Open the Black Box of AI:

Saliency Map of DUI Sentencing and Legal XAI

Hsuan-Lei SHAO[*1] Wei-Hsin WANG[*2] Sieh-Chuen HUANG[*2] Kuan-Ling SHEN[*2]

[*1] Dept. of East Asian Studies, National Taiwan Normal University
[*2] College of Law, National Taiwan University

This article is to construct an AI model to predict drunk driving (DUI) sentencing cases in Taiwanese Judgments. We provide a textCNN model for the four-classification sentencing range with 72% accuracy and make it explainable AI (XAI) by visualized saliency maps. The method is to observe the" saliency value" by the final output differential by every word vector. We succeed in establishing a model which can input Chinese words and pick up" salient" words. More specifically speaking, phrases such as" alcohol rate in his/her breath," "highly dangerous," and "recidivist" have higher saliency values. They happen to echo the provisions of the Criminal Code (the DUI article §185-3 I, the sentencing article §57, and the recidivist §47). The result of this paper can be coherent with the legal domain knowledge, being the first step in the XAI approach to legal analytics.

1. Research Background

"Drunk driving," aka driving under the influence of alcohol or drug (hereinafter referred to as DUI), is illegal in many countries, including Taiwan, because DUI can lead to severe accidents, injuries, and fatalities on the road. Precisely speaking, there were 160 deaths immediately and 4,778 injuries because of DUI in 2021 Taiwan. It is not only a judicial problem, but also a serious social problem in Taiwan. The DUI behavior can let the drivers to prison themselves. We penalize DUI behavior by criminal law (Criminal Code of the Republic of China, hereinafter referred to as "CCRC"). Precisely speaking, in Chapter 11, Offenses Against Public Safety, Article 185-3[1]. That can be a juridical case that involves a bunch of human labor, such as judges, prosecutors, lawyers, court clerks, and mediators. There were 59,454 DUI cases in the court in 2021. Therefore, this study is to construct a model to predict the sentencing of a DUI case. This automated model can alleviate the labor costs incurred by many DUI cases. In addition, this study will provide a different approach to dealing with Chinese text data.

On the other hand, we will try to let the model "explain" the reason for the prediction. It is the so-called explainable AI (hereinafter referred to as XAI). Especially in the legal AI field, researchers and legal professionals will not readily accept a "black-box" system into legal practice. They want to know why the AI model works in the legal field.

2. Literature Review and Research Design

There have been a couple of studies on drunk driving in the field of empirical legal studies, discussing whether it is possible to reduce the number of drunk driving cases by increasing sentences. In the case of Chile, raising the penalty may have a short-term deterrent effect, but the long-term caseload will gradually return to its original level [2]. The other approach is to study sentencing factors. For example, the study in Colorado, U.S. involves the use of criminal record and blood alcohol levels as the benchmark [3]. This is also the present system in Taiwan.

Another related literature is "Prediction Model for Drunk Driving Sentencing: Applying TextCNN to Chinese Judgment Texts" [4] This research attempts to train a deep-learning model to predict sentences by inputting the section of "recidivist and facts of the judgment." This paper established a pre-processing method on unstructured Chinese judgment texts without word segmentation and providing a TextCNN model reached a 73% accuracy rate in four-category sentencing prediction. This model (aka model 0) as a baseline that we can compare with this new saliency model. We expect it opens the possibility of applying different machine learning techniques to legal texts.

3. Data Sources and Research Design

3.1 Data Source and Input Data Processing

We used Taiwan's DUI judge text dataset by Central News Agency (the CNA dataset)[5]. In the dataset, judgments were labeled according to judgment date, jurisdiction, whether it is a recidivist, whether the punishment may be commuted to a fine, perpetrators' education level, vehicle type, alcohol concentration, sentence, and fine. We think the "fact" part of the judgment is the most concerning part of sentencing.

Moreover, because the "recidivist" can be an important factor on sentencing (Article 47 of the CCRC), we add the "recidivist" label and remove too short judgments which have insufficient information to improve the performance. After the above data preprocessing, the total number of effective samples reduces to 33,129, with an average of 468 words, a standard deviation of 235 words, and a third quartile of

Contact: Corresponding Author: Sieh-Chuen HUANG, Prof., National Taiwan U., Orcid: 0000-0003-3571-5236, schhuang@ntu.edu.tw

522 words. The comparison of samples before and after processing is shown in Table 1. The reason for categorizing sentences into four types will be elaborated on in the next section.

Table 1: Comparison of sample quantities before and after processing

Sentence-Four Classifications	Original sample size	Sample size after processing
<=2 months	15,263 (33.0%)	10,998 (33.2%)
>2 months, <=3 months	13,948 (30.2%)	10,921 (33.0%)
>3 months, <=6 months	15,359 (33.2%)	9,926 (30.0%)
>6 months	1,551 (3.4%)	1,284 (3.8%)
Total samples	46,228 (100%)	33,129 (100%)

Finally, we make up the datasets. Of the total 33,129 samples, we choose 80% for training, 10% for validation, and 10% for testing (comprising 3,312 samples).

3.2 Research Design I: Sentence prediction as Text Classification

In this paper, we will three textCNN-based model [6], input by Chinese words (model 1) or Chinese characters(model 2 and 3). Basically those models contained one convolution layer, one hidden layer, and a softmax layer, achieving about 73% accuracy. It is different when we explain by taking saliency of the Chinese "characters" (model 1), or Chinese "words" (model 2 and 3). In our intuition, it is easy to explain and understand by "words" meaning. Therefore, we design the process at the beginning with "torchtext." We use the SpaCy package to word-segment the text, then build a dataset to import data, build a vocabulary to embed words into word vectors, and finally build an iterator to output to the model for training. First, splicing the previously constructed word vectors into a matrix of 10002*300 dimensions and inputting it into the next convolutional layer, the size of the convolutional kernel is [2,3,4,5]*300 dimensions, outputting 100 channels as model 2, [2,3,4]*300 dimensions outputting 250 channels as model 3. Then the pooling layer takes the maximum value of the results of each channel (MaxPooling). The output results of the pooling layer are concatenated into the fully connected layer and connected to softmax for four classifications. Get the scores of train and valid, and find the best result (test score) according to the loss value. Model 2 and model 3 can fit n-gram feature extraction.

Finally, we calculate the word vector, normalize the whole and obtain the strength of the word, divide the sentence into sentences according to the punctuation marks, and then draw a picture. According to the graph, we can quickly understand which words will affect the judgment result.

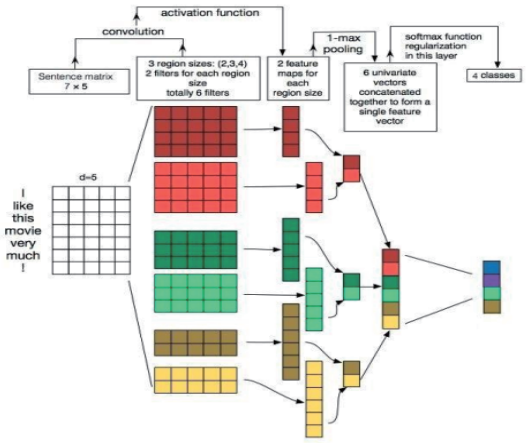


Figure 1: Text Classification Using Deep Learning [7]

3.3 Research Design II: Word Importance by Saliency Map

We will figure out the importance of each Chinese word by the idea of a "saliency map," which is a type of image analysis that highlights areas of an image or video that stands out from the rest. It is used in computer vision applications to identify the most important or salient regions of an image that a human would focus on by assigning a value to each pixel that indicates how important the pixel is to the overall scene.

As well as highlighting areas of an image, we can "highlight" the important word, which could affect the classification score "confidence." After preprocessing, we can use a word vector of Chinese words embedded by the "SpaCy" package in 300 dimensions as the input. After we have finished training the model, we can have each output (the classification and its confidence score) of the test data. Then we define "saliency" as the gradient to measure how a change in a single-word vector dimension weight or bias affects the overall cost of the network. In order to be visible, we normalize by its range in each dimension.Then, we assume an average value of saliency can present each word' s importance.

A sentence can be easier to visualize than the whole text corpus. We split by Chinese punctuation marks: Exclamation Mark, Full Stop, Question Mark and Semicolon. Then output them as bar plots. The entire process can be figured in figure 2. below.

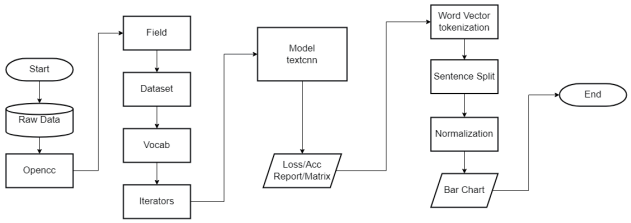


Figure 2: Research Process

Finally, we plot a saliency map of Chinese text classification in the CS/IE technology on the corpus of the DUI judg-

ments. These saliency maps can point out the "keywords" that affect the final prediction more; they are legally explainable AI.

4. Research Result

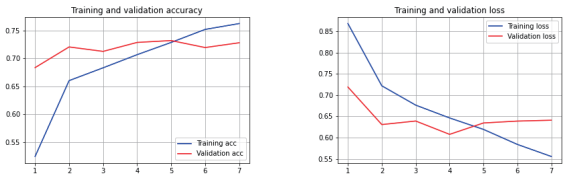
4.1 Sentence Prediction Model's Efficiency and Evaluation

In this paper, we compared three kinds of models. They are all input "recidivist+crime fact." Model 1 is by Chinese character and one kernel with a window-size of five, then the model 2 and model 3, we try to finetune the different kernel combinations and output channel. We also enlarged the word embedding this time.

Table 2: Comparing of model structure.

	segmentation	kernel	channel	embedding
model 1	character	5	256	64
model 2	words	[5,4,3,2]	100	300
model 3	words	[4,3,2]	250	300

Generally speaking, there are only tiny differences between model 2 and model 3. In the training stage, we observe that both the accuracy and loss values of Model 2 and Model 3 can converge into steady. However, not as well as the training dataset, validation datasets cannot converge very well and even go to little overfitting. Therefore, we are early-stopped in the best performance epoch. The model 2 and model 3 have a similar training process as figure 3.



(a) accuracy (b) loss

Figure 3: model 2 training processing log

In the evaluation stage, the performance of models are almost as good as each other. We also can evaluate those three models. Though model 1 can be few better than others.

Table 3: evaluation results of models.

Sentences	model 1 test dataset			model 2 test dataset			model 3 test dataset			support
	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score	
n<=2	0.78	0.87	0.82	0.77	0.74	0.75	0.74	0.76	0.75	1070
2<n<=3	0.64	0.59	0.61	0.64	0.52	0.57	0.64	0.51	0.56	1022
3<n<=6	0.77	0.72	0.74	0.74	0.89	0.81	0.75	0.86	0.80	1101
n>6	0.69	0.84	0.76	0.71	0.76	0.73	0.69	0.82	0.75	119
Macro avg	0.72	0.75	0.73	0.72	0.73	0.72	0.71	0.74	0.72	3312
Weighted avg	0.73	0.73	0.73	0.72	0.72	0.71	0.71	0.72	0.71	3312

In conclusion, both models can work well and are very similar in training processing and performance. All of them can converge quickly in three of four epochs. It might be because the DUI cases have a similar structure. However, the validation (test) performance could be better than the training dataset; they still contain about 70% and cannot gradient descent continuously. We cannot know how to come right now. If only from an efficiency of view, this comparison encourages us to use a simple text CNN model rather than a complex one. Our models show sacrifice accuracy in the trade of explainability. It proves the rules of their conflict.[8]

4.2 Saliency Maps and their Explianablity

Our primary purpose is to request the model (AI) to explain how they predict the sentence from what words are their "reason." We visualize the saliency of each word. The process is as follows:

Firstly, to calculate the word vector and word segmentation, throw the word vector into the model and return it to the source for backpropagation (differentiation), take the absolute value and convert it into numpy (saliency) for standardized calculation of word strength, and then combine the word segmentation results with word strength Make sentences according to punctuation marks, and finally draw a picture of each sentence according to the strength and participle.

Moreover, we shall provide the main sentence reason in Taiwan, such as the DUI article (§185-3 I), which is mainly by the alcohol rate in his/her body[1] or the article of sentencing range (§57), which mainly by No.8. The seriousness of the offender's obligation violation. 9. The danger or damage caused by the offense. 10. The offender's attitude after committing the offense. Furthermore, with the article of a recidivist (§47), we could try to find how those factors affect saliency maps below. We use model 2 to draw these figures. The fact item is the primary basis for AI to make judgments.

4.2.1 Alcohol Rate in the Body

In figure. 4, "Alcohol Rate" (0.25毫克以上) is the peak of the saliency of this sentence.

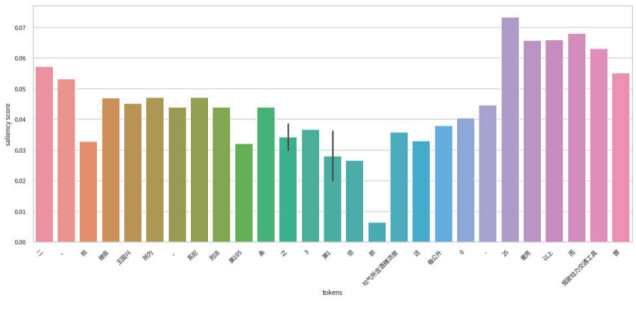


Figure 4: "Alcohol rate" as the most important factor

4.2.2 Sentencing Range

Another example is Figure 5 gives us more information about explainability. Firstly, the alcohol rate(0.53毫克, for the DUI article, §185-3 I) "can be hazardous" (具有高度危險性, for §57 IX) and endanger road safety (危害道路安全, for §57 IX)

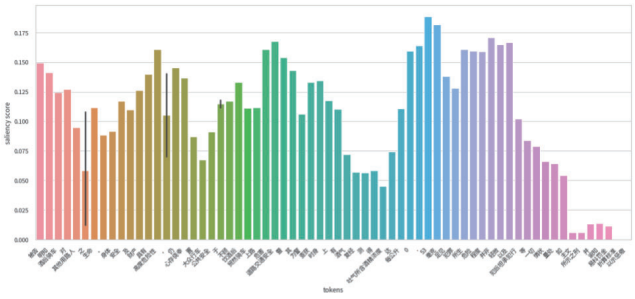


Figure 5: The saliency map of “highly dangerous”

4.2.3 Recidivism

Punishment for a recidivist shall be increased (§ 47). It can be seen in our system. In fig. 6, This person has three times DUI records （曾犯3次酒後駕車）.

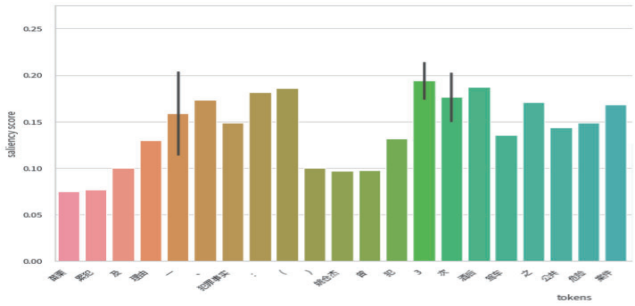


Figure 6: The saliency map of “recidivist”

5. Conclusion: Discuss Explainable AI and its Limitation

We try to make a deep learning model explainable in this paper. Observing the value of the final output softmax by every word vector affecting, we define a ”saliency value” of the appointing word. After collecting those saliency values, we can figure out which word is more important than others.

In our dataset of DUI, we succeed in making a model which can be input by Chinese words. Moreover, pick up ”salient” words. With the relative legal article of the CCRC (the DUI article § 185-3 I, the sentencing article § 57, and the recidivist § 47), we can see those related word saliency values are higher which means the model can tell humans the reason for its prediction, such as ”alcohol rate in his/her breath（酒精濃度）,” highly dangerous” （高度危險性）,”recidivist” （累犯）. The calculation result can be coherent with the legal domain knowledge and specific expert’s common sense.

Therefore, we claim we can glance at the AI black box from a small slit. Especially in the field of legal studies and practice. The judge often only fully believes the AI model with reason. Our research can be a first step of the XAI in empirical legal studies in the AI era.

On the other hand, legal texts (judgments) usually have specific usage. Different judges or different areas can use the same word with different content. It is so-called ”discretion.” So, can or cannot our explanation conclude each style of those ”styles”? It is our next challenge. Finally and philosophically, is it a real ”reason” for the AI model? or just a ”re-explain” by researchers? We could try to approach closer and closer to the truth.

Acknowledgment

Hsuan-Lei Shao, “Knowledge Graph of China Studies: Knowledge Extraction, Graph Database, Knowledge Generation” (MOST 110-2628-H-003-002-MY4, Ministry of Science and Technology) in Taiwan.

Sieh-Chuen Huang, “Digital Court, Legal Tech, and Access to Justice” (MOST 110-2423-H-002-003, Ministry of Science and Technology) in Taiwan.

References

[1] Article 185-3I , the CCRC: A person who drives a motor vehicle in any one of the following circumstances shall be sentenced to imprisonment for not more than three years... 1. the person’s exhalation contains alcohol of 0.25 milligrams per liter or more, or the person’s blood alcohol concentration is 0.05 percent or more.

[2] Garca-Echalar, A., Rau, T. (2020). The Effects of Increasing Penalties in Drunk Driving Laws-Evidence from Chile. International Journal of Environmental Research and Public Health, 17(21), 8103. https://doi.org/10.3390/ijerph17218103

[3] Lange, T. J., Greene, E. (1990). How Judges Sentence DUI Offenders: An Experimental Study. The American Journal of Drug and Alcohol Abuse, 16(1-2), 125—133. https://doi.org/10.3109/00952999009001577

[4] Shao, Hsuan-Lei and Huang, Yu-Ying and Huang, Sieh-Chuen, Prediction Model for Drunk Driving Sentencing: Applying TextCNN to Chinese Judgement Texts (November 14, 2021). https://www.kl.itc.nagoya-u.ac.jp/jurisin2021/, Available at SSRN: https://ssrn.com/abstract=4306412

[5] Ke, H. (2019). Outline of drunk driving in Taiwan seen from 490,000 judgments, CNA English News, https://www.cna.com.tw/project/20190719-drunkdriving/epilogue.html, last accessed 2021/08/10.

[6] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882

[7] Boukil, S., Biniz, M., El Adnani, F., Cherrat, L., El Moutaouakkil, A. E. (2018). Arabic text classification using deep learning technics. International Journal of Grid and Distributed Computing, 11(9), 103-114.

[8] Philipp Hacker et al., Explainable AI under Contract and Tort Law: Legal Incentives and Technical Challenges, Artificial Intelligence and Law (2020).