Criminal Charge Prediction Based on Legal Text Classification Methods

QING-ZHE CUI[1], WEI XU[1,✉], KE XU[2,✉], QING-HAO CAI[3], XIN YUAN[4], YAN-RUI CHEN[5], KAI-FANG LI[6], LI-XIANG XU[7]

[1] Faculty of Mathematics and Computer Science, University of Münster, Münster 48149, Germany
[2] School of Law, Anhui Medical University, Anhui 231200, China
[3] School of Artificial Intelligence, Anhui University, Anhui 230601, China
[4] School of Electrical and Mechanical Engineering, The University of Adelaide, Adelaide, SA 5005, Australia
[5] School of Transportation, Southeast University, Nanjing 211189, China
[6] School of Advanced Manufacturing Engineering, Hefei University, Hefei 230601, China
[7] School of Artificial Intelligence, Hefei Institute of Technology, Hefei 230601, China
✉ Corresponding Authors' E-MAIL: wei.xu@uni-muenster.de, 351954525@qq.com

Abstract:

Criminal charge prediction is an important task in legal artificial intelligence, aiming to automatically identify the corresponding charge from the fact description of a criminal case. In this paper, we formulate charge prediction as a legal text classification problem and conduct comparative experiments on the CAIL2018 legal judgment dataset. Single-charge samples are selected from the original dataset, and the top 10 most frequent charge categories are used to construct a balanced dataset. We compare three traditional machine learning methods based on character-level TF-IDF features with two deep learning methods, TextCNN and BERT. The experimental results show that both traditional and deep learning methods can achieve strong performance on this task, while TF-IDF with Linear SVM obtains the best overall result. This study provides a lightweight and complete experimental reference for criminal charge prediction based on legal text classification.

Keywords:

Criminal Charge Prediction; Legal Text Classification; CAIL2018; TF-IDF; TextCNN; BERT.

1. Introduction

Legal judgment prediction [1] is an important task in the field of legal artificial intelligence [2], aiming to automatically analyze legal case documents and predict corresponding judicial outcomes. Among different legal prediction tasks, criminal charge prediction [3] is a fundamental problem, because the charge of a case directly reflects the legal nature of the defendant's behavior and provides important references for subsequent sentencing and judicial reasoning.

In criminal cases, the case fact description contains key information about the behavior, object, consequence, and subjective intention involved in the case. Therefore, predicting criminal charges from case fact descriptions can be regarded as a typical legal text classification task. With the development of natural language processing technology, both traditional machine learning methods and deep learning methods have been widely used for text classification. Traditional methods usually combine TF-IDF [4] features with classifiers such as Naive Bayes [5], Logistic Regression [6], and Support Vector Machine (SVM) [7]. Deep learning methods, such as TextCNN [8] and BERT [9], can further learn semantic representations from text.

Although deep learning models usually have stronger representation learning ability, traditional machine learning methods may still perform competitively in tasks where category boundaries are relatively clear. Therefore, it is necessary to compare these methods under the same dataset and evaluation setting, especially for criminal charge prediction based on legal fact descriptions.

To investigate this problem, this paper conducts a criminal charge prediction experiment based on the CAIL2018 [10] legal judgment dataset. We select samples with only one charge label from the original training set and choose the top 10 most frequent charge categories as the research objects. To reduce the influence of class imbalance, 5,000 samples are randomly selected for each charge category, forming a balanced dataset with 50,000 samples. The dataset is divided into training, validation, and test sets according to the ratio of 8:1:1.

In this study, five methods are designed and compared for

criminal charge prediction. The traditional machine learning methods include TF-IDF with Multinomial Naive Bayes, TF-IDF with Logistic Regression, and TF-IDF with Linear Support Vector Machine. The deep learning methods include TextCNN and BERT. Accuracy, Macro-F1, and Weighted-F1 are used as the main evaluation metrics.

The main objective of this paper is to compare the effectiveness of traditional machine learning and deep learning methods on the criminal charge prediction task. Specifically, this study aims to answer the following questions:

(I) How do traditional machine learning methods perform on charge prediction based on legal case fact descriptions?

(II) Can deep learning methods such as TextCNN and BERT achieve better performance than TF-IDF-based linear classifiers?

(III) Which method is more suitable for a balanced Top 10 criminal charge classification task under limited computational cost?

Through this experimental study, we aim to provide a lightweight but complete reference pipeline for legal text classification, including data preprocessing, sample balancing, model construction, performance evaluation, and result visualization.

2. Related Work

Legal judgment prediction has become an important research direction in legal artificial intelligence. Its goal is to predict judicial results, such as relevant law articles, criminal charges, and prison terms, according to the fact descriptions of legal cases. The CAIL2018 [10] dataset provides a large-scale benchmark for this task and has been widely used in Chinese legal judgment prediction research. Since charge prediction can be formulated as a text classification problem, many existing methods focus on learning effective representations from legal fact descriptions.

Traditional text classification methods usually represent documents with statistical features and then apply machine learning classifiers. TF-IDF [4] is commonly used to transform text into sparse feature vectors, which can be combined with classifiers such as Naive Bayes [5], Logistic Regression [6], and Support Vector Machine [7]. Among them, SVM has been shown to be effective for text categorization because of its ability to handle high-dimensional sparse features. These methods are simple and efficient, making them suitable baselines for legal text classification.

Deep learning methods can automatically learn semantic features from text. TextCNN [8] applies convolutional neural networks [11] to sentence classification and captures local n-gram features through convolutional filters. BERT [9] further improves text representation by using bidirectional Transformer-based pre-training, and it can be fine-tuned for downstream classification tasks with only a task-specific output layer. Compared with traditional methods, deep learning models usually have stronger representation ability, but they also require more computational resources and training time.

In this paper, we compare both traditional machine learning methods and deep learning methods for criminal charge prediction. Specifically, TF-IDF with Naive Bayes, Logistic Regression, and Linear SVM are selected as traditional baselines, while TextCNN and BERT are used as deep learning models. This comparison helps evaluate whether deep models can bring clear advantages over lightweight traditional classifiers on a balanced Top 10 charge classification task.

3. Methodology

In this study, criminal charge prediction is formulated as a multi-class legal text classification task. Given a case fact description x, the objective is to predict its corresponding charge label y from a predefined charge set $\mathcal{Y} = \{1, 2, \ldots, K\}$, where K denotes the number of charge categories. As shown in Fig. 1, we compare traditional machine learning methods and deep learning methods under the same experimental setting.

3.1. Text Representation

For traditional machine learning models, the input case fact text is represented by character-level n-gram TF-IDF features. Character-level representation is adopted because Chinese legal texts may suffer from word segmentation errors, while character n-grams can directly capture local expressions in case descriptions. The feature extraction process can be written as:

$$\mathbf{v} = \mathrm{TFIDF}(x), \tag{1}$$

where x denotes the case fact text and $\mathbf{v}$ is the corresponding sparse feature vector.

For deep learning models, the input text is transformed into dense representations. TextCNN uses character embeddings, while BERT uses token embeddings and contextual encoding. These representations are then used for charge classification.

3.2. Traditional Machine Learning Models

Three traditional machine learning classifiers are used in this paper, including Multinomial Naive Bayes, Logistic Regres-

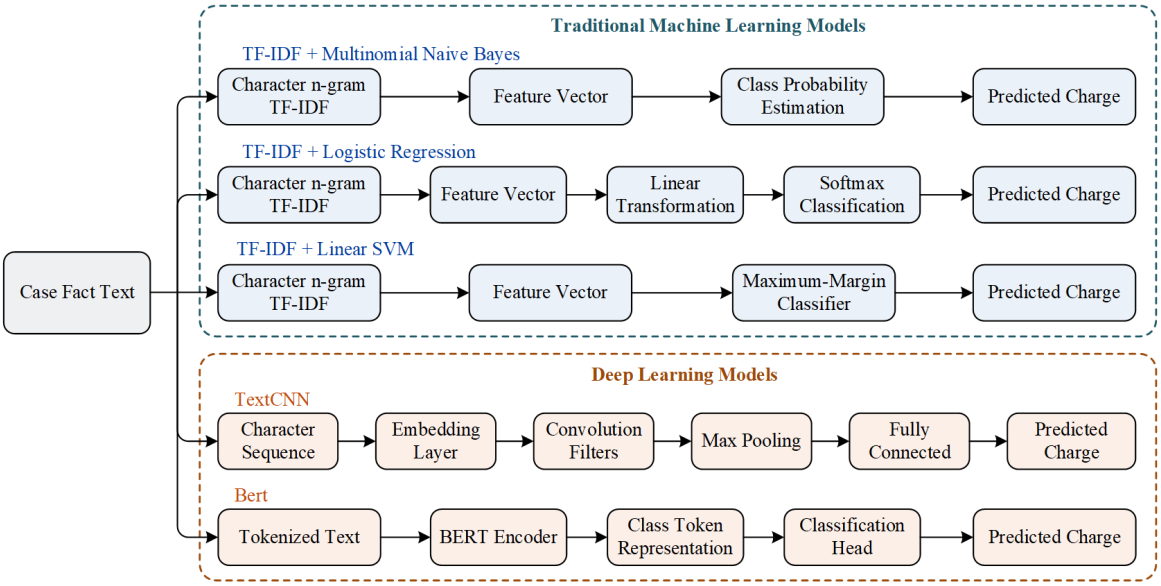

FIGURE 1. Overall Framework of Criminal Charge Prediction Based on Legal Text Classification.

sion, and Linear Support Vector Machine. All of them take the TF-IDF feature vector as input.

For TF-IDF with Multinomial Naive Bayes, the predicted charge is obtained by selecting the class with the maximum posterior probability:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(y|\mathbf{v}). \tag{2}$$

For TF-IDF with Logistic Regression, the TF-IDF vector is first mapped to class scores by a linear transformation, and the final prediction is obtained by the softmax function:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \text{softmax}(\mathbf{W}\mathbf{v} + \mathbf{b})_y, \tag{3}$$

where $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters.

For TF-IDF with Linear SVM, the classifier learns a maximum-margin decision boundary in the high-dimensional feature space. The prediction process is defined as:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} (\mathbf{w}_y^T \mathbf{v} + b_y), \tag{4}$$

where $\mathbf{w}_y$ and b_y denote the weight vector and bias term for class y.

3.3. Deep Learning Models

TextCNN is used to capture local semantic patterns from case fact descriptions. Given the input character sequence x, each character is first mapped into an embedding vector. Convolutional filters are then applied to extract local features, followed by max pooling and a fully connected layer for classification. The overall process can be expressed as:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \text{softmax}(\mathbf{W}_c \cdot \text{Pool}(\text{Conv}(\text{Embed}(x))) + \mathbf{b}_c)_y. \tag{5}$$

BERT is adopted to obtain contextual semantic representations of legal texts. The tokenized case fact text is fed into the BERT encoder, and the representation of the classification token is used for final prediction:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \text{softmax}(\mathbf{W}_b \cdot \mathbf{h}_{[\text{CLS}]} + \mathbf{b}_b)_y, \tag{6}$$

where $\mathbf{h}_{[\text{CLS}]}$ denotes the contextual representation of the classification token generated by BERT.

3.4 Classification Objective

For all trainable models, cross-entropy loss is used as the optimization objective:

$$\mathcal{L} = -\sum_{i=1}^{N} \sum_{y=1}^{K} \mathbb{I}(y_i = y) \log p(y|x_i), \tag{7}$$

where N is the number of training samples, K is the number of charge categories, y_i is the true label of the i-th sample, and $p(y|x_i)$ denotes the predicted probability of class y.

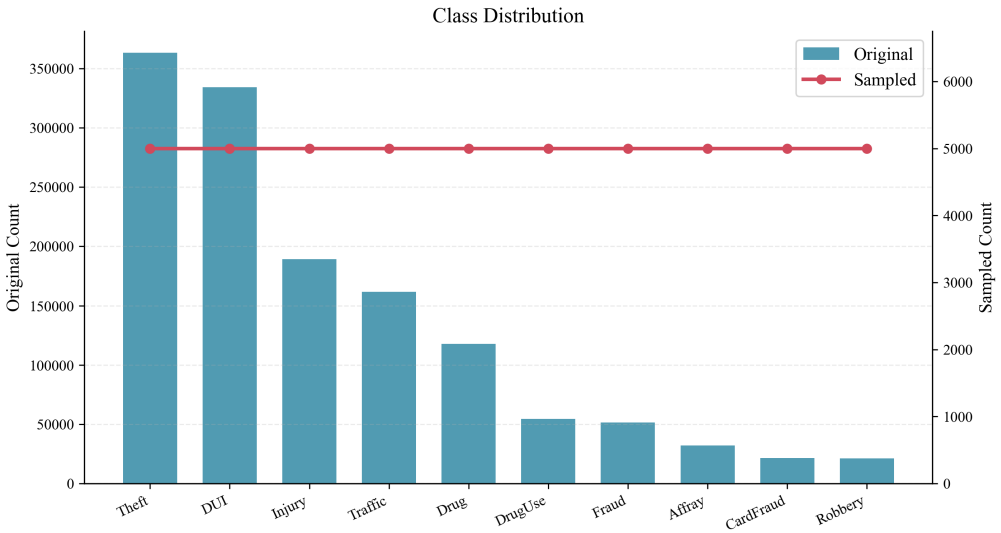

FIGURE 2. Class distribution before and after balanced sampling.

4 Experimental Results and Analysis

4.1 Dataset

The dataset used in this experiment is derived from the training set of the CAIL2018 [10] legal judgment dataset. This study focuses on criminal charge prediction based on case fact descriptions. Each sample contains a fact description and its corresponding charge label.

Since the original dataset contains cases with different numbers of charge labels, we first filter out samples that contain multiple charges and only retain samples with a single charge label. Then, the top 10 most frequent charge categories are selected as the research objects. To reduce the influence of class imbalance, 5,000 samples are randomly selected from each charge category. Finally, a balanced dataset with 50,000 samples is constructed.

The dataset is divided into training, validation, and test sets according to the ratio of 8:1:1. Specifically, 40,000 samples are used for training, 5,000 samples are used for validation, and 5,000 samples are used for testing. The training set is used to optimize model parameters, the validation set is used for model selection, and the test set is used for final evaluation.

As shown in Fig. 2, the original class distribution is highly imbalanced. Categories such as theft and dangerous driving have much larger sample sizes than other categories. After sampling, each selected charge category contains the same number of samples, which makes the comparison among different models more reliable.

4.2 Implementation Details

In this experiment, five charge prediction methods are compared, including three traditional machine learning models and two deep learning models. For traditional machine learning methods, character-level TF-IDF is used to represent case fact texts. The n-gram range is set from 2 to 4, and the maximum number of features is set to 200,000. Based on the TF-IDF features, Multinomial Naive Bayes, Logistic Regression, and Linear SVM are trained as classification models.

For TextCNN, the input length is set to 512 characters. The vocabulary size is about 8,000, and the embedding dimension is set to 128. Dropout with a rate of 0.5 is used to reduce overfitting. The model is trained with the Adam optimizer, with a learning rate of 1×10^{-3} and a batch size of 128.

For BERT, we use the `bert-base-chinese` pre-trained model. The maximum input length is set to 512 tokens. The batch size is set to 32, and the learning rate is set to 2×10^{-5}. The model is trained for 20 epochs. All deep learning experiments are conducted on an NVIDIA GeForce RTX 3090 GPU.

Accuracy, Macro-F1, and Weighted-F1 are adopted as the main evaluation metrics. Accuracy measures the overall classification correctness. Macro-F1 calculates the average F1-score of all categories equally, while Weighted-F1 considers the number of samples in each category.

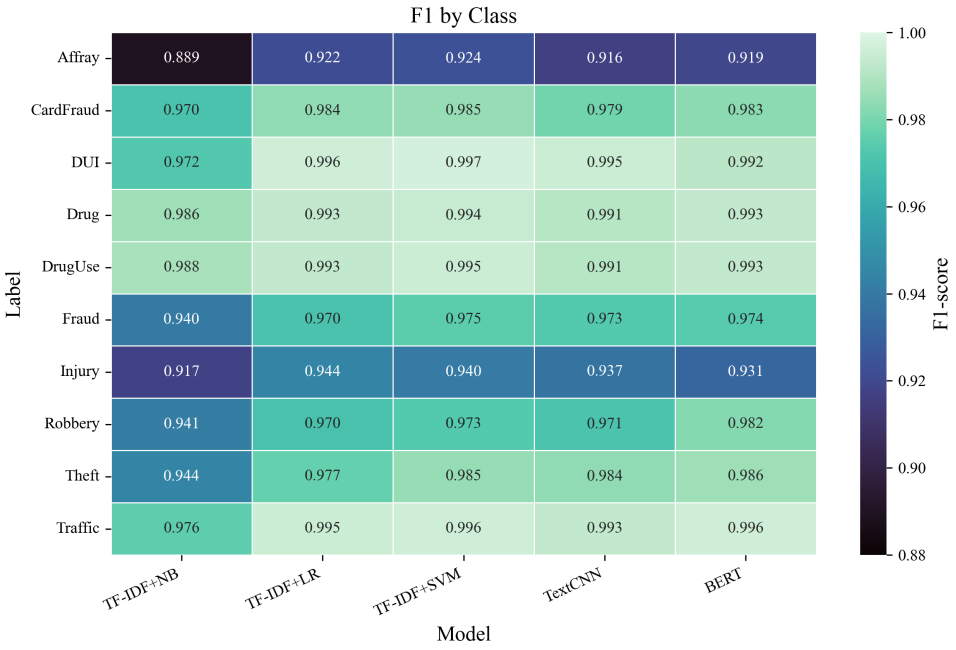

FIGURE 3. Class-level F1-score comparison of different models.

4.3 Results and Analysis

The overall experimental results are shown in Table 1. Among all compared methods, TF-IDF with Linear SVM achieves the best performance, with Accuracy, Macro-F1, and Weighted-F1 all reaching 0.9764. TF-IDF with Logistic Regression also performs well, achieving a Macro-F1 of 0.9744. These results show that character-level TF-IDF features combined with linear classifiers are highly effective for this charge prediction task.

TABLE 1. Experimental results of different models on the test set.

Model	Accuracy	Macro-F1	Weighted-F1
TF-IDF+NB	0.9524	0.9523	0.9523
TF-IDF+LR	0.9744	0.9744	0.9744
TF-IDF+SVM	0.9764	0.9764	0.9764
TextCNN	0.9730	0.9730	0.9730
BERT	0.9750	0.9749	0.9749

Compared with traditional machine learning methods, the deep learning models also achieve strong performance. TextCNN obtains a Macro-F1 of 0.9730, which is close to the results of Logistic Regression and Linear SVM. This indicates that convolutional neural networks can effectively capture local textual patterns in legal fact descriptions. BERT achieves a Macro-F1 of 0.9749, slightly higher than TextCNN but still lower than Linear SVM. Although BERT has stronger contextual representation ability, its advantage is not obvious in this balanced Top 10 charge classification setting.

The class-level F1-scores of different models are shown in Fig. 3. Most models perform well on categories such as dangerous driving, drug-related crimes, and traffic offenses. However, the category of affray obtains relatively lower F1-scores across all models, indicating that this category may have more ambiguous fact descriptions or stronger overlap with other charge categories. TF-IDF with Naive Bayes performs worse than the other models on several categories, especially affray and injury, which leads to its lower overall performance.

Overall, the experimental results show that traditional machine learning methods remain highly competitive in criminal charge prediction based on legal text classification. In particular, TF-IDF with Linear SVM achieves the best performance while maintaining relatively low computational cost. Deep learning models such as TextCNN and BERT also perform well, but they do not show a clear advantage over TF-IDF-based linear classifiers in this experimental setting.

5. Conclusions

In this paper, we conducted a comparative study on criminal charge prediction based on the CAIL2018 legal judgment dataset. The task was formulated as a legal text classification problem, and a balanced Top 10 charge dataset was constructed from single-charge case samples. Several traditional machine learning and deep learning methods were compared under the same experimental setting. The results show that traditional TF-IDF-based methods, especially the linear classifier, remain highly competitive for this task. Although TextCNN and BERT can also achieve strong performance, their advantages are not obvious in the current balanced classification scenario. Overall, this study demonstrates that simple and efficient text classification methods can still provide strong baselines for criminal charge prediction. In future work, we will further explore more complex settings, including imbalanced data, larger charge categories, multi-charge prediction, and domain-specific legal pretrained models.

References

[1] Chuyue Zhang and Yuchen Meng, "Bridging the divide: technical research and application on legal judgment prediction," *Artificial Intelligence and Law*, pp. 1–32, 2025.

[2] Al-Hareth Alhalalmeh and Alalddin Al-Tarawneh, "Artificial intelligence and the law: the complexities of technology and legalities," in *Intelligence-Driven Circular Economy: Regeneration Towards Sustainability and Social Responsibility—Volume 2*, pp. 641–649. Springer, 2025.

[3] Weikang Yuan, Kaisong Song, Zhuoren Jiang, Junjie Cao, Yujie Zhang, Chengyuan Liu, Jun Lin, Ji Zhang, Kun Kuang, and Xiaozhong Liu, "A multi-agent framework with legal event logic graph for multi-defendant legal judgment prediction," *Information Processing & Management*, vol. 63, no. 1, pp. 104319, 2026.

[4] Shahzad Qaiser and Ramsha Ali, "Text mining: use of tf-idf to examine the relevance of words to documents," *International journal of computer applications*, vol. 181, no. 1, pp. 25–29, 2018.

[5] Geoffrey I Webb, Eamonn Keogh, and Risto Miikkulainen, "Naïve bayes.," *Encyclopedia of machine learning*, vol. 15, no. 1, pp. 713–714, 2010.

[6] Michael P LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[7] Shan Suthaharan, "Support vector machine," in *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pp. 207–235. Springer, 2016.

[8] Xiaonan Xu, Zheng Xu, Zhipeng Ling, Zhengyu Jin, and Shuqian Du, "Comprehensive implementation of textcnn for enhanced collaboration between natural language processing and system recommendation," in *International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2024)*. SPIE, 2024, vol. 13180, pp. 1527–1532.

[9] Anna Rogers, Olga Kovaleva, and Anna Rumshisky, "A primer in bertology: What we know about how bert works," *Transactions of the association for computational linguistics*, vol. 8, pp. 842–866, 2020.

[10] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al., "Cail2018: A large-scale legal dataset for judgment prediction," *arXiv preprint arXiv:1807.02478*, 2018.

[11] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 6999–7019, 2021.