear classifiers. Kaur and Bozic (2019) used deep neural networks based on convolutional networks. O'Sullivan and Beel (2019) used word embeddings pretrained on documents from the ECtHR data. Chalkidis et al. (2019) introduced Legal-BERT, a BERT-style model pretrained on legal data (legislative texts, ECtHR cases and American contracts), and evaluated it on the ECtHR dataset. Other common dataset include decisions from the US supreme court, e.g. Katz et al. (2014) use decision trees to predict the votes of Judges. We refer the reader to Medvedeva et al. (2023) for a general survey on judicial decision prediction. In contrast to these publications, we focus on French, and on a very specific type of litigation: the habitual residency of children of divorced parents. Moreover we use both linear classifiers and deep nets based on fine-tuning pretrained language models.

As regards legal NLP applied to French, Şulea et al. (2017) and Sulea et al. (2017) evaluated SVM classifiers on the prediction of the area of a case, as well as its ruling, on decisions from the French Supreme Court (*Cour de Cassation*). Salaün et al. (2020) used pretrained language models such as Flaubert (Le et al., 2020) and Camembert (Martin et al., 2020) to predict the outcome of cases about conflicts between a tenant and a landlord in Québec. They observed that pretrained language models outperformed linear classifiers. Our dataset is on another type of litigation (habitual residency of children) and we focus on the manual construction of our dataset, instead of automatically constructing it. Finally, Douka et al. (2021) introduced JuriBERT, trained on *Légifrance*, an official website publishing all French law and evaluated it on topic classification tasks for documents from the Cour de Cassation (highest court in France).

3 Data

In France, when parents of minor children separate, they must decide on their children's place of residency. If the parents are unable to agree on a place of residency, the matter is referred to a court which, after hearing each parent, issues a ruling. We study the three principal outcomes: children live with their mother, their father, or in alternating residency. If the outcome of the lower court is not suitable for at least one parent, they can make an appeal to retry the case. New judges re-examine the facts and confirm or overturn the judgment handed down by the first court. These are the rulings we'll be using for the classification experiment.

We collected 987 rulings from French Courts of Appeal from the Jurica[1] databasis, that we accessed through a legal publisher (LexisNexis). The rulings have the following structure:

1. The metadata of the trial (date, place of appeal court, trial number, etc.);
2. The *facts*: the parties (the parents) are introduced along with their lawyers, children, and the composition of the court.[2];
3. The prior *judgment* under appeal is described, and the parties claims' and arguments are set out;
4. The *reasons* of the Court of Appeal's decision: the judge explains the decision to come based on sections of the law and the facts provided by the parties;
5. The reasons lead to the **decision** taken by the court: the reversal or confirmation of the first-instance judgment, from which we can infer the label we want to predict.

Although the structure of court rulings in sections (facts, judgment, reasons, decision) is conventional, judges' practices vary a lot in the length of these sections. For example, the facts section might be extremely short or very verbose, and its content

[1] Jurica is no longer maintained due to a recent change in policy about legal open data.

[2] The publisher replaced names by their initial in an effort to comply with French personal data laws (1978 Data Protection Act).

may not be restricted to what lawyers define as facts. Moreover information about *facts* or *reasons* may be scattered through the document and not only occur in their dedicated section. Therefore, instead of relying on the structure of the documents to extract information (and let the model be exposed to potential 'spoilers' about the decision), we rely on the manual annotation of every segment of the document as either (i) fact (ii) judgement (iii) reason (iv) decision, or (v) not relevant. The manual annotations were carried out by two law professors (who are coauthors of the paper and designed the annotation guidelines), one law post-doc researcher and one law graduate students who were trained for the task.

Due to the time required to carry out annotations, each document was annotated by a single annotator, so we did not compute interannotator agreement scores.

Annotating the rulings was quite challenging, especially to differentiate the facts and the reasons, because depending on the judge's method, the parties' claims and arguments can be set out in the reasons to justify the verdict. Therefore, in several rulings, the *facts* category is limited to the names and birth dates of parents and their children.

Prediction Task Our goal was to predict the outcome of appeal cases ruling on the residency of children among three possible outcomes: they live with their mother (label M, 50%), their father (F, 33%) or in alternate residency (B for both, 17%).[3]

We trained and evaluated classifiers settings in several settings, depending on the information input to the model: (i) F+R: Facts+Reasons (ii) F+R+J: Facts+Reasons+Judgment (iii) Decision (iv) Decision+Judgment. The last two settings are meant to assess the model's ability to extract the information about the decision,[4] whereas the first two assess the model's ability to make inferences about the judges' reasoning.

4 Models

Baseline We use various classical classifiers as baselines, namely:

- a multinomial naive Bayes (MNB);
- a k-Nearest Neighbors (k-NN) with Manhattan distance;
- a multiclass Support Vector Machine (SVM);
- a forest of decision trees, with a predefined number of trees created, where the final decision is made on the basis of the label most predicted by the trees;
- a Multi-Layer Perceptron (MLP) with 1 hidden layer of 100 units, with a *ReLU* activation function.

We vectorize the texts using bag-of-n-grams methods: either unweigthed, or weighted by TF-IDF. We use both unigrams, trigrams or both (based on preliminary experiments). For each algorithm, we made a Cross Validation to determine the best hyperparameters of each classifier. All baselines were implemented with Scikit-Learn library (Pedregosa et al., 2011).

Language Models As regards pretrained language models, we used FlauBERT (Le et al., 2020) and CamemBERT (Martin et al., 2020), two generic-purpose pretrained models for French, as well as JuriBERT (Douka et al., 2021), a language model trained only on data from the legal domain.

Because of the 512 tokens limitation of language models, the classifiers could not read the entire data in input, and thus did not make the difference between F+R and F+R+J, or didn't have all the argumentation of the judge. Therefore, we chunked each document based on line breaks, encode each chunk separately with the language model, and run a bi-LSTM (2 layers of 256 units) on the chunk representations to obtain a fixed-size vector for the whole document. We use dropout before feeding the representation to a classification layer.

Finally, to tokenize the data, we used the tokenizers associated with each language models, provided by the platform *Hugging Face*, and the associated library *transformers* (Wolf et al., 2020).

5 Experiments

Before starting data processing, we split the corpus into three stratified subsets: training (68%), testing (20%) and validation (12%).

[3] An alternative way of framing the task with two labels would be to predict whether the first judgment is (i) confirmed (ii) reversed. However, these two labels do not provide the same information as the three-label task we chose, since labels have to be interpreted with regards to the first judgment, and the second label (reversed judgement) is not as informative (2 possible outcomes).

[4] Sometimes the *decision* text only mentions that the first judgment is confirmed or overturned, making it necessary to know what the first judgment is to infer the label correctly.

Input	Language Model	Accuracy	F1-Score
Facts + Reasons	MNB+ bag-of-unigrams without TF-IDF	**0.57**	**0.54**
	FlauBERT	0.51	0.39
	CamemBERT	0.52	0.44
	JuriBERT	0.48	0.42
Facts + Reasons + Judgment	MLP + bag of-trigrams without TF-IDF	**0.56**	**0.50**
	FlauBERT	0.48	0.36
	CamemBERT	0.49	0.43
	JuriBERT	0.55	0.42
Decision	MLP + bag of-trigrams without TF-IDF	0.89	0.88
	FlauBERT	**0.90**	**0.89**
	CamemBERT	0.88	0.88
	JuriBERT	0.85	0.86
Decision + Judgment	MLP + bag of-trigrams without TF-IDF	0.74	0.69
	FlauBERT	0.95	0.94
	CamemBERT	0.92	0.92
	JuriBERT	**0.96**	**0.96**

Table 1: Accuracy and F1-Score score for the best baseline in each setting and for the three classifiers based on language models: FlauBERT, CamemBERT & JuriBERT. MLP : Multi-Layer Perceptron ; MNB : multinomial naive Bayes.

The BERT-based classifiers were trained for 80 epochs, with a batch size of 1, on a GPU. We used the AdamW optimizer (Loshchilov and Hutter, 2019) to minimize the loss function (negative log likelihood), and a scheduler with warmup, to vary the learning rate for a few warm-up periods. We tuned the learning rate ($\{10^{-6}, 2 \cdot 10^{-6}, 5 \cdot 10^{-6}\}$ for FlauBERT, $\{10^{-5}, 10^{-6}, 2 \cdot 10^{-6}, 5 \cdot 10^{-6}\}$, $\{10^{-5}, 2 \cdot 10^{-5}, 2 \cdot 10^{-6}, 5 \cdot 10^{-6}\}$ for JuriBERT) and the dropout value (0.15, 0.25)

For each classifier, we selected the best models based on validation F1 score (macro-average over classes. We also report accuracy in results.

Results and Discussion The test results for the best baseline (as selected on the validation set) in each setting and for the three classifiers based on language models are presented in Table 1. As regards the best baseline, we note that the best results were obtained on the *decision* and the *judgment + decision*, i.e. the parts of the rulings with the outcome explicitly stated. We found out that the MLP with bag-of-trigrams without TF-IDF weighting performed the best on all categories, except for the *facts + reasons*, where the multinomial naive Bayes algorithm with bag-of-unigrams without TF-IDF weighting reached the highest score. Finally, the results for F+R and F+R+J input are below 60%, slighlty above the most-frequent-label baseline which scores 50%.

For the language models, the categories with the highest results were those with the verdict stated. These results are even better than those of the baseline, exceeding 90% accuracy. The F1-score is also very high, showing it has nothing to do with the imbalance of the corpus.

Overall, we conclude that the classifiers based on pretrained language models have no difficulty reading over the decision and inferring the label (judgment+decision setting), however they struggle to infer what the decision will be when they are fed only the facts, reasons and judgment information. We hypothesize that the difficulty stems from the length of documents (on average 1095 tokens for F+R+J input). Moreover the key information usable to make a reasonable prediction is often scarce: most of the reasons cite the relevant articles and laws (that are not predictive of the output but make counfounding factors), which makes it hard to distinguish information that is important for the decision from irrelevant information. Finally, the documents do not follow standardized writing and differ considerably between each other (depending on the writer and the tradition of the local court).

The mismatch in results between settings with and without *decision* is an argument in our opinion in favour of the need to manually curate datasets for legal NLP, rather than only relying on automatic construction.

Human performance on the task As a control experiment, we assessed the performance of humans (either law experts or laypeople) on the task, on a sample of 100 documents, where humans were shown the F+R input, and had to make a guess at the output. Both law experts and laypeople had accuracies between 89 and 96%, which shows that the inputs contain enough information to predict the label.

6 Conclusion

Our experiments have shown the limitations of language models on judicial documents, since classifiers based on these models fail to understand the textual input data. They perform well when the verdict is already given, but fail when the text is more important and requires a deeper analytical reading, during the court's argumentation for example.

On the question of children's habitual residency, the subject is far too sensitive to entrust the task of deciding to an AI, and we can only imagine complementary tools, making it possible to reduce the duration of procedures, to support citizens in putting together their case or to help the parties' lawyers. By successfully processing data in such a way that a classifier can identify the most relevant terms and phrases, it would be possible to find the 'winning arguments,' those that convince the judge or that can tip a case over the edge. This could then help lawyers to build their arguments, by giving them access to these elements in cases similar to their own.

Ethics Statement

NLP applied to legal decisions raises several key ethical questions. Processing legal decision is important for understanding how legal reasoning work, and might also be used to analyse judges' biases in their decisions. However, as all technological tools, it is prone to dual use (Hovy and Spruit, 2016). The main risk is that institutions implement such models to replace judges, with harmful consequences, since machine learning systems are known to amplify biases they are exposed to in the training data, and are oftentimes not easily interpretable in their predictions, which makes unusable in a context where they may have an impact on humans.

Regarding personal data, the legal publisher who provided the data did not anonymize the decisions to a GDPR compliant standard. Indeed, they only replaced surnames with initials, but kept firstnames and town names unchanged. As a result, we are unable to release the annotated data. We acknowledge that not doing so is a hindrance for reproducibility and open science. In this case, we judge that personal data safety should be prioritized.

Acknowledgements

We thank Chen Chen for working on a pilot study that led to this paper, as well as Nour Safa (PhD student at Université Grenoble Alpes) and Djohar Sidhoum-Rahal (then a post-doc at Université Grenoble Alpes) for their work on the annotation of the dataset. We acknowledge the financial support of MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ computer science*, 2:e93.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2021. JuriBERT: A masked-language model adaptation for French legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–101, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Daniel Martin Katz, Michael J Bommarito II au2, and Josh Blackman. 2014. Predicting the behavior of the supreme court of the united states: A general approach.

Arshdeep Kaur and Bojan Bozic. 2019. Convolutional neural network-based automatic prediction of judgments of the european court of human rights. In *AICS*, pages 458–469.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Zhenyu Liu and Huanhuan Chen. 2017. A predictive performance comparison of machine learning models for judicial cases. In *2017 IEEE Symposium series on computational intelligence (SSCI)*, pages 1–6. IEEE.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the european court of human rights. *Articial Intelligence and Law*, 28(2):237–266.

Masha Medvedeva, Martijn Wieling, and Michel Vols. 2023. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31(1):195–212.

Conor O'Sullivan and Joeran Beel. 2019. Predicting the outcome of judicial decisions made by the european court of human rights. *CoRR*, abs/1912.10819.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Olivier Salaün, Philippe Langlais, Andrés Lou, Hannes Westermann, and Karim Benyekhlef. 2020. Analysis and multilabel classification of quebec court decisions in the domain of housing law. *Natural Language Processing and Information Systems*, 12089:135 – 143.

Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu, and Josef van Genabith. 2017. Exploring the use of text classification in the legal domain. *CoRR*, abs/1710.09306.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017. Predicting the law area and decisions of French Supreme Court cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722, Varna, Bulgaria. INCOMA Ltd.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.