Investigating the State-of-the-Art Performance and Explainability of Legal Judgment Prediction

Rohan Bhambhoria[†, ‡, ⋄, *], Samuel Dahan[†, ⋄, §], Xiaodan Zhu[†, ‡, ⋄]
[†] Ingenuity Labs & [‡] ECE, Queen's University
[⋄] Conflict Analytics Lab, Queen's University
[§] Cornell Law School

Abstract

In the past decade deep learning models have achieved impressive performance on a wide range of tasks. However, they still face challenges in many high-stakes problems. In this paper we study Legal Judgment Prediction (LJP), which is an important high-stakes task utilizing fact descriptions obtained from court cases to make final judgements. We investigate the state-of-the-art of the LJP task by leveraging the most recent deep learning models, *longformer*, and demonstrate that we obtain the state-of-the-art performance, even with a limited amount of training data, benefiting from the advantage of pretraining and the long-sequence modeling capability of *longformer*. However, our analyses suggest that the improvement is due to the model's fitting to *spurious correlations*, in which the model makes correct decisions based on information irrelevant to the task itself. We advocate that caution should be seriously exercised when explaining the obtained results. The second challenge in many high-stakes problems is interpretability required for models. The final predictions made by deep learning models are useful only if the evidences that support the models' decisions are consistent with those used by subject-matter experts. We demonstrate that by using post-hoc interpretation, the conventional method XGBoost is actually capable of providing explainable results with a performance comparable to the *longformer* model, while not being subject to the spurious correlation issue. We hope our work contributes to the line of research on understanding the advantages and limitations of deep learning for high-stakes problems.

Keywords: Interpretability, Explainability, Legal Judgment Prediction (LJP)

1. Introduction

Legal Judgment Prediction (LJP) is an important task which involves utilizing fact descriptions obtained from court cases in order to make decisions of the final outcome. The development of these models is crucial as they can reduce the time taken by legal professionals in determining the outcome of an ongoing case [1]. Alternatively, they may be used by these same professionals to reinforce their opinion on a decision to be made. By analyzing the bias which may have been present, practitioners in the legal field can identify decisions which may have previously been made due to various other commonly present factors, such as nationality or gender [2]. This factor is essential in determining a lesser number of cases which are able to generalize well and contain an acceptable amount of human bias for further annotations by professionals, as this task may be expensive and time-consuming. Additional insights in the model behaviour are essential in high-stakes decision making fields such as healthcare, finance, and law. For this reason, we are not only interested in analyzing predictions which are made by models, and obtaining models with capabilities of high performance, but we are also interested in interpretability. Previous works utilizing fact descriptions in predicting final outcomes under binary classification and multi-label classification settings [3] mainly focus on performance, and interpretability is largely ignored. Few works [4] tackle the challenge of interpretability for legal judgment prediction, although they are mainly focussed on datasets in the Chinese language.

[*] r.bhambhoria@queensu.ca

In this paper, we obtain state-of-the-art results by using a transformer-based model called longformer [5], capable of modelling long sequences. To the best of our knowledge, we are the first at utilizing the longformer for the LJP task. Previous models such as BERT [6], were restricted to modelling sequence lengths of 512 tokens. The task in the legal domain we tackle for achieving these results is the high-stakes decision making problem of LJP, which comprises of binary classification for determining the presence of a violation in a human rights article. As per our analyses, we find that improvement in the models' predictive capabilities is due to the fact that these results were reliant on spurious correlations. This fact is echoed in our results, as state-of-the-art performance is obtainable even under the consideration of a limited amount of training data, leading to sparse data conditions. In other words, we must be wary of the capabilities of transformer-based models producing state-of-the-art results in the legal judgment task, as they may be due to the wrong reasons. Through making use of post-hoc interpretation methods such as LIME [7], we are able to address the challenge of high-stakes decision-making having the requirement of interpretability. We exhibit the capabilities of the conventional XGBoost model [8] in providing explainable results with similar performance to that of the best results produced by the longformer [5].

2. Related Work

Language modelling architectures have been created for modelling long sequences [5], which are particularly useful for legal tasks, as court cases comprise of long documents. Most current legal document classification works utilize hierarchical architectures [3], while other works simply pose the long document challenge [9]. In this work, we use a recently introduced transformer-based model for modelling long sequences, called longformer [5].

Previous works have also dealt with elements pertaining to explaining specific model predictions, and this can be seen as an important goal in research for interpretability. However, it is non-trivial to explain why a deep non-linear neural network model makes a certain decision, a task categorized under the umbrella of ante-hoc methods [10]. For the uncertainty, and lack of ability of these methods to draw out a cross-comparison between models chosen in this paper, we refrain from utilizing ante-hoc methods. On the other hand, post-hoc interpretability is the application of interpretation methods to a trained model. We utilize LIME [7], and Layer Integrated Gradients [11] to obtain insights on the decision-making path from input text to outputs of our models for classification. In the legal field, [12] highlights the importance of using such methods, although the application has largely been unexplored. In this work, we explore the application of these post-hoc interpretation methods to the task of legal judgment prediction.

3. Dataset

We utilize the ECHR dataset [3] for experiments. The dataset is split into training, validation and testing sets. The training and validation datasets are roughly balanced, and the test set is imbalanced, with 66% of cases having violations. The ground truth corresponds to the presence/absence of violations in cases. Following the original work [3], we also utilize an anonymized dataset to identify tokens which may be influencing predictions based on demographic or factual information. This setting simply makes use of a named entity recognizer to replace tokens related to location, gender, etc. with type tags [3].

4. Methodology

We utilize a combination of different methods for models and interpretations methods to identify the best performing pair in terms of performance, and explainability. We emphasize that in this paper, we treat both, the models, and these interpretability methods

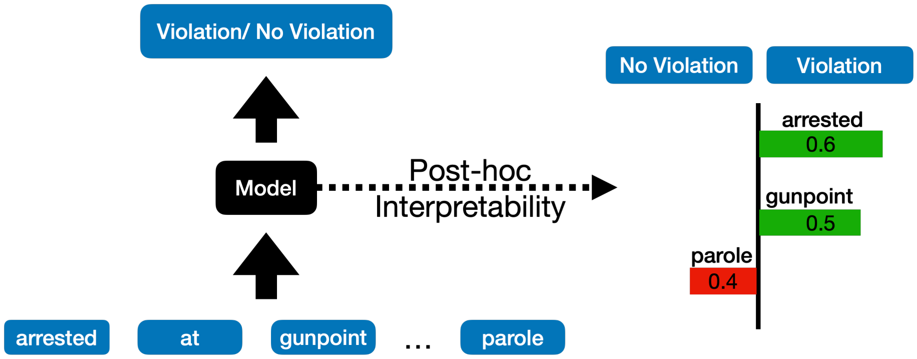

Figure 1. The model is trained on input sequences for the binary classification task of predicting "Violation/No Violation". Through post-hoc interpretation methods, we are able to derive impact of tokens on probabilities influencing the model output

as black-boxes. Hence we do not detail specifics about perturbations on instances of our test dataset. Instead, we simply utilize the interpretability methods as tools for producing simpler models which may be capable of providing explanations.

XGBoost: This model is created by an ensemble of weak prediction decision trees [8]. In addition to performance, XGBoost is highly interpretable, making it an ideal candidate for experiments.
Longformer: The recently introduced transformer-based model is able to achieve state-of-the-art results on long sequences. It makes use of a linearly scalable attention mechanism, which combines a local windowed attention with a task motivated global attention [5]. Due to the performance achieved on long sequences, we select this model to gain insights on attributes influencing decisions made.
LIME: Local Interpretable Model-agnostic Explanations (LIME) [7] is a widely used method. It works by learning a sparse linear model around a perturbed instance, which we consider as our test case. The idea behind creating a simpler model to provide an interpretation is that it may be more difficult to obtain an interpretation globally, whereas obtaining the same locally is simpler. In this paper, we use LIME for both models XGBoost, and Longformer.
Layer Integrated Gradients: A strong axiomatic attribution method [11] which can be used for identifying input features which attribute to a prediction of a deep neural network. We use this method for Longformer to gain additional insights in the model behaviour.

5. Experimental Setup

For the XGBoost model, we utilize TF-IDF vectors with an ngram range of (1, 5), following the hyperparameter values set in the original paper [3]. For purposes of providing interpretation, we use LIME [7] out-of-the-box with default parameters, which produces 10 features under the consideration of 5000 samples.

For the longformer model, we run experiments with a batch size of 1, learning rate of 2e-5, and for 10 epochs on two RTX 2080Ti GPUs. Training time for 100% of the data is $\approx$ 8 hours. For purposes of providing interpretation for the transformer-based model using LIME, we set the number of features calculated to 5, and the number of samples to 50. For the layer integrated gradient [11] method, we set the number of attribution steps to 500.

6. Results and Discussion

Results on both, non-anonymized and anonymized datasets are shown in Table 1. The results of the majority class are identical to those shown in [3]. Under the non-anonymized data setting, our XGBoost model achieves stable, and high performance which is comparable to the results of our transformer-based model for the same setting. The longformer, being capable of modelling longer sequences achieves better results consistently as expected. We attain state-of-the-art results by a margin of $\approx 1\%$ for the F1 score when utilizing as less as 10% of the training data for the longformer model. However, we must be wary of the results produced by utilizing lower amounts of training data here, as the performance may be a direct result of spurious correlations.

For the anonymized setting, the XGBoost model consistently outperforms the longformer under consideration of all variations of subsamples of data. Again, without utilizing the entire training dataset, the best results are obtained using $\geq 50\%$ of the training dataset for both models considered. Under the anonymized data setting, we observe that the results produced by the XGBoost model, again, are stable with variations in the amount of the training data utilized. In addition to this, the XGBoost model consistently outperforms the longformer. This indicates that the high performance of the longformer in the non-anonymized setting may have been due to tokens such as gender and location which are now replaced by type tags by anonymization. This directly accounts for decisions being made for incorrect reasons, under the well-known problem of spurious correlations. On the other hand, the performance of the XGBoost model in contrast to the longformer is largely unaffected by the anonymized data setting. Further insights into decisions being made by transformer-based models and the comparison with the conventional model can be drawn from interpretation methods which perturb inputs and deal with the models themselves as black boxes.

Table 1. F1 scores (macro-averaged) on the binary prediction task for presence of violation in court cases. Results for Hier-BERT, and HAN are taken from [3].

	F1 Score	
Majority Class	0.397	
	Non-Anonymized	Anonymized
XGBoost (10%)	0.799	0.798
XGBoost (25%)	0.810	0.811
XGBoost (50%)	0.808	0.814
XGBoost (100%)	0.809	0.814
Longformer (10%)	0.832	0.539
Longformer (25%)	**0.833**	0.775
Longformer (50%)	0.827	0.797
Longformer (100%)	0.826	0.791
Hier-BERT (100%)	0.820 ± 0.009	-
HAN (100%)	-	**0.802** ±0.027

We use a post-hoc interpretation method [7] for the XGBoost model (LIME [7]). We can observe from figure 2 that the most commonly occurring tokens which account for the predictions are relatively unchanged in terms of impact on probability with respect to the percentage of data utilized. This suggests that a limited amount of data may suffice in determining not only the predictions, but also for the right reasons under the non-anonymized dataset for the XGBoost model.

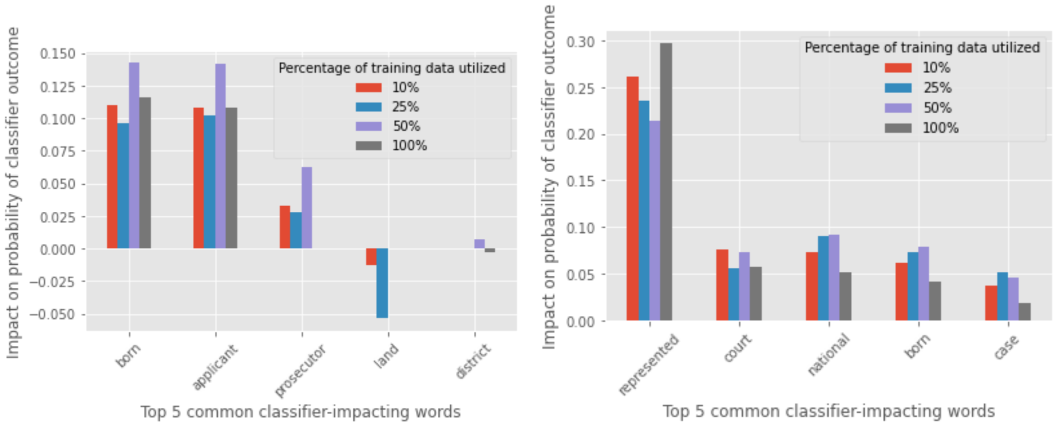

Figure 2. Impact of common tokens on XGBoost model towards the presence (left), and absence (right) of violation in binary classification for the non-anonymized dataset.

For the longformer model, we observe that post-hoc interpretation methods such as LIME [7], and Layer Integrated Gradients [11], when used out-of-the-box, highlight the importance of the beginning of sentences, punctuation, and numbers. In terms of interpretability, this may suggest that the models may be making correct predictions for incorrect reasons. On the other hand, it may suggest that invalid reasons are being picked up due to the computational resource constraints of methods required for interpreting large transformer-based models. Comparing it to the results for the longformer under the anonymized setting in Table 1, we can observe that for state-of-the-art results obtained when utilizing 10% of the training dataset may mostly have been due to the confidence of the model in the choice of a select number of tokens which consistently appear in text leading to violations.

7. Conclusion

In this paper, we produce state-of-the-art results on a binary classification task of legal judgment prediction for identifying violations in English under the non-anonymized data setting using the longformer model. We observe that a limited number of training samples are sufficient to achieve these results. Furthermore, under this data setting we can see that the XGBoost model is capable of achieving comparable performance. We also experiment with the anonymized data setting and conclude that conventional XGBoost models are capable of outperforming transformer-based models under circumstances where the model is unable to learn from tokens which lead to bias, such as nationality and gender. We can conclude that the XGBoost model is largely unaffected by the introduced bias, whereas the longformer is prone to learning from these bias-inducing tokens. This can be observed in the reduced performance when comparing the anonymized setting of data for the longformer with the non-anonymized setting. Finally, on experimentation with post-hoc interpretation methods, we observe that predictions made by the XGBoost model at the bare-minimum, offer transparency. On the other hand, the interpretability obtained by utilizing these methods indicate that the transformer-based models may have been making predictions for invalid reasons (spurious correlations).

Acknowledgement

This research is supported by CRCC New Frontiers in Research Fund (NFRF) and NSERC Discovery Grants and Accelerator Supplement.

References

[1] H. Zhong, Z. Guo, C. Tu, C. Xiao, Z. Liu, and M. Sun. “Legal Judgment Prediction via Topological Learning”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 3540–3549. DOI: 10.18653/v1/D18-1390. URL: https://www.aclweb.org/anthology/D18-1390.

[2] K. Chaloner and A. Maldonado. “Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories”. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 25–32. DOI: 10.18653/v1/W19-3804. URL: https://www.aclweb.org/anthology/W19-3804.

[3] I. Chalkidis, I. Androutsopoulos, and N. Aletras. “Neural Legal Judgment Prediction in English”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4317–4323. DOI: 10.18653/v1/P19-1424. URL: https://www.aclweb.org/anthology/P19-1424.

[4] H. Ye, X. Jiang, Z. Luo, and W. Chao. “Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1854–1864. DOI: 10.18653/v1/N18-1168. URL: https://www.aclweb.org/anthology/N18-1168.

[5] I. Beltagy, M. E. Peters, and A. Cohan. *Longformer: The Long-Document Transformer*. 2020. arXiv: 2004.05150 [cs.CL].

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 2016, pp. 1135–1144.

[8] T. Chen and C. Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

[9] J. Soh, H. K. Lim, and I. E. Chai. “Legal Area Classification: A Comparative Study of Text Classifiers on Singapore Supreme Court Judgments”. In: *Proceedings of the Natural Legal Language Processing Workshop 2019*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 67–77. DOI: 10.18653/v1/W19-2208. URL: https://www.aclweb.org/anthology/W19-2208.

[10] K. Sokol and P. Flach. “Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 56–67. ISBN: 9781450369367. DOI: 10.1145/3351095.3372870. URL: https://doi.org/10.1145/3351095.3372870.

[11] M. Sundararajan, A. Taly, and Q. Yan. “Axiomatic Attribution for Deep Networks”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML’17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 3319–3328.

[12] P. Hacker, R. Krestel, S. Grundmann, and F. Naumann. “Explainable AI under Contract and Tort Law: Legal Incentives and Technical Challenges”. In: *SSRN Electronic Journal* (Jan. 2020). DOI: 10.2139/ssrn.3513433.