Classifying Crime Types using Judgment Documents from Social Media

Haoxuan Xu*‖, Zeyu He*‖, Mengfan Shen†, Songning Lai*, Ziqiang Han†§ and Yifan Peng‡§
*School of Information Science and Engineering, Shandong University, Qingdao, China
†School of Political Science and Public Administration, Shandong University, Qingdao, China
‡Department of Population Health Sciences, Weill Cornell Medicine, New York, USA
Email: {202020120237,202000120128,202000120172}@mail.sdu.edu.cn, mengfan@mail.sdu.edu.cn,
ziqiang.han@sdu.edu.cn,yip4002@med.cornell.edu

***Abstract*—The task of determining crime types based on criminal behavior facts has become a very important and meaningful task in social science. But the problem facing the field now is that the data samples themselves are unevenly distributed, due to the nature of the crime itself. At the same time, data sets in the judicial field are less publicly available, and it is not practical to produce large data sets for direct training. This article proposes a new training model to solve this problem through NLP processing methods. We first propose a Crime Fact Data Preprocessing Module (CFDPM), which can balance the defects of uneven data set distribution by generating new samples. Then we use a large open source dataset (CAIL-big) as our pretraining dataset and a small dataset collected by ourselves for Fine-tuning, giving it good generalization ability to unfamiliar small datasets. At the same time, we use the improved Bert model with dynamic masking to improve the model. Experiments show that the proposed method achieves state-of-the-art results on the present dataset. At the same time, the effectiveness of module CFDPM is proved by experiments. This article provides a valuable methodology contribution for classifying social science texts such as criminal behaviors. Extensive experiments on public benchmarks show that the proposed method achieves new state-of-the-art results**

***Index Terms*—BERT, Crime, Chinese, NLP, Judgement Documents**

I. Introduction

Criminal behavior is the behavior exhibited by an offender that perpetuates an illegal act [1]. This behavior can be classified into many types based on factors such as the scope and length of the prediction [2]. In terms of scope, criminal behavior can be classified into macro-prediction and micro-prediction of crime [3], according to the scope of the prediction determined by the predicted objects.

Macro prediction of crime involves using the macro-level crime phenomena and types within a specific time and space to make predictions. Its main purpose is to reveal the trends and dynamics of criminal phenomena in terms of quality and quantity, by providing evidence-based suggestions for crime prevention policy or enhancing the overall efficacy of criminal justice.

Micro-prediction of crime uses scientific techniques to predict the likelihood of an individual's initial or recurring criminal behavior within a specific period in the future. Its main purpose is to make early predictions of individuals' potential initial or recurring criminal behavior under certain spatial and temporal conditions, enabling the selection of the most appropriate form and severity of punishment to achieve particular preventive purposes while administering specific penalties.

This study aims to create a system for categorizing crime types based on behavioral facts associated with criminal activities. Previous research has primarily focused on methods such as native Bayesian, convolutional neural networks (CNN), and long short-term memory networks (LSTM). For example, Almanie et al. used decision tree algorithms and naive Bayesian classification to predict potential crime types [6]. Sivaranjani et al used the KNN and K-means methods to compare the clustering results and find the most suitable crime type [7]. Stalidis et al. employed a CNN-based deep learning framework to predict crime types for multiple open-source datasets [8]. Yi et al. learned the nonlinear correlation mapping of different regional inputs and outputs based on LSTM to use spatial correlation for crime modeling [9]. Although these methods achieved successful outcomes, they have limitations such as high text training costs, low accuracy, and weak generalization ability.

Recently, significant progress has been made in developing large language models (LLMs) that can be applied to various natural language processing tasks in the general and biomedical domains. But studies have yet explored their usage in social science problems. In addition, the impact of data imbalance makes it challenging to directly apply LLMs to this task. To bridge gaps, we propose a transformer-based method with a novel Crime Fact Data Preprocessing Module (CFDPM) that rebalances the training samples through data augmentation.

To validate the effectiveness of the proposed method, we evaluate our method on two datasets: one is a publicly available dataset extracted from non-classified judgment documents of the Supreme Court of China (CAIL-big), and the other is an in-house dataset extracted from Police Briefings from the police department's social media. The experimental results demonstrate that our method outperforms serval baselines and achieves state-of-the-art performance.

§ Corresponding authors
‖ These authors contributed equally to this work and should be considered co-first authors.

Fig. 1. Flow chart of crime type classification.

II. Methods

A. Our Framework

To overcome the problem of uneven distribution of data unique to crime facts, we propose a new module and a fine-tuning approach for wider adaptation to more data sets.

Fig 1 shows our framework. Given a text, we first applied three data augmentation methods (in the Crime Fact Data Preprocessing Module [CFDPM]) to augment the text. We then fine-tuned a BERT-based model [22] with dynamic masking to predict labels for each text. The specific module introduction will be introduced below.

B. Crime Fact Data Preprocessing Module

The data imbalance presented in the datasets [11] , caused by varying probabilities of criminal events, can significantly impact the effectiveness of the model. For example, the top 10 crime types account for 79.0% of cases in CAIL-big, whereas the bottom 10 crime types cover a mere 0.12% of cases. Thus, one prominent challenge is to create a more balanced distribution within the dataset. To address this challenge, we designed a Crime Fact Data Preprocessing Module (CFDPM) for data augmentation and balance.

First, we introduced a sample threshold for each crime type to decide whether to activate CFDPM.

$$X_i^{'} = \begin{cases} \text{Random\_Sampling } (X_i) & \text{if } n_i > K \\ \text{Data\_Augmentation } (X_i) & \text{if } n_i \leq K \end{cases} \quad (1)$$

$X_i^{'}$ represents the processed $i-$ th sample set, X_i represents the i-th sample set of the original data, n_i represents the number of samples in the i-th sample set, and K is the sampling threshold for each crime type. If the total number of samples in a certain type of crime is greater than , then samples are randomly selected from that type. Otherwise, new samples are generated using data augmentation methods.

Three methods are used to generate new sample data: synonym replacement, adding random noise, and back-translation (Fig. 2).

Synonym Replacement. We replace words with their synonyms to generate different training samples [12]. W_{ij} represents the i-th word in the j-th sample set and $W_{ij}^{'}$ represents the word for the same sample after being replaced by a synonym.

$$W_{ij} \Rightarrow W_{ij}^{'} \quad (2)$$

Random noise addition. With probability p, we replace words in the training set with blank characters ' ', to generate new training samples [13]. p represents the probability of being replaced by '_'. This method is suitable for Chinese, which is written without spaces between successful characters and words.

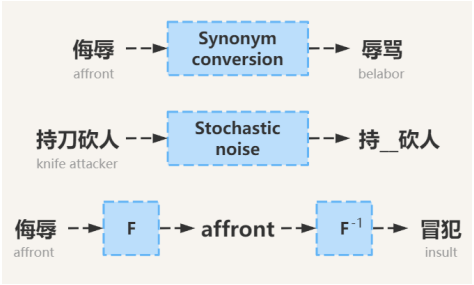

Fig. 2. Three methods used to generate new sample data.

$$W_{ij} = \begin{cases} \text{`\_'} & p \\ w_i & 1-p \end{cases} \quad (3)$$

Back translation. We translate Chinese into English, and then translate it back into the original language to generate new samples [14]. S_{ij} represents the sentence of the j-th sample in the i-th sample set, $S_{ij}^{'}$ represents the sentence after the back translation of the sample. F is the function of translating the original language to the target language, and F^{-1} represents the function of translating the target language to the original language.

$$S_{ij}^{'} = F^{-1}\left(F\left(S_{ij}\right)\right) \quad (4)$$

C. Dynamic Masking

Masking is a widely used operation in deep learning, particularly for pretraining transformer models like BERT. It involves covering a mask layer over the original tensor to select or shield specific elements [17].

The original BERT implementation adopted the static masking approach. BERT performs masking only once during the data preprocessing stage, which means that the same input masks are fed to the model on every single training epoch.

Different from traditional training paradigms, we adopt a dynamic masking method in this task. Dynamic masking does not mask out data in the preprocessing stage, but the masking is performed every time a sequence is fed to the model. In this way, the model sees different versions of the same sentence with masks on different positions [17].

D. Baselines

In this study, we compared our method with four baseline models. **TextCNN** is a simple and fast approach for text classification based on N-grams and softmax [20]. **LADAN** distinguishes confusing law articles by distinguishable features from similar law articles using a graph-based method.extracting [21]. **BERT** is a contextualized word representation model that is pre-trained based on a masked language model, using bidirectional Transformers [24]. We initialized BERT with pre-trained BERT provided by Devlin et al [22]. We then continue to fine-tune the model, using the corpus. **XLNet**

is a pretraining language model based on autoregression and autoencoder [23].

E. Evaluation Criteria

To evaluate our model's prediction of crime types and compare it with other models, we have adopted four evaluation criteria to analyze the strengths and weaknesses of our model and other models. Accuracy (ACC), f1-score (F1), Precision (P) and Recall (R) were respectively used to evaluate our model

$$ACC = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$F1 = 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{8}$$

Among them, the number of false positives, false negatives, true positives and true negatives are recorded as FP, FN, TP, and TN.

III. RESULTS AND DISCUSSION

A. Datasets

We included two benchmark datasets in this study. Table I shows the detailed statistics of the datasets.

TABLE I
STATISTICS OF THE DATASETS.

	Ours	CAIL-big
# Training Cases	3,793	101,619
# Test Cases	542	13,619
# Validation Cases	1,083	26,749
# Charges	48	119

1) CAIL-big: The Supreme People's Court of China published all non-classified judgement documents online to improve the transparency and accountability of criminal justice work[1]. The Chinese of AI and Law challenge task (CAIL) [10] collected and released these documents continually to encourage further explorations on advanced legal intelligence algorithms. In this study, we used the open dataset CAIL-big. Statistics on some of the major offenses in CAIL-big are shown in Table II. We used 150,000 documents for training, and 16,000 documents for testing.

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

TABLE II
MAJOR OFFENSES IN CAIL-BIG

Category	Train	Test
Larceny	31,748	4,618
Dangerous driving	29,388	4,028
Intentional assault	16,939	2,402
Traffic offences	14,182	1,944
Smuggling and manufacturing narcotic drugs	10,173	1,764
Tolerating drugs	4,987	486
Swindling	4,827	468
Cause trouble	2,842	360
Robbery	2,197	331
Credit Card Fraud	1,854	265
Possession of guns and ammunition	1,636	226
Interference with public servant	1,696	129
Possession of drugs	1,562	150
Opening casinos	1,422	172
Conceal the proceeds of crime	1,144	133
Bribery	1,165	97
Illegal denudation	986	131
Gambling	955	131
Vandalism	959	81
forcible seizure	736	84
bribery	719	89

2) Ours: Our own dataset was collected from the Police Briefing from all the police department's social media (Weibo) posts in 2020, which included 5,000 criminal activities and they were manually annotated as training samples for processing the unknown data. We followed the data annotation style of CAIL-big and supervised the annotation of our own dataset.

One challenge in our approach is how to use the CAIL-big dataset to help train the model on our in-house dataset. For this purpose, we need to align the two datasets. The format of the CAIL-big dataset is shown in the following:

- Fact: 犯罪嫌疑人王某... (The defendant Wang...)
- Charges: 过失杀人 (manslaughter)
- Prison Term: 16个月 (16 months)
- Defedant: 王某 (Wang)

As can be seen from the example above, CAIL-big mainly consist of four components: Fact, Charges, Prision Term, and Defedant. (1) **Fact**: We used the crawler to extract facts from China Judicial Documents Network[2]. Here, we limit the word count to 300. (2) **Prison Term and Defedant**: We used the regular expressions to extract prision term and defedant from the text of fact. (3) **Charges**: We first trained a model to classify Charges on CAIL-big data set, and then applied the model to our in-house dataset. Finally, we manually checked each label. In this way, the manual labeling effort can be greatly reduced.

B. Implementation Details

We used LTP as a word segmentation tool to decompose Chinese into words [3]. The pre-trained GloVe was used to initialize word embedding in the neural network, and the maximum encoding length was set to 400.

[2]https://wenshu.court.gov.cn
[3]https://ltp.ai/

In the data augmentation module CFDPM, for the data with a small sample size, we carry out three augmentation methods in a cyclic manner until sufficient samples are generated. Synonym replacement thesaurus uses pyltp replacement thesaurus. The substitution of blank words is the probability that each word has a p is replaced with a blank character. Here we set p to 0.1. For back translation, we called Baidu Translation English translation API interface [4].

During the training process, we employed the Adam optimizer with a learning rate of 2e-3 and a batch size of 64. For models not requiring pre-training, we consistently used 60 epochs. Conversely, for models necessitating a pre-trained model, we utilized 10 epochs.

During the fine-tuning process, we initially trained the network using the large dataset (CAIL-big) and the mentioned hyperparameters. Subsequently, we froze all the other layers of the model, leaving only the final MLP layer for further training. At this stage, the training set should consist of a small, unfamiliar dataset. We utilized our own dataset as mentioned earlier, with a learning rate of 2e-5 and 20 epochs. The Adam optimizer was also selected for fine-tuning.

C. Main Results

To assess the effectiveness of our module and approach, we conducted comparisons between the baseline models and our models on the CAIL-big dataset. We also incorporated our proposed CFDPM module in to some of the baselines for comparative analysis (Table III).

TABLE III
MAIN RESULTS ON THE CAIL-BIG.

Modules	Acc	F1	P	R
TextCNN	86.45	84.75	84.56	84.94
LADAN	87.50	85.9	85.55	86.26
BERT	89.16	87.94	88.21	87.68
XLNet	88.37	87.41	87.60	87.24
TextCNN+CFDPM	**87.23**	**85.84**	**85.93**	**85.76**
LADAN+CFDPM	**88.15**	**86.75**	**86.04**	**87.51**
BERT+CFDPM	**90.23**	**89.07**	**89.11**	**89.04**
XLNet+CFDPM	**89.50**	**88.61**	**89.12**	**88.10**
Ours	**91.60**	**90.14**	**89.87**	**90.42**

Firstly, in comparison to all other models, our model achieved the highest accuracy and F1-score, reaching 91.6 and 90.14, respectively. This is 1.37 higher in accuracy and 1.07 higher in F1-score than the next best models. The experiment demonstrates that our enhanced model can effectively predict crime facts.

Secondly, we carried out a comparative experiment with the proposed CFDPM module. It was observed that the addition of this module led to varying degrees of improvement in accuracies and F1-scores for different models. The average accuracy increased by 0.82, while the average F1-score rose by 0.95.

[4]https://fanyi-api.baidu.com/

D. Impact of dynamic masking

As shown in Table IV, our model performs better than the one without a dynamic masking mechanism.

This approach yields a moderate improvement. The rationale behind choosing this model for the task is our belief that the attention mechanism of the BERT model can delve deeper into the inherent features present in crime fact text. By further integrating the underlying semantic information, we can extract more implicit features that are superior for crime type classification compared to conventional models.

TABLE IV
MODEL PERFORMANCE WITH AND WITHOUT DYNAMIC MASKING.

	ACC	F1
Ours w/o Dyanamic masking	90.34	90.14
Ours	**91.60** (↑1.26)	**91.12** (↑0.98)

E. Impact of Fine-tune

We gathered our own data from Police Briefings posted on various police department social media accounts (Weibo). It's worth noting that the facts in CAIL-big differ slightly from those in social media postings.

We evaluated the feasibility of the fine-tuning strategy using BERT, XLNet, and our own models. The non-fine-tuned model is trained solely on our dataset, whereas our fine-tuning strategy involves initial training on CAIL-big, followed by freezing the other layers and training only the last fully connected layer.

As show in Table V, we observed that after fine-tuning, both accuracy and F1-score significantly improved across all three models. We believe that the differences between the two datasets are not substantial in nature. However, compared to the published dataset, our own dataset is smaller in size, which results in suboptimal precision before fine-tuning due to overfitting. In practice, collecting vast amounts of data can be resource-intensive and may not always be a feasible option. Therefore, for the crime facts task, employing a strategy that involves training on an open dataset and then fine-tuning with one's own dataset is a viable approach.

TABLE V
MODEL WITH AND WITHOUT FINE-TUNING.

Modules	Acc	F1
BERT	78.62	74.15
XLNet	79.21	74.71
Ours	81.40	75.80
BERT+fine-tune	**86.58**	**83.21**
XLNet+fine-tune	**86.74**	**83.68**
Ours+fine-tune	**89.21**	**84.92**

IV. CONCLUSION

This study develops and applies an NLP approach to the task of crime type classification in judgment documents and social

media posts. We proposed a new module named CFDPM for data augmentation to address data imbalance, and proposed to adopt a Dynamic Masking mechanism to replace BERT's static masking paradigm. Through experiments, our proposed model shows superior performances over several baseline methods on two benchmark datasets. We further proved the effectiveness of our CFDPM module and the Dynamic Masking mechanism through ablation studies.

In the future, we propose to dive deep into the domain adaptation issues among different language styles and plan to improve model generalizability.ability [25] of our models to adapt to different styles of datasets.

REFERENCES

[1] Jeffery C R. Criminal behavior and learning theory[J]. Contemporary Masters in Criminology, 1995: 175-186.

[2] Gottfredson D M. Prediction and classification in criminal justice decision making[J]. Crime and justice, 1987, 9: 1-20J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Douglas J E, Burgess A W, Burgess A G, et al. Crime classification manual: A standard system for investigating and classifying violent crime[M]. John Wiley & Sons, 2013.K. Elissa, "Title of paper if known," unpublished.

[4] Deng J, Zhou J, Sun H, et al. Cold: A benchmark for chinese offensive language detection[J]. arXiv preprint arXiv:2201.06025, 2022.

[5] Lai S, Hu X, Han J, et al. Predicting Lysine Phosphoglycerylation Sites using Bidirectional Encoder Representations with Transformers& Protein Feature Extraction and Selection[C]//2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE, 2022: 1-6.

[6] Almanie T, Mirza R, Lor E. Crime prediction based on crime types and using spatial and temporal criminal hotspots[J]. arXiv preprint arXiv:1508.02050, 2015.

[7] Sivaranjani S, Sivakumari S, Aasha M. Crime prediction and forecasting in Tamilnadu using clustering approaches[C]//2016 International Conference on Emerging Technological Trends (ICETT). IEEE, 2016: 1-6.

[8] Stalidis P. Semertzidis T, Daras P. Examining deep learning architectures for crime classification and prediction[J]. Forecasting. 2021, 3(4): 741-762.

[9] Yi F, Yu Z, Zhuang F, et al. Neural Network based Continuous Conditional Random Field for Fine-grained Crime Prediction[C]//IJCAI. 2019: 4157-4163.

[10] Xiao C, Zhong H, Guo Z, et al. Cail2018: A large-scale legal dataset for judgment prediction[J]. arXiv preprint arXiv:1807.02478, 2018.

[11] Liu P, Wang X, Xiang C, et al. A survey of text data augmentation[C]//2020 International Conference on Computer Communication and Network Security (CCNS). IEEE, 2020: 191-195.

[12] Chen, X., Zhu, J., & Dai, Z. (2019). Simple and effective text matching with rich semantic features based on Synonym Replacement. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3962–3972).

[13] Wei J, Zou K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks[J]. arXiv preprint arXiv:1901.11196, 2019.

[14] Coulombe C. Text data augmentation made simple by leveraging nlp cloud apis[J]. arXiv preprint arXiv:1812.04718, 2018.

[15] Beddiar D R, Jahan M S, Oussalah M. Data expansion using back translation and paraphrasing for hate speech detection[J]. Online Social Networks and Media, 2021, 24: 100153.

[16] Sun C, Qiu X, Xu Y, et al. How to fine-tune bert for text classification?[C]//Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18. Springer International Publishing, 2019: 194-206.

[17] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[18] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv:1404.2188, 2014.

[19] Yin W, Kann K, Yu M, et al. Comparative study of CNN and RNN for natural language processing[J]. arXiv preprint arXiv:1702.01923, 2017.

[20] Chen Y. Convolutional neural network for sentence classification[D]. University of Waterloo, 2015.

[21] Xu N, Wang P, Chen L, et al. Distinguish confusing law articles for legal judgment prediction[J]. arXiv preprint arXiv:2004.02557, 2020.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, andKristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[23] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. Advances in neural information processing systems, 2019, 32.

[24] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.1

[25] Ramponi A, Plank B. Neural unsupervised domain adaptation in NLP—a survey[J]. arXiv preprint arXiv:2006.00632, 2020.

[26] Chen X, Cong P, Lv S. A long-text classification method of Chinese news based on BERT and CNN[J]. IEEE Access, 2022, 10: 34046-34057.

[27] Lin B Y, He C, Zeng Z, et al. Fednlp: Benchmarking federated learning methods for natural language processing tasks[J]. arXiv preprint arXiv:2104.08815, 2021.