Reducing Judicial Inconsistency through AI: A Review of Legal Judgement Prediction Models

Yifan Wu

School of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, Chongqing, China

Abstract. Ensuring equitable sentencing is a fundamental objective of the judicial system. However, disparities in law enforcement standards, policies, and personnel competence across regions can lead to divergent sentencing outcomes for similar cases. This inconsistency undermines the integrity of justice and diminishes public confidence. With the development of AI technology, especially in the field of NLP, more and more researchers are focusing on the role that AI can play in legal judgements, and the LJP model has been developed. The LJP model is widely expected to help reduce the judicial inconsistency that currently exists, and better help to maintain the fairness and justice of the law. This paper summarizes the latest developments in the field of LJP, introduces and compares some of the current representative works, including the advantages and disadvantages of current technology. After that, it discusses possible future research directions and considers the significance of the development of this field.

1 Introduction

In current legal practice, fairness in sentencing is a core objective of the judicial system. However, due to factors such as differences in law enforcement standards, policies and the quality of law enforcement personnel, there are inconsistencies in the sentencing outcomes of similar cases.

Two illustrative cases are provided below for reference: In a restaurant in Region A, Alice stole the victim's an electric bicycle worth 1860 yuan. Alice fled to the entrance of a nearby alley and was discovered and captured by the victim. The offender was sentenced to five months in prison and a fine of 2000 yuan； When passing by a cinema in Region B, Bob found an unlocked electric bicycle(valued at RMB 2071) parked near the cinema and stole it. The owner of the bicycle found it and reported it to the police. The next day, Bob was arrested by the police. The offender was sentenced to six months in prison and a fine of RMB 4000.

In both cases, the perpetrators exhibited similar behaviour and the value of the stolen items was comparable. However, the sentences handed down were disparate. Such inconsistencies not only impact the perceived fairness of the judicial system but also erode public trust in it.

Corresponding author: 2022212705@stu.cqupt.edu.cn

The analysis of this legal inconsistency has been going on for decades. Anderlini et al [1] theoretically analysed and compared the judicial consistency of different legal systems through mathematical methods. Some researchers analysed real-world data from a case-by-case perspective. In recent years, the Legal Judgment Prediction(LJP) model having developed considerably, Wang et al.[2] used the JLP model to establish the Legal Inconsistency Coefficient(LInCo) to analyse the consistency of sentencing.

The evolution of the JLP model has been influenced by the advancements in the domain of Natural Language Processing (NLP). Current LJP models use a variety of different technical principles, which makes the performance of different models somewhat different. At present, the more excellent models in this field, such as the Legal Attention-based Document Analysis(LADAN) model [3]. The model incorporates an attention mechanism that facilitates enhanced comprehension of pivotal information within legal texts, thereby enhancing the precision of judgment prediction. The NeurJudge model [4] proposes a crime plot perception behaviour separation technique and constructs a new neural judge framework. These techniques help the model can process complex legal texts, predict judgment results, and perform well on multiple legal datasets.

The main work of this paper is as follows: The author describes the current work in the field of LJP from the perspective of NLP and legal professionals. Also elaborates on methods including LM or those based on embedding-based and symbol-based methods. Then discusses possible future developments in the field of LJP. Additionally, the author delineates the prevalent applications of LJP, including judgment prediction and case matching. Finally, the author discusses the value of developing LJP models.

2 Key technologies of LJP models

The author will commence by introducing the current analysis of legal inconsistency, then outline the key technologies and representative models that are significant for the evolution of the LJP model.

2.1 Legal consistency analysis

There have been numerous discussions on the issues related to similar cases and similar judgments. These discussions have been based on the comparison and analysis of individual cases. And they have shown that in legal practice, when faced with similar cases, there are two situations that lead to significant differences in the results of the judgments: The first is when disparate charges lead to significant discrepancies in sentencing. The second is when the same crime is accepted, due to social opinion, local protectionism, differences in the understanding of the law by the subject of the trial, and the judge's broad discretion, the final result will be significantly different. Wang et al. initially employed the LJP model to propose the LInCo coefficient and subsequently analysed a substantial corpus of legal judgments in China.

The current research indicates that there are significant differences in the judgments of similar cases in different regions, especially in the sentencing of minor crimes. And the impact of the difference in the region where the case belongs is greater than the difference in gender. This difference not only violates the judicial principle of ‘similar cases are judged similarly,’ but it may also give rise to instances of judicial injustice. Therefore, people currently need a more pertinent and fairer system to assist in the judgment, which to some extent promotes the development of the LJP field.

2.2 LJP models

Current LJP model construction ideas include, but are not limited to, the following: using mainstream text classification models such as BERT, or additional pre-training on specific domain corpora based on this model to improve the model's understanding of the unique language patterns in the legal field; using label embedding technology to integrate the semantic information of legal articles and crimes into the factual description text through vector representation; and integrating causal knowledge into neural networks to enhance the model's reasoning ability, thereby supporting more reasonable decision-making. These ideas have all contributed to the progress of the LJP model.

2.2.1 Pre-training language model（PLM）

Current PLMs, even for large models with hundreds of millions of parameters, are difficult to analyze complex texts in a single text domain, let alone multiple domains. Therefore, it is valuable to build domain-specific corpora to pre-train models for specific domains [5].

Based on this, current researchers have proposed language models such as LEGAL-BERT [6] and LEGAL-ROBERTA [7] that have been pre-trained in the legal field to predict the outcome of a judgment. However, due to the differences between the texts used by existing PLMs and legal texts, the performance of PLMs directly applied to legal tasks is unsatisfactory [8].

Therefore, researchers have proposed that more machine learning-friendly datasets can be constructed, or that structured prediction methods can be used to achieve better prediction results.

2.2.2 Element extraction technology

Some researchers believe that clearer classification and judgment results can be achieved by better extracting the characteristics of easily confused legal articles and case descriptions. In the past, researchers have had several classic encoding models for element extraction, such as LSTM [9], DPCNN [10], and BERT [11].

Currently, in the specific field of law, new frameworks have also been proposed. The two relatively excellent models currently available are as follows:

Xu et al. proposed the LADAN framework, which introduces a graph distillation operator to extract discriminative features from similar legal articles, thereby effectively distinguishing between easily confused legal provisions.

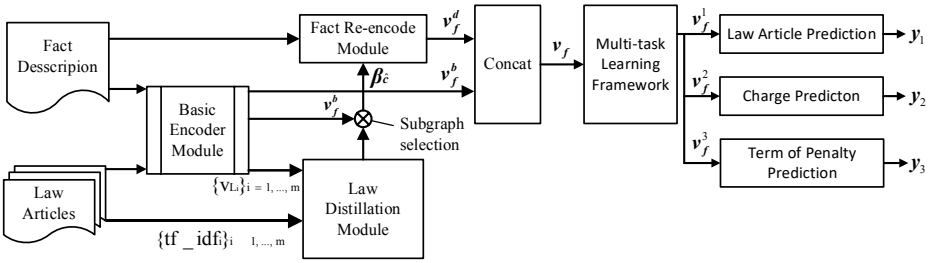

Fig. 1. An overview of the LADAN framework.

Fig. 1 shows that LADAN accepts case fact descriptions and legal article texts as input, first generating a basic representation v_f^b and a distinguishing representation v_f^d from the fact descriptions via a basic encoder and re-encoder. These two representations are subsequently merged to perform the prediction tasks.

Graph distillation operator (GDO) divides similar legal articles into multiple legal article communities. On this basis, LADAN processes the case description text through the following two representations: Base Representation: extracted directly from the description text and used to match the legal community. Discriminative Representation: extracted from the description text through the fact re-encoding model and used to distinguish between different legal articles within the same community.

Li et al. proposed NeurJudge, a plot-aware neural framework. This framework divides the fact description text into two main parts: Adjudging Circumstance(ADC) and Sentencing Circumstances(SEC). NeurJudge first predicts the applicable legal provisions in the case through the analysis of ADC. This step is based on the content of the conviction circumstances, and the goal of the model at this stage is to select the most appropriate legal provisions. After that, NeurJudge further divides SEC into Statutory Sentencing Circumstances(SSC) and Discretionary Sentencing Circumstances(DSC), and uses them to predict the specific sentence.

Additionally, Li et al. designed an extension called NeurJudge+ to address the issue of confusing verdicts. NeurJudge+ builds upon NeurJudge and enhances legal case prediction performance through the use of graph-based techniques. Specifically, NeurJudge+ constructs of labels which are built to aggregate confusing charges and articles, respectively. They propose a Graph-based Label Embedding (GLE) method consisting of Graph Decomposition Operation (GDO) and Label-to-Fact (L2F) attention. The GDO method effectively extracts distinguishing features for labels from similar labels on the label similarity graphs. The L2F attention mechanism then interacts with these features and the fact vectors, enhancing the final fact representation. This method effectively improves the performance of charge and article prediction.

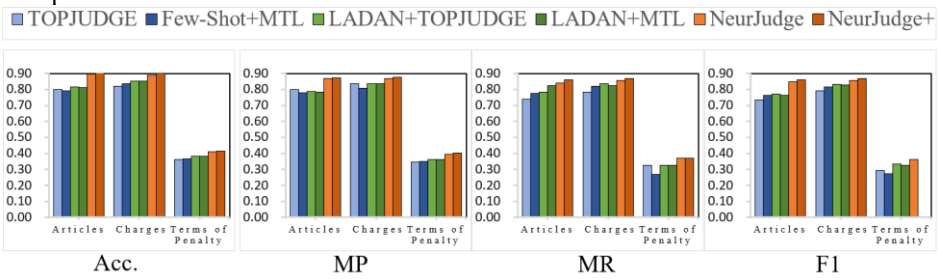

Fig. 2. Results of different models and their variants of CAIL-small on four metrics

From the results of Fig. 2, the paper could get the following observations: (1) The latest models are gradually evolving and improving, and are increasingly able to make more accurate predictions. (2) The performance of models on different tasks is indicative of the difficulty of the tasks. It is evident that current models lack the ability to accurately predict the terms of penalty.

The idea behind both frameworks is similar: the given text is divided into legal provisions and then further subdivided. This is done by extracting elements and characteristics multiple times to better match the corresponding detailed fields and make more accurate predictions. The possible drawback of this approach is that it is not very explanatory and does not clearly explain the logic behind the predictions. Furthermore, the amount of training required for this method is usually large, which is not conducive to the prediction of niche cases that do not occur frequently.

2.2.3 Similar case matching technology

In countries that adopt the common law system, judicial decisions are based on similar and representative cases in the past. Therefore,
how to match and use similar cases to make predictions has also become an important topic in the field of LJP. Currently, many researchers are focusing on building an easy-to-use legal search engine. Vu Tran et al. [12] combined the method of lexical features and latent features embedded in abstract attributes to construct a method for modelling case abstracts as a continuous vector space, which helps to build an efficient legal case retrieval system [13]. Arian et al. [14] designed a series of techniques for automatic abstracting of legal documents, and based on this, combined with models such as Vanilla BERT and Longformer, they achieved excellent results in case retrieval.

In addition to the above research, many other excellent case matching techniques have been proposed. However, the current technology may encounter the following challenges: First, in the special field of law, case matching is different from common semantic matching. Compared with the similarity between words and sentences, it is more necessary to consider the similarity of legal concepts and legal knowledge, which requires a clearer and more explicit explanation and marking of key symbols in law. Second, it is difficult to apply scattered legal knowledge to the model. Although there are some knowledge graphs in this field, such as Li et al. establishing a text-guided legal knowledge graph reasoning, the effect is far from satisfactory. There is a need to build better knowledge graphs and find a reasonable way to apply them to the legal field.

2.2.4 Other related work

Chen et al. [15] analyzed the shortcomings of the LJP model using the theory of structural causal models (SCMs), including data imbalance, lack of human knowledge, and incomplete testing data. Liu et al. [16] constructed causal graphs from factual descriptions. Reducing manual intervention and enabling the model to make causal inferences can help legal practitioners make correct decisions efficiently. In the field of legal text retrieval, incorporating internal and external structural information into the task is also a new idea (Shao et al. [17], Li et al. [18]).

Several datasets have also been established. Tu et al. established the Cail2018 dataset [19], which has become the most commonly used Chinese judicial judgment prediction dataset. It contains 2.6 million criminal cases published by the Supreme People's Court of China, which is several times larger than other datasets in existing judgment prediction work. After that, they constructed CAIL2019-SCM [20] for similar case matching in the legal field. Ilias et al. [21] established the LexGLUE dataset for English legal language understanding. In addition, there are general legal corpora such as MultiLegalPile [22], LexFiles [23], and CaseHOLD [24].

3 Challenges and prospects

3.1 The limitations of preceding research

Although the LJP model has made some progress in the field of legal judgment prediction, there are still many limitations in the current technology. The following are some noteworthy points:

First, during the development process, researchers should focus on enhancing the interpretability of LJP model decisions. Most current LJP models are based on complex

neural network structures. This has improved prediction accuracy to some extent, but it also leads to insufficient interpretability of the model. Being able to give a convincing explanation of the decision made is very important in the process of legal practice. Judicial decisions require not only correct results, but also a transparent and explainable process. This is to avoid the parties not being able to understand or challenge the legitimacy of the decision. At the same time, the existing models require a large amount of data during the training process to achieve the desired prediction effect. However, for those uncommon and niche cases, the scarcity of data will lead to problems such as overfitting. This will prevent the model from learning effectively, which will affect the accuracy of the prediction.

The integration of legal knowledge is also a challenge. The specificity and complexity of legal texts make it difficult to effectively integrate legal knowledge into LJP models. Although technologies such as knowledge graphs have performed well in other fields, their application in the legal field has not yet met expectations. Existing technologies cannot fully capture the deep connections between legal provisions, which limits the performance of the model in complex cases

The development and application of LJP models raises specific ethical issues. Because LJP models can have a profound impact on the parties involved and society as a whole, ethical issues must always be considered in their development. During training, models may inherit or amplify biases in the training data. These biases may stem from historical data that is unfair in terms of race, gender, socioeconomic status, etc., and ultimately result in hidden algorithmic discrimination. At the same time, if the model makes a wrong judgment, it is difficult to determine whether the responsibility should be borne by the developer, the judge, or the model itself.

3.2 Future development directions

In view of the limitations of the current technology, future research and development can be carried out in the following areas:

The future development of LJP models should focus on the interpretability of the model, making the decision-making process of the model more transparent, so as to ensure that it can improve the trust of legal practitioners and the public in judicial applications, and also provide feedback support for the optimisation of the model. In addition, it is very important to reduce the cost of model training. Researchers can explore data augmentation techniques, transfer learning or few-shot learning to improve the model's adaptability to small samples. At the same time, they can improve the model's transferability between different legal domains through multi-domain pre-training, multi-task learning or mixed expert model(MoE) techniques, so that it can still maintain efficient and accurate prediction when dealing with cross-domain cases.

At the same time, researchers can explore the construction of a more detailed and comprehensive legal knowledge graph, and enhance the model's understanding and application of the relationship between legal provisions and legal concepts through reinforcement learning or symbolic reasoning.

For ethical issues, first, the training data for LJP models should include diverse cases, including representative data from different ethnic groups, genders, socioeconomic backgrounds, etc. During the training phase, the model can be made less biased by balancing the data to reduce the inequality factors. It is also important to introduce a specialised algorithm to detect bias in the training data and model output. Researchers need to detect bias and correct it in a timely manner.

Last but not least, the LJP model cannot ignore the opinions of practitioners such as judges and lawyers in its practical application. Researchers can pay more attention to the specific manifestations in legal practice to inspire the construction of the model.

4 Conclusion

This paper reviews the current state of the art and development of LJP-related tasks, and discusses what researchers can do in the future. Current models have made significant progress in improving the accuracy and consistency of judgment prediction, and have shown considerable potential and application prospects, but current technology still faces many challenges.

Interpretability, knowledge modelling, and legal reasoning are fundamental tasks that LJP models can serve in real-world legal work, and ethical issues are always a threshold that cannot be bypassed. Existing methods are advancing the resolution of these issues, but researchers need to make more efforts.

In the future, researchers need to combine more diverse and in-depth methods to focus on solving related basic problems. At the same time, they should try to develop and establish high-quality, large-scale data sets and continuously expand them to ensure that model development and training can keep pace with the times. In addition, techniques such as multi-domain pre-training, multi-task learning, or hybrid expert models can be used to reduce the training cost of models and increase their generalisation ability.

It is also important to note that LJP models should play as a supporting role to help the legal system, rather than attempting to supplant it. For example, they can be employed by parties or judges to refer to the outcome of a judgment, reducing the time required for individual cases. The review of the model's prediction results should never be relaxed, and relevant experts should intervene in a timely manner if necessary.

References

1. L. Anderlini, F. Leonardo, and R. Alessandro. "Legal efficiency and consistency." *European Economic Review* 121 (2020).
2. Y. Wang, et al. "Equality before the law: legal judgment consistency analysis for fairness." *arXiv preprint arXiv:2103.13868* (2021).
3. N. Xu, et al. "Distinguish confusing law articles for legal judgment prediction." *arXiv preprint arXiv:2004.02557* (2020).
4. Yue, Linan, et al. "Neurjudge: A circumstance-aware neural framework for legal judgment prediction." *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. (2021).
5. S. Gururangan, et al. "Don't stop pretraining: Adapt language models to domains and tasks." *arXiv preprint arXiv:2004.10964* (2020).
6. I. Chalkidis, et al. "LEGAL-BERT: The muppets straight out of law school." *arXiv preprint arXiv:2010.02559* (2020).
7. S. B. Majumder, D. Das. Rhetorical role labelling for legal judgments using ROBERTA. *FIRE (Working Notes)*, (2020).
8. H. Zhong, et al. How does NLP benefit legal system: a summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*, (2020).
9. S. Hochreiter, and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735-1780, (1997).
10. R. Johnson, and T. Zhang. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (2017).

11. J. Devlin. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, (2018).
12. V. Tran, M. L. Nguyen, and K. Satoh. Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, (2019).
13. V. Tran, et al. Encoded summarization: summarizing documents into continuous vector space for legal case retrieval. *Artificial Intelligence and Law*, 28:441-467, (2020).
14. A. Askari, et al. Combining lexical and neural retrieval with Longformer-based summarization for effective case law retrieval. *DESIRES*, (2021).
15. H. Chen, et al. "Knowledge is power: understanding causality makes legal judgment prediction models more generalizable and robust." *arXiv preprint arXiv:2211.03046* (2022).
16. X. Liu, et al. "Everything has a cause: Leveraging causal inference in legal text analysis." *arXiv preprint arXiv:2104.09420* (2021).
17. Y. Shao, et al. "BERT-PLI: Modeling paragraph-level interactions for legal case retrieval." *IJCAI*. 2020.
18. H. Li, et al. "Thuir@ coliee 2023: Incorporating structural knowledge into pre-trained language models for legal case retrieval." *arXiv preprint arXiv:2305.06812* (2023).
19. C. Xiao, et al. "Cail2018: A large-scale legal dataset for judgment prediction." *arXiv preprint arXiv:1807.02478* (2018).
20. C. Xiao, et al. "Cail2019-scm: A dataset of similar case matching in legal domain." *arXiv preprint arXiv:1911.08962* (2019).
21. I. Chalkidis, et al. "LexGLUE: A benchmark dataset for legal language understanding in English." *arXiv preprint arXiv:2110.00976* (2021).
22. J. Niklaus, et al. "Multilegalpile: A 689gb multilingual legal corpus." *arXiv preprint arXiv:2306.02069* (2023).
23. I. Chalkidis, et al. "LeXFiles and LegalLAMA: Facilitating English multinational legal language model development." arXiv preprint arXiv:2305.07507 (2023).
24. L. Zheng, et al. "When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings." *Proceedings of the eighteenth international conference on artificial intelligence and law*. (2021).