Automatic Discovery of Controversial Legal Judgments by an Entropy-Based Measurement

1st Jing Zhou
Center for Applied Statistics and School of Statistics
Renmin University of China, China
jing.zhou@ruc.edu.cn

2nd Shan Leng
Department of Statistics
University of Wisconsin-Madison, USA
shan.leng@wisc.edu

3rd Fang Wang*
Data Science Institute
Shandong University, China
wangfang226@sdu.edu.cn

4th Hansheng Wang
Guanghua School of Management
Peking University, China
hansheng@pku.edu.cn

***Abstract*—The judgment of controversial cases has always been an important judicial issue, but it is not easy to discover them in practice. In this paper, based on 1,361,354 legal instruments data collected from China Judgments Online, we adopt a deep learning framework to classify 147 different kinds of crimes. The proposed method has three critical steps: 1) We adopt a deep learning model to predict crime categorization; 2) With the trained model, each case is given a score vector which represents the probability that it belongs to each crime; 3) With the probability score, we develop an entropy-based index to measure the controversy of each case. We find that the larger the entropy, the more inconsistent the result given by the model based on the first instance judgment. To verify the proposed entropy measure, we provide 1) two-sided evidence based on second instance judgments; 2) comparison with some baseline models. Both confirm the practical usefulness of the entropy measure. Our results indicate that the proposed framework has an ability to discover potentially controversial cases. It should be noted that the goal of this study is not to substitute the model result for the judge's decision, but to provide a guiding reference for the judicial practice of sentencing.**

***Index Terms*—legal judgment prediction, controversial case discovery, entropy-based measure**

I. Introduction

The criminal law is one of the most important foundations of a nation's legal system. It grants the most powerful protection to every person's right to access to justice. Good practice of criminal law is beneficial to national security, social stability, economic development, and tranquility. Otherwise, justice can scarcely be realized in practice. However, enforcing criminal law is not a trivial task. This is particularly true for a country with a huge population and a large number of legal cases like China. According to the statistics from the Supreme People's Court, 1.12 million first instance cases were judged by courts in 2020 [1]. Many of those went through a second instance (about 3.73%). This suggests that a large amount of litigation was not terminated (no appeal or protest). These are the controversial cases, that is, litigation where judges might hold very different opinions. Ideally, controversial cases would not occur because different judges should come to the same judgment. However, since China's legal system is statutory, judges often have different opinions on the same or similar cases due to the ambiguity or uncertainty of written legal provisions, which leads to controversial rulings.

The discovery of controversial cases is critical to legal system development. Controversial cases are misjudged in the first place sometimes. To bring them to the attention of the highest-level judge can maximize their likelihood to be rectified, leading to better justice in practice. Even for correctly judged controversial cases, it is still of great importance to bring them to the attention of legal practitioners so that cases of similar types can be better judged in the future. But the identification of controversial cases is not easy. Consider, for example, Among the 1.12 million criminal first instance cases judged in China in 2020, a nontrivial number of cases warrant attention. The question is how to discover them.

We propose a deep learning framework for the automatic discovery of controversial cases. The proposed research framework is based on two components. The first is the database. As mentioned before, since 1996, every trial case in China has been documented in text format, and most of these are published on China Judgments Online. We developed a database of 1,361,354 trial records in text documents, covering 268 kinds of legal judgments occurring in China in 2018. A text document usually includes details such as the facts, the court's opinion, and the verdict. This constitutes the data foundation. The second basis is deep learning models. Since the trial cases are recorded in natural language, the capability to understand natural language is of great importance. In this regard, we benefit from development in deep learning related natural language processing skills. A number of important models are studied and further modified so that they can work better for controversial cases discovery.

We are not the first to use deep learning models to study legal problems. Various researchers have used this technique

DOI: 10.18293/SEKE23-035
*Corresponding author
[1]http://www.court.gov.cn/zixun-xiangqing-290831.html

to explore the issues of legal judgment prediction (LJP) [1], [10]. Some studies try to extract criminal elements or domain concepts from legal documents to predict legal results [3], [8]. For example, [6] investigated several discriminative legal attributes to improve the accuracy of low-frequency charges. Other studies focus on jointly modeling multiple subtasks by some novel multi-task legal judgment prediction frameworks in LJP [15], [16]. [17] proposed a TopJudge model, which made use of the relationship between subtasks in LJP. [12] used a multiscale attention mechanism in charge prediction with multi-defendants. Another group of researchers has focused on similar case matching [14]. They concern the finding of pairs of similar past cases. Typical research topics are information extraction [7] and similarity calculation [14]. To calculate similarity at the semantic level, information such as citations [13] and legal concepts [11] are uesed. Recently, [2] proposed a heterogeneous graph embedding method for Chinese legal document similarity measure. Other issues include legal question answering and legal assistant systems. [18] provided a detailed literature review of the use of deep learning methods for legal problems.

To summarize, we highlight the following contributions. The quantitative relationship between facts and crimes is first extracted by a deep learning model. Second, based on model results, we propose an entropy-based index to measure the controversy of cases, leading to the discovery of controversial cases, the analysis of which can provide a basis for justice, as well as guide the evolution of criminal law and theory.

II. Legal judgement data

A. Data description

The data were collected from the China Judgments Online website[2], which contains the public legal instruments of criminal cases occurring in mainland China in 2018. There are 1,361,354 legal instruments covering 268 types of crimes. Each legal instrument contains various variables, where the court investigation is the most important information. It represents the opinions of the judge, which are supported by facts and evidence. The court investigation includes the constitution of the crime and circumstances for sentencing that must be considered at the trial. For example, the crime of murder must include elements such as intention, action, results, and capacity of criminal responsibility. Sentencing must consider surrender, confession, and other circumstances, which should be clearly written. To this end, we extract two key parts from the court investigation, which consist of text content beginning with "verified by trial" and "this court held that." We aim to find facts verified by the court from the first part of the text, and information about crime constitution and sentencing circumstances from the second part.

B. Data preprocessing

The original dataset is noisy, and preprocessing is required. We follow the next four steps to exclude some samples. First, we exclude legal instruments that do not contain the final judgements. Second, we focus on first instance cases, i.e., the first trial for a case. Third, we eliminate legal instruments that involve multiple crimes or defendants. Finally, crimes with very small sample sizes (e.g., less than 30) are also exlcuded. The final sample used for crime classification contained 731,454 legal instruments, covering 147 crimes.

[2]https://wenshu.court.gov.cn/

III. Deep learning models for crime classification

A. RNN and LSTM models

Before introducing the model structure, we formulate the crime classification problem as follows. Let X_i be the description of the ith judgment, $1 \leq i \leq N$, where N is the number of judgments, which can be represented by a word sequence as $X_i = \{X_{it} : 1 \leq t \leq T_i\}$, where T_i is the length of X_i, and X_{it} is a word generated from a fixed vocabulary W. Recall that we extracted text content beginning with "verified by trial" and "this court held that" as our target for analysis. In this case, the vocabulary is constructed as all the unique words obtained using word segmentation for that target content. Next, let Y_i be the associated class label (i.e., the crime type). Then, using a deep learning model, we wish to predict Y_i by X_i. Remarkably, fact descriptions X_i might have different lengths as a word sequence. So we calculate the length of each X_i and their percentiles, and find that about 98%-99% of the text information in the original content can be retained if the maximum length of the word sequence is set to 1,000. Therefore, we use this as the maximum length of a word sequence and fix T_i at 1,000. For sequences with T_i less than 1,000, we do padding to achieve that length.

We next consider how to construct a classical model based on X_i. Theoretically, we should map each keyword in the vocabulary to a vector with a high dimension d [9]. We choose $d = 128$ and obtain a sequence of vectors $Z = f(X_{it})$ for some mapping function. We consider how to predict $X_{i(t+1)}$ by incorporating information from both current text and historical states such that the semantic information from the entire fact description up to the tth keyword can be fully used to predict $X_{i(t+1)}$. We hope to establish a functional relationship between X_{it} and Y_i. Hence an RNN [4] model is constructed, with eight layers: one input layer of text content with dimension 1,000; one embedding layer with 128 hidden nodes; one simple RNN layer with 64 hidden nodes; one global max pooling layer; two dropout layers; and two fully connected layers with dimensions of 512 and 147 respectively. This model has 57,047,123 parameters in total.

The LSTM model is a more improved model to balance long- and short-term dependencies [5] compared to RNN. Our LSTM structure is similar to that of an RNN, with the RNN layer replaced by an LSTM layer with 64 hidden nodes. This gives the LSTM model 57,084,179 unknown parameters.

B. Training process and results

As for training of the proposed RNN and LSTM models, we randomly divide the data into training and validation sets at an 80:20 ratio. To minimize the loss function, a standard

mini-batch gradient descent algorithm is utilized with a mini-batch size of 256. The learning rate is determined as 0.001 and the Adam optimization algorithm is applied.

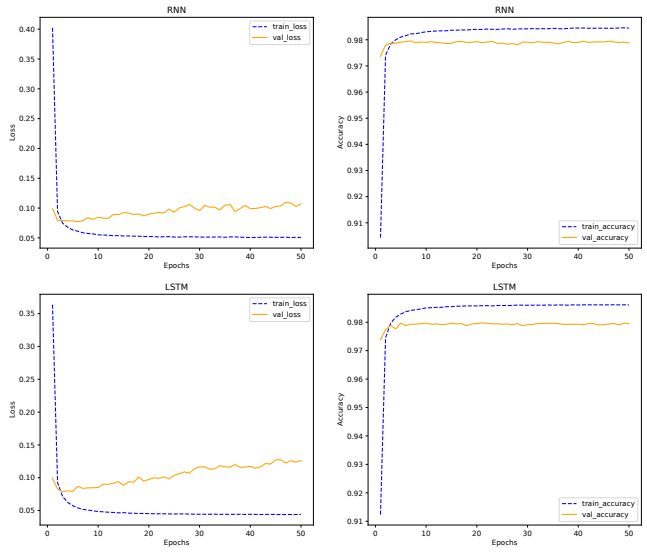

Fig. 1. Left panel: loss curves on training and validation sets for RNN and LSTM models; Right panel: accuracy curves on training and validation sets for RNN and LSTM models.

To obtain stable results, we train each model for 50 epochs. The loss curves on the training and validation sets for the two models are presented in the left panel of Figure 1, and the prediction accuracy curves are shown in the right panel. The final prediction accuracy on the validation set is about 0.9787 for RNN and 0.9795 for LSTM. It can be seen that there is little difference between the two models in terms of accuracy. This suggests that deep learning models have a certain ability for crime classification.

C. An entropy-based measurement

To define an appropriate measure for the discovery of controversial cases, we consider two representative cases. One is the crime of illegally transferring or reselling land use rights. For this case, the top five crimes predicted by both RNN and LSTM are illegally transferring or reselling land use rights; illegal acquisition, transportation of illegal logging, and deforestation; purchasing abducted women and children; privately dividing state assets; and destroying computer information systems. Their predicted probabilities are respectively 20%, 10%, 7%, 6%, and 5%. As one can see, both RNN and LSTM are confused about which crime it should be classified as, because no crime's predicted probability dominates the others. This seems to be a strong indication of controversial cases, at least for RNN and LSTM models. In contrast, both RNN and LSTM correctly predict the crime of illegal medical practice with nearly 100% probability, strongly indicating that both models are affirmative about this crime, which therefore is less likely to be controversial.

From the above discussion, we find that if the predicted probability are concentrated on one category, then this is a case of clear judgment, and little controversy should be found. In contrast, a controversial case should involve at least two crime types, and should confuse both algorithms. Accordingly, their predicted probabilities should be comparable, i.e., no crime type's predicted probability dominates the others. Therefore, a measure to quantify how the predicted probability is distributed across different crime types might be useful to discover controversial cases. This suggests an entropy-type measure for the ith judgment, $H_i = -\sum_{r=1}^{R} p_{ir} \cdot \log(p_{ir})$, where p_{ir} is the probability that the ith judgment belongs to the rth crime type. Based on this, we can conclude that the larger the entropy value, the more dispersed the predicted probability distribution, and the more likely that a judgment will be controversial.

We next apply this measure to the validation set, which leads to a total of 146,291 entropy scores. All the judgments are sorted by descending entropy scores, and are divided into 10 equal-sized groups. For each group, we report the misclassification rates of the RNN and LSTM models in Figure 2, from which we find that the first group has the largest misclassification rate, which is 20.06% for RNN and 19.02% for LSTM. This is expected because cases with large entropy scores should be the most difficult to predict. Accordingly, these cases are the most confusing for both deep learning algorithms and human experts.

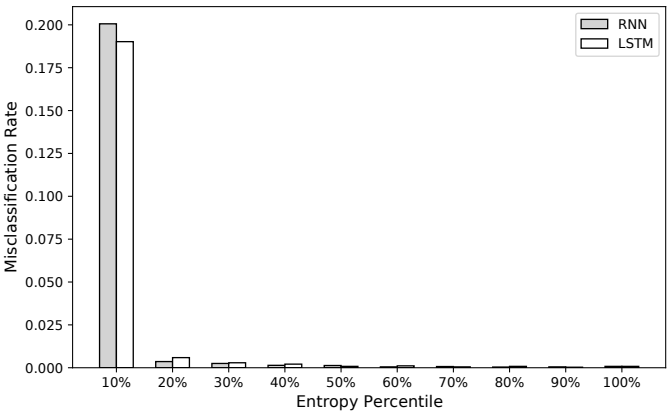

Fig. 2. Bar plot of misclassification rate for RNN and LSTM models according to different entropy percentiles

D. Comparison with other methods

To further verify the practical usefulness of the proposed entropy-based measure, we first present an analysis based on second instance judgments, i.e., second trials by a higher court based on an appeal or a prosecutor's protest. Such a verdict or judgment is a strong legal testimony, indicating that the judgment of the first instance might be controversial. We expect that if a first instance judgment with large entropy value is controversial, then its second instance is more likely to rescind the original judgment and remand the case to the original court or directly change the conviction. Specifically, we present two-sided evidence to support the proposed entropy

score measure. First, we calculate the number of cases with an appeal or protest in the first group, as shown in Figure 2, which is 275, accounting for 10.24% of the judgments in this group. However, the corresponding proportion is only 3.73% in the whole dataset. Second, for those 275 second instances, the number of cases remanded to the original court for retrial, or whose original conviction is directly changed, is 248, accounting for 90.18% of the second instances, while in the whole dataset this ratio is only 13.1%. Both confirm that the proposed entropy measure should be of great practical use.

Moreover, we compare our method with some baselines. An intuitive baseline could be a model that directly predicts whether a case is controversial. So a binary classifier predicting whether a case is controversial or not is developed. Such a classifier can be trained over the text of cases regarding whether they are appealed against or not. We use LSTM, RNN, Bert, and Transformer to train the binary classifier respectively. For each baseline, we apply the best model to the validation set to compute a probability score (i.e., how likely it is appealed against) for each judgement. All the judgements are sorted by descending probability scores, and divided into 10 equal-sized groups. For each group, we calculate the appeal rate. Table I displays the results in the top group for each baseline. We can see that the best baseline is Transformer with the largest appeal rate of 18.6%, which is lower than the misclassification rate obtained by our method (i.e., 20.06%). These direct binary classifiers do not perform as well as the proposed method. This indicates that simple binary classification can not identify the controversy of cases efficiently and further confirms the practicability of the entropy-based measure.

TABLE I
APPEAL RATE OF THE TOP 10% PROBABILITY SCORE GROUP

Model	LSTM	RNN	Bert	Transformer
Appeal Rate	11.14%	12.59%	5.00%	18.60%

IV. CONCLUSION

Based on a large amount of judicial document data, we proposed an entropy-based measure to discover controversial cases. Our results can assist judges to identify potentially controversial cases. However, there are still some limitations. First, we removed charges with small sample sizes during preprocessing, which may be treated as rare events worthy a separate study. Second, we only considered RNN and LSTM models, to the exclusion of more advanced deep learning models. This should be another direction for future research. Finally, we did not consider the problem of sample imbalance, which may bring some issues in the classification. It would be of great interest to solve this problem.

V. ACKNOWLEDGEMENT

Zhou's research is supported in part by the National Natural Science Foundation of China (No. 72171226, 11971504), the Beijing Municipal Social Science Foundation (No. 19GLC052), the Fundamental Research Funds for the Central Universities,and the Research Funds of Renmin University of China, No.21XNA027. Fang's research is supported in part by the National Natural Science Foundation of China (No. T2293773), Hansheng's research is partially supported by the National Natural Science Foundation of China (No. 12271012, 11831008) and the Open Research Fund of the Key Laboratory of Advanced Theory and Application in Statistics and Data Science (KLATASDS-MOE-ECNUKLATASDS2101).

REFERENCES

[1] Ahmad, S., Asghar, M., Alotaibi, F., and Al-Otaibi, Y. (2022). A hybrid CNN+BILSTM deep learning-based DSS for efficient prediction of judicial case decisions. *Expert Systems with Applications*, 209, 118318.

[2] Bi, S., Ali, Z., Wang, M., Wu, T., and Qi, G. (2022). Learning heterogeneous graph embedding for Chinese legal document similarity. *Knowledge-Based Systems*, 250, 109046.

[3] Chen, H., Wu, L., Chen, J., Lu, W., Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Information Processing and Management*, 59(2), 102798.

[4] Elman, J.L. (1990) Finding structure in time. *Cognitive science*, 14(2), 179-211.

[5] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

[6] Hu, Z., Li, X., Tu, C., Liu, Z., and Sun, M. (2018). Few-Shot Charge Prediction with Discriminative Legal Attributes. In *Proceedings of COLING*.

[7] Liu, B., Wu, Y., Zhang, F., Liu, Y., Wang, Z., Li, C., Ma, S. (2022). Query Generation and Buffer Mechanism: Towards a better conversational agent for legal case retrieval. *Information Processing and Management*, 59(5), 103051.

[8] Lyu, Y., Wang, Z., Ren, Z., Ren, P., Chen, Z., Liu, X., Song, H. (2022). Improving legal judgment prediction through reinforced criminal element extraction. *Information Processing and Management*, 59(1), 102780.

[9] Mikolov T., Chen K., Corrado G., et al. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[10] Mumcuoğlu, E., Öztürk, C. E., Ozaktas, H. M., and Koç, A. (2021). Natural language processing in law: Prediction of outcomes in the higher courts of turkey. *Information Processing and Management*, 58(5), 102684.

[11] Opijnen, M., and Santos, C. (2017). On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law*, 25(1), 65-87.

[12] Pan, S., Lu, T., Gu, N., Zhang, H., and Xu, C. (2019). Charge prediction for multidefendant cases with multi-scale attentions. In *CCF Conference on Computer Supported Cooperative Work and Social Computing*, 766-777. Springer.

[13] Raghav, K., Reddy P., Reddy V. (2016). Analyzing the extraction of relevant legal judgments using paragraph-level and citation information. *Artificial Intelligence for Justice*, 30.

[14] Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Zhang, T., Han, X., Wang, H, and Xu, J. (2019). Cail2019-scm: A dataset of similar case matching in legal domain. *arXiv:1911.08962*.

[15] Yang, S., Tong, S., Zhu, G., Cao, J., Wang, Y., Xue, Z., Sun, H., and Wen, Y. (2022). MVE-FLK: A multi-task legal judgment prediction via multi-view encoder fusing legal keywords. *Knowledge-Based Systems*, 239(5), 107960.

[16] Yao, F., Sun, X., Yu, H., Yang, Y., Zhang, W., and Fu, K. (2020). ECHR-OD: Gated hierarchical multi-task learning network for judicial decision prediction. *Neurocomputing*, 411(21), 313-326.

[17] Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., and Sun, M. (2018). Legal judgment prediction via topological learning. In *Proceedings of EMNLP*.

[18] Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., and Sun, M. (2020). How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. *arXiv:2004.12158v5*.