Extraction of Elements of Protest Based on BERT Model and TextTeaser improved algorithm

Yueguo Zhang¹, Lili Dong²* and Shenghong Li¹

¹ School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai, 20020, China
² Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 20020, China
*Corresponding author’s e-mail: zhyg@sjtu.edu.cn

Abstract. A protest is a document that the procuratorial organ exercises the power of legal supervision and submits it to the judicial organ. Element extraction is the basis for tasks such as case card filling, document correction, cause analysis, and case recommendation. At present, there are no effective tools and methods for extracting the elements of criminal protests. Most of them are manually marked, with low accuracy and low efficiency. In response to this problem, this paper is to propose a method to extract the elements of the criminal protest by sub-category, based on Bert model, using natural language processing technology. The experiments are demonstrated to illustrate the effectiveness of the proposed method.

1. Introduction
Judicial supervision is an important system commonly implemented in the world. The People's Procuratorate has the right to file a protest and form a protest if any legally effective judgment or ruling is found in China. A protest is a written document that the procuratorial organ exercises the power of legal supervision and submits it to the judicial organ. Its main contents include the organ that filed the protest, the reasons for protest, the basis and requirements, etc. At present, the procuratorial system is undergoing profound changes and facing major opportunities and challenges in China. Technologies such as artificial intelligence and natural language processing will greatly enhance the development of smart inspection services, replacing many inefficient manual processing methods. The technique of extracting the elements of protest is of great significance for intelligent prosecution [1-2].

This paper is to propose a method to extract the elements of the criminal protest, using three Natural language processing (NLP) tasks of label classification, named entity recognition, and text summarization. BERT, Bidirectional Encoder Representation from Transformers, was announced by Google. It has achieved advanced results in some natural language processing tasks. This paper is to fine-tune BERT to perform three types of NLP tasks including label classification, named entity recognition, and text summary to extract the elements in the criminal protest.

Tag classification is one of the tasks of natural language understanding. Since the era of deep learning, recurrent neural networks (RNN) and their variants have gradually become the main composition method of classification models. The deep LSTM network proposed by Shi et al. improves the sentence feature learning method and accuracy of classification [3]. Lin R. combines the RNN and CNN network to propose the RCNN network [4], which improves the model learning the problem of slow speed and non-convergence, then the attention mechanism was proposed. Zhou et al.
further combined the LSTM network with the attention mechanism \cite{5} to solve the problem of cross-language classification. After the emergence of BERT, the label classification task record was refreshed again. With the development of time, the research of named entity recognition has changed from the early labor-consuming methods based on rule and dictionary \cite{6} to traditional machine learning methods which support vector machine SVM, conditional random field CRF, and then deep learning methods. Through continuous practice, CRF-assisted neural network for named entity recognition has become the longest used method \cite{7-8}. Now the emergence of BERT has ushered in a new climax in the research of named entity recognition. At present, extractive summarization algorithms include the experience-based Lead3 algorithm, graph-based textrank algorithm \cite{9}, topic model-based LSA \cite{10}, LDA algorithm, and the feature-based TextTeaser algorithm \cite{11}. The extractive summarization has strong generalization power and high flexibility. As the research of natural language processing enters the era of deep learning, the seq2seq method shines in machine translation. Combined with the attention mechanism, Alexander M. et al. applied the seq2seq+attention structure to the generative summary \cite{12}, and achieved good results. After BERT was proposed, how to apply BERT’s ideas to generative abstracts became a research hotspot.

2. RELATED WORKS

2.1. Criminal protest

According to the type of litigation, the protest can be divided into criminal protest, civil protest and administrative protest. Among them, the criminal protest can be divided into the second instance procedure protest and the trial supervision procedure protest. These two types of criminal protests are similar in text organization, grammar, and logic and this paper mainly analyses the criminal protest. A criminal protest letter refers to a document prepared by the People’s Procuratorate when it protests against a criminal judgment or ruling made by the People’s Court in accordance with the law. A standard criminal complaint, the logic of the writing consists of several parts, such as the head of the document, original judgment, review opinion, analysis of the situation, and argumentation of reasons, the summary of resistance decision, tail and notes.

2.2. BERT model and its fine-tuning

BERT simplifies the work of natural language processing tasks by further completing the task of natural language processing in the pre-training word vector stage. The user can fine-tune the pre-trained word vector, and then use it to complete different tasks. The BERT model is also a word vector model, which is used to convert symbolic text into a computable numerical vector. Word2vec converts words into word vectors \cite{13-15}. Compared with other seq2seq models, transformer is completely implemented by the attention network in BERT model. The BERT bidirectional language model has two training tasks, Mask LM and Next Sentence Prediction (NSP). The performance of machine translation and other language understanding tasks is also better than other models. For different NLP tasks, BERT gives corresponding calling methods. This paper uses BERT fine-tuning for multi-label classification to generate the counter-appeal tag elements in the protest, and for named entity recognition to extract the subject elements of the criminal protest.

2.3 Text summarization algorithm

Text summaries refer to extracting from texts and text collections, refining the key content and reorganizing and displaying them in a fluent language. Representative extractive summaries include experience-based Lead3 algorithm, graph-based textrank algorithm, topic model-based LSA and LDA algorithm, and feature-based TextTeaser algorithm. TextTeaser is a feature-based extractive summarization algorithm. When extracting summaries, TextTeaser considers the position feature score of the sentence in the text, the length feature score of the sentence, Density-Based Selection score of the sentence, and Summation-Based score of the sentence Selection, title score of the sentence. This
paper conducts an experiment of extractive summaries and improves the TextTeaser extractive summaries algorithm, which is used to extract case description elements of criminal protests.

3. Proposed Method
This paper proposes an element extraction method based on BERT fine-tuning. In the process of extracting elements, this paper mainly conducts the research of the following technical methods.

- Fine-tune BERT to generate protest label and train the model to infer the ability, for example, it is not only when the four words "sentenced are abnormally light" appear directly in the text that the label is generated, for "three years sentenced, too light", "the sentencing is not strong enough", the label "sentenced are abnormally light" can also be output through semantic understanding. The data set is labelled based on the original text of the criminal protest, and the training data is manually labelled. The network structure of the model is optimized to output the probability value convert to Chinese label.
- Fine-tuning BERT is used to extract case subjects, implement named entity recognition, select network structure, select data sets, and process output results.
- Modify the TextTeaser algorithm to generate case descriptions. Design a calculation method for the feature value of the criminal protest summary. Select the length of the summary and the number of sentences.
- Obtaining and establishing a case dictionary, combined with the selection of other auxiliary extraction methods except dictionary matching and realizing case extraction.
- Analyse the structure of the article, and use different regular expressions to extract different elements at different positions in the text.

This paper is to extract the following 12 elements from the criminal protest: protest number, judgment number, procuratorate, court, subject of the case, original judgment, cause of the case, review opinion, case description, label of the protest, legal basis, time of the protest, and time of elements extraction. According to the standardization of the criminal prosecution, the location of different elements is determined and the elements of the criminal prosecution are extracted. The subject of the case refers to the subject of the offender, the review opinion refers to the review opinion given by the procuratorial organ, such as the fact that the judgment is unclear, the application of the law is wrong, and the sentence is improper. The description refers to the prosecution’s the interpretation of the review opinion and the analysis of the facts are the summary of the main part of the criminal protest. There are nine types of protest labels, namely, application of law error, improper sentencing, procedural violation, error in determining facts, improper application of probation, improper recovery, and refund Improper, unusually light sentencing, and unusually severe sentencing.
3.1. Feature extraction based on BERT fine-tuning
After obtaining the text of the criminal protest, the paragraph features of the protest are used to divide the paragraphs. Then different methods are used to extract the elements of the corresponding paragraphs, and finally the extracted elements are merged and output as the element extraction of the criminal protest. In addition to using the BERT model, the TextTeaser algorithm is also improved to extract the elements of the protest.

From the head paragraph of the text, this method extracts elements of counter appeal number, element of judgment number, element of court name, element of procuratorate name, element of case subject, element of original judgment, element of cause of action. The method extracts the elements of review opinion from the paragraph of review opinion, obtain the case description elements, protest label elements from the paragraph on the reasons for the protest, and extracts the elements of the time of protest from the closing paragraph. The elements of legal basis appears in both the review opinion paragraph and the paragraph of the reasons for the protest. Therefore, it can be extracted and merged from the two paragraphs.

3.2. BERT Finetune multi-label classification generates counter-claim labels
The input is the paragraph of the reason for protest after the sentence and the tags are read and predicted sentence by sentence. After all sentences in the paragraph are processed, the merged result is output as the classification label. There are nine labels in total, namely: wrong application of law, improper sentencing, procedural violation, wrong identification of facts, improper application of probation, improper recovery, improper return of compensation, extremely light sentencing, extremely heavy sentencing. The output can be one label or multiple labels, or empty. When dividing the text to obtain the paragraph of the reasons for the protest, the beginning feature of the paragraph is like "the reasons are as follows" and "the details of the review are as follows" and the ending features of the paragraph are like "to sum up", "to maintain judicial justice, a protest is specifically raised", etc. The ending feature of a paragraph may be the starting feature extracted from another paragraph.

3.2.1. Fine-tuning multi-label classification
When using BERT fine-tune for single sentence classification tasks, we use BERT to get the output of the last layer of [CLS], and then calculate the loss by fully connecting sigmoid and one-hot code of the label. Because it is a multi-label classification, sigmoid is used as the activation function instead of the softmax function. The sigmoid function is defined by the following formula:

\[ S(x) = \frac{1}{1 + e^{-x}} \] (1)

\[ S'(x) = \frac{e^{-x}}{(1 + e^{-x})^2} = S(x) \cdot (1 - S(x)) \] (2)

3.2.2. Model optimization
This paper optimizes the model through three ways:

- Additional pre-training. BERT uses general text for pre-training, which is less involved in the judicial field of this research. If the Mask LM and Next Sentence Prediction task is to use the model to further understand the text in the judicial field, it will be used in the label generation of the criminal protest to get better performance [16]. This paper uses 1640 criminal protests as texts for additional pre-training. The learning rate is set to "3e-5", batch_size is 32, and train_steps is 50000.

- Loss function modification. In order to solve the problem that the imbalance of positive and negative samples in the data set leads to poor prediction of some labels, the original binary classification CE cross entropy is changed to Focal loss [17].

- Focal loss was originally proposed to improve the effect of image object detection. By reducing the contribution of easily classified samples to loss, it effectively solved the problem of imbalance between positive and negative samples. In this paper, the two label samples such as procedural
violation and excessive sentencing are extremely unbalanced, with a positive-negative ratio of 1:1000, which is an opportunity for Focal loss to show its talents. The CE losses are as follows:

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{otherwise} \end{cases}$$ (3)

Let

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases}$$ (4)

Equation 4 is equal to:

$$CE(p, y) = CE(p_t) = -\log(p_t)$$ (5)

Focal loss is calculated as follows:

$$FL(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t)$$ (6)

$$\alpha_t = \begin{cases} \alpha & \text{if } y = 1 \\ 1 - \alpha & \text{otherwise} \end{cases}$$ (7)

In the formula, the definition of $p_t$ is the same as before, $\gamma$ is the number in the interval [0,5], $\alpha$ is the number in the interval [0,1]. In this paper, $\gamma$ is 2, $\alpha$ is 0.25.

The final result of the model output is the probability value of the interval [0,1]. The probability value indicates the probability that the corresponding label is 1. In general, when the probability value is greater than 0.5, the corresponding label is output. In order to improve the accuracy of the model, the method of threshold shift is adopted. We use the training set to train the model, then use the test set to test the model, and filter the threshold to make the label conversion result the best. The thresholds for each label are different.

In this paper, the initial threshold of each label is 0.5. First search for the best threshold for the first label, starting from 0.1, then the threshold list becomes [0.1,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5], test the set, get the test result and calculate the score. After getting the score, increase 0.1 by 0.1 and test again. By analogy, until all the thresholds between 0.1 and 0.9 are tested, the threshold corresponding to the highest score is obtained and the threshold list is modified to the threshold of the first label. Then the second tag threshold test is performed until all tags have the best threshold. When actually generating the label of the criminal protest letter, the optimal threshold is used for output.

3.3. BERT Finetune named entity recognition and extraction of case subjects

The input is the head paragraph of the text after the sentence, read sentence by sentence and extract the entities. After processing all sentences in the paragraph, we merge the screening results to obtain the main elements of the case. There are mainly two types of entities which are PER and ORG.

- Fine-tuning named entity recognition. This paper uses BERT fine-tune to do named entity recognition tasks and process the data into the required format. When obtaining the output, use the mode method in BERT to obtain the word vector of each word, and then input the word vector into the constructed network for training.

- Model optimization. CRF enables the model to supervise output features, and CRF can also learn many hidden features. At each time step, the hidden state tensor of BiLSTM is passed to CRF, so that BiLSTM can learn new content under the limited characteristics of CRF. This paper uses the BERT-BiLSTM-CRF structure and performs additional pre-training on the BERT model. The model structure is shown in Figure 2.
3.4. **TextTeaser algorithm improvement to extract case description elements**

When applying the TextTeaser algorithm to extract the elements of criminal protest, there are two main problems. One is improper calculation of the feature score of the sentence's position in the text. The sentence of the reason for protest will get a lower score if TextTeaser algorithm is be used. The other is lack of weight calculation for the summary of the criminal protest. Some sentences should not appear in Abstract. For words such as "constitute", "violation", and "not adopted", special weights should be calculated. This paper has made the following improvements to the TextTeaser algorithm:

- Delete the weight calculation of the sentence position and add the judgment of the listed sentence. Specifically, when the sentence appears in the form of "one is", "second" can be judged as listed sentences. These sentences themselves have strong generalization power and can be directly used as the content of the case description.

- Increase the calculation of the characteristic score (SK) of the criminal protest. Specifically, when "should", "verify", and "constitute" etc. appear in the sentence, we increase the weight of the sentence. When "the following circumstances" etc. appear in the sentence, we reduce the weight of the sentence. When increasing and decreasing the weight, we increase and decrease the weight value of 1.

\[
\text{TotalScore}(S) = TS \cdot 0.5 + SL \cdot 0.5 + KF \cdot 2.0 \quad + \quad KS \qquad (8)
\]

In the formula, all parameter definitions are the same as the previous definitions.

- The input of the paragraph title is the elements of the extracted review opinions. If the elements of the review opinion are extracted, "the judgment is indeed wrong", "the facts are wrong", then this sentence is used as the heading of the entire paragraph and to calculate the sentence heading scores of the remaining sentences.

To extract case description, we calculate the weight of all sentences, take out the sentence in the paragraph, then sort the other sentences in descending order of weight and output them as the case description elements of the criminal protest.

3.5. **Case dictionary matching**

The input is the head paragraph of the text and the output is the cause element. The grammatical structure analysis and the dictionary matching of the case are combined to extract the cause of the case. The cause of the case is obtained by combining the two methods and the cause of the case is output as the element of the criminal protest. The head paragraph of the text has two ways to express the cause of the case. Take the crime of dangerous driving as an example, "something is suspected of dangerous driving" and "something is guilty of/constitutes a crime of dangerous driving". A total of 842 cases
were crawled on the Chinese Judgment Documents website. These cases are made into a dictionary to judge whether each cause in the dictionary appears in a paragraph one by one, and extract it as the cause if it appears, for example, "crime of arson", "crime of poisoning". There may be both "crime of illegally renting and lending guns" and "crime of lending guns" at the same time. In this case, the longer case is reserved for "crime of illegally renting or lending guns".

Other elements are mainly used to extract corresponding elements in specific paragraphs through regular expressions, and the elements of the protest letter number are extracted. In addition, the entire review opinion paragraph is the review opinion element, which can be obtained directly after the paragraph is divided. The extracted review opinion elements are also used as the title in the TextTeaser improved algorithm to calculate the sentence title feature score and also used to calculate the sentence title feature score.

4. Experiments

4.1. Data set and experimental design

In the experiment, 1640 public criminal protests have been collected as experimental data. When the multi-label classification experiment is conducted, 1476 criminal protests are used as the training set and 164 are used as the test set. When conducting named entity recognition experiments, we use public data sets. In the BERT Fine-tune multi-label classification experiment, the basic BERT Fine-tune multi-label classification is compared with the algorithm in this paper, and the BERT+RCNN structure is tested. The BERT+RCNN structure is based on the algorithm in this paper, and the fully connected layer in this paper is replaced by a layer of RCNN network [18].

At present, the commonly used evaluation standard for multi-label classification and named entity recognition is F-Measure. The F value is a comprehensive calculation of the precision rate and the recall rate. The calculation formula is as follows:

\[ F_\beta = \frac{(1 + \beta^2) \times P \times R}{\beta^2 \times P + R} \]  

In the formula: \( \beta \) is a constant, usually taken as 1, and the F value at this time is also called the F1 value. \( P \) is the accuracy rate, which is the ratio of the number of correct positive examples in the test results to the number of all classified positive examples. That is, \( P = \text{correct number/(correct number+wrong number)} \). \( R \) is the recall rate, which is the ratio of the number of correct positive cases in the test results to the number of actual positive cases, which is \( R = \text{correct number/number of positive examples in the data set} \). For the extraction of other elements, the extraction rate is calculated:

\[ F = \frac{C}{A} \]  

\( C \) is the amount of content extracted. \( F \) is the amount of content contained in the text.

4.2. Experimental results and analysis

4.2.1. BERT Finetune multi-label classification

The basic BERT Finetune multi-label classification, BERT+RCNN structure, the results of this classification method on the test set are as follows. Because BERT's fitting ability is already very strong, connecting other classification networks (RCNN, etc.) will influence the existing fitting effect. In this task, the classification features are obvious and the difficulty is low. The best results can be achieved by directly using the fully connected layer.

| Label | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | Result |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|--------|
| 0.9984 | 0.9786 | 0.9863 | 0.9932 | 0.9512 | 0.7  | 0.9  | 1.0  | 0.0  | 0.9096 |
4.2.2. BERT Finetune named entity recognition

The recognition results of the model on the ORG (organization) and PER (person) entities are as follows:

|                | ORG | PER |
|----------------|-----|-----|
| Named entity recognition based on BERT | 0.8968 | 0.9758 |

The PER effect is obviously better. Compared with PER, ORG has a poorer recognition effect. The reason is the lack of a named entity recognition task data set in the legal field. This paper uses the public data set for training. Although the judicial field text is added to the BERT pre-training, there is no judicial field related corpus for the subsequent named entity recognition training task. Regarding the name of the court, there is still room for improvement in the recognition of the name of the procuratorate.

4.2.3. Other factors extraction results

Among the 1,640 criminal protests, the selection rate of the protest number, judgment number, procuratorate, court, subject of the case, original judgment, cause, review opinion, legal basis, and time of the protest is as follows:

|                | Protest number | Judgement number | Procuratorate | Court | Subject of the case | Original judgment | Cause of case | Review opinion | Legal basis | Time of protest |
|----------------|----------------|------------------|--------------|-------|---------------------|------------------|---------------|----------------|-------------|----------------|
| Decimation rate | 0.9987         | 0.9957           | 1.0          | 0.9963| 0.9859              | 0.9963           | 0.9914        | 1.0            | 1.0         | 0.9975         |

Among them, the extraction rate of the protest number, judgment number, and time of the protest did not reach 1.0. The reason is that there is an irregular criminal protest. When writing the number “0”, it will use the Russian character “О” and the English letter “O”. The normal extraction rate of criminal protests should reach 1.0.

This paper details the algorithm experiment of extracting elements of criminal protest. From the experimental results, the method proposed in this paper has achieved good results and can effectively assist the case-handling personnel in scoring, analysis, and information collection.

5. Conclusion and further work

This paper analyses the trial supervision process, conducts an in-depth analysis of a large number of typical cases and data, and clarifies the elements in the criminal protest. Based on the BERT model and Text Teaser algorithm, an algorithm for extracting elements of criminal protests is proposed. After obtaining the text of the criminal protest, we use different methods to extract corresponding elements in different paragraphs. This paper uses Fine-tune BERT to complete multi-label classification. This paper proposes some optimization methods such as network structure increase, loss function modification, additional pre-training, etc. Additional pre-training methods are used and the network structure is modified to BERT-BiLSTM-CRF. This paper improves the TextTeaser algorithm for extractive summaries, increase the sentence feature scores for criminal protests, and reduce the sentence location feature scores in the algorithm.

In total, this paper proposes and implements a method for extracting elements of criminal protests, and conducts a large number of experiments. The experiments show that the proposed method is
reasonable and efficient. In the future, researchers can further analyse and dig deep-level information in criminal protests based on the extraction results of the elements, and promote the construction of smart justice. In addition, researchers can study the application of the method proposed in this paper to the extraction of elements of other types of protest letters.

Acknowledgments
This work was supported by the National Key Research and Development Program of China (2018YFC0832000, 2018YFC0830800).

References
[1] Zhao Z.G., Jin H.H.(2019) Enlightenment from the practice of "smart prosecution" under the strategy of intelligent politics and law[J]. People's Procuratorate, 8: P29-32.
[2] Tao J.P.(2020) Analysis and prospects of the construction of "smart prosecution"[J]. Chinese prosecutors, 1:69-72.
[3] Shi Y.Y., Yao K.S, et al.(2016) Deep LSTM based Feature Mapping for Query Classification[C]. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.1501-1511.
[4] Lin R.H., Fu C.Z., et al.(2018) Academic News Text Classification Model Based on Attention Mechanism and RCNN. CCF Conference on Computer Supported Cooperative Work, and Social Computing. Singapore, pp.507-516.
[5] Zhou X.J., Wan X.J., Xiao J.G.(2016) Attention based LSTM Network for Cross Lingual Sentiment Classification[C]. empirical methods in natural language processing, pp.47-256.
[6] Xie R.B., Liu Z.Y., Jia J., et al. (2016) Representation Learning of Knowledge Graphs with Entity Descriptions[C]. In AAAI, pp.2659-2665.
[7] Liu L.Y., Shang J.B., Xu Frank F.(2018) Empower Sequence Labeling with Task-aware Neural Language Model[C]. In AAAI, pp.5253-5260.
[8] Greenberg N., Bansal T., et al. (2018) Marginal Likelihood Training of Bilstm-crf for Biomedical Named Entity Recognition from Disjoint Label Sets[C]. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp.2824-2829.
[9] Mihalcea R., Paul T. (2004) TextRank: Bringing Order into Texts.". empirical methods in natural language processing, pp.404-411.
[10] Makbule G. O., Ferda N. A., Ilyas C. (2011) Text summarization using Latent Semantic Analysis, In Journal of Information Science, 37(4):405-417.
[11] Dadang G.W., Ayodhia P.P. et al. (2017) Automatic Text Summarization for Indonesian Language Using TextTeaser[J]. IOP Conference Series: Materials Science and Engineering, pp.1-7
[12] Alexander M. R, Rush S.C., Jason W.(2015) A neural attention model for abstractive sentence summarization. In EMNLP, pp.379-389.
[13] Mikolov T., Chen K., Corrado G.,Dean J.(2013) Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, pp.1-12.
[14] Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L.(2018) Deep contextualized word representations, In NAACL, pp.2227-2237.
[15] Li Z. J.,Fan Y.,Wu X. J.,Summary of research on pre-training technology for natural language processing[J].Computer Science,2020,47(03):162-173.
[16] Hochreiter S., Schmidhuber J.(1997) Long short-term memory[J]. Neural Computation, 9(8):1735-1780.
[17] Chen S.D.,Ouyang X.Y.(2020)Summary of Named Entity Recognition Technology[J].Radio Communication Technology, 46(03):251-260.
[18] Girshick R., Donahue J., Darrell T., Malik J.(2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.580-587.