Research on the Text Classification of Legal Consultation Based on Deep Learning

ZuoQiang Du
School of Computer and Information Engineering,
Harbin University of Commerce,
Harbin, China

***Abstract*—In view of the existing traditional legal service, practitioners are unable to meet the huge demand; a large number of citizens are unable to determine the scope of the problems when they encounter infringement or require various legal assistance. Based on this, an automatic classification model of legal consultation based on Deep Learning is proposed in this paper. A KP+BiLSTM+Attention model is proposed. The Keyword Parser is introduced to extract key information. TF-IDF and part of speech tagging are used to filter out the important information in the user's legal problem description. The extracted keywords are given a weight value, and the other information weights are set to zero. The text information is transferred into two parallel word vector embedding layers. One of the word vector embedding layers transfers the results to the fusion layer for splicing, difference and point multiplication after the key information is converted into vector form. The output results are respectively connected with the results obtained from the other embedding layer as residuals. The final results are transferred to the BiLSTM+Attention model for training. The test results show that KP+BiLSTM+Attention model has significantly improved the accuracy and F1 value of the best benchmark method for text classification tasks of legal consulting. Therefore, KP+BiLSTM+Attention method has better performance in dealing with the classification of legal consulting issues.**

***Keywords*—Text classification; legal consultation; deep learning; KP+BILSTM+ATT model; word embedding layer**

I. Introduction

Legal consultation is to determine the scope of the problem encountered by the public and to provide legal solutions applicable to the relevant field. How to put forward effective opinions and suggestions quickly and effectively for their own problems is the focus of people's attention, and it is also an important issue facing the popularization of legal awareness [1][2][3]. In view of this problem, the demand of users is classified correctly for lawyer advice is the first task. Because it can complete automatically the classification and recognition of text information [4][5][6], Deep Learning (DL) is widely used in the field of Natural Language Processing (NLP) [7][8][9][10]. The problem category is determined accurately after adaptive training according to relevant information of legal provisions and customer problem descriptions [11][12][13].

Text classification is a basic work in the field of NLP, which aims at classifying text information [14][15][16][17]. Salton [18] proposed the word vector space model. However, this model needed to define a large number of rules for each category which were highly dependent on the professionals. The Q-AND-A method based on knowledge graph had achieved good results in the fields of public security information analysis, medical consultation and drug prescribing [19]. Zhang [20] used the fine-granularity question and answer model based on Bi-LSTM+CRF to select entities. Li [21] used the entity recognition model based on Bi-LSTM+CNN+CRF to establish a human-computer interactive question and answer system to improve the employment rate. In the legal field, a question and answer system was proposed based on the legal field by Huang [22], which introduced a small number of samples and a transfer learning model. Zhang et al. [23] introduced DL into the legal judgment task, and a better effect was achieved in the open data set of the real legal judgment prediction task.

In this paper, the original word segmentation algorithm TF-IDF is improved, and combined with the DL model. The brand new Keyword Parser (KP) is proposed based on common keyword extraction approach. Firstly the Chinese word segmentations are calculated based on TF-IDF, and the weights assigned according to the degree of importance. Then, it enhances the extracted key information according to part-of-speech screening, and the important words in each piece of data can be further screened and retained. The retained information is passed to two embedding layers at the same time. The fusion layer embedded in one embedding layer will process the incoming data by splicing and dot multiplication, and the results will be passed to the other Embedding one as the residual connection to improve the credibility of the information. Finally, the classification results were obtained by training Bi-LSTM+Attention model. Based on the legal consultation database, comparing common Machine Learning (ML) text classification models and DL text classification models, the experimental results show that KP-BiLSTM+Att model performs better on the legal consulting data set.

II. Structure of KP-BiLSTM-Att Model

A. Chinese Segmentation

Unlike English words, which are naturally separated by spaces, Chinese takes words as the basic unit and there is no obvious separation mark between words [24]. The main task of Chinese word segmentation [25] is to divide a complete sentence into individual words. There are three main types of mainstream Chinese word segmentation methods: word segmentation based on string matching, word segmentation based on understanding and word segmentation based on statistics [26]. In this paper, jieba segmentation is used as a word segmentation tool, and the sentence is cut into data information X using the built-in precise mode.

B. *Structure of Legal Consultation Text Classification Approach Based on KP-BiLSTM-Att*

The overall structure of the legal consultation text classification approach based on KP-BiLSTM-Att proposed in Fig. 1.

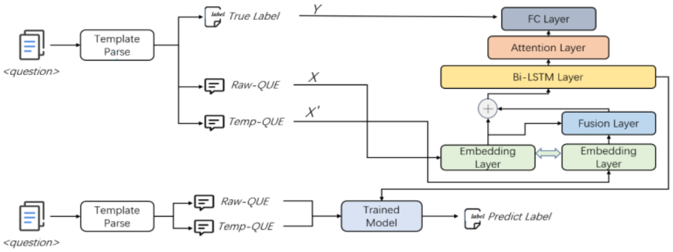

Fig. 1. Structure of the legal consultation text classification based on KP-BiLSTM-Att.

The text information X which has not been processed and the information X` which has been processed by keyword processors are passed into the two embedding layers for vectorization operations. A fusion layer is added to the embedding layer that is responsible for dealing with X' to carry out vector operations on the vectorization results of X and X', and then the results are connected with X as residuals. The vector results are passed into the BiLSTM+Attention model for training. The training results were used by linear layer and softmax function to predict the scope of legal consulting problems. Finally, Focal Loss is introduced to solve the unbalance problem of all kinds of samples in the dataset.

III. Model Training Stage

There are four modules during model training stage: Keyword Parse, two parallel embedding layers, Bi-LSTM layer and attention layer.

The main task of the KP is to receive the pre-processing results, and obtain the keyword information of the problem description by TF-IDF algorithm and [mask] label. Two parallel word embedding layers receive the original word segmentation result X and the keyword information X' after processing by the KP respectively. Meanwhile, a fusion layer is added the word embedding layer which is receiving X' to calculate the output vector information of two embedding layers. The final calculation results are passed into Bi-LSTM layer and attention layer for training.

A. *Keyword Parser*

In this module, the text information obtained after data preprocessing is received by the parser, and the weight value of each word in the word segmentation result about the category is calculated based on TF-IDF. The importance of each word for the category is obtained. Then, with the idea of part-of-speech tagging [27], all words are labeled with a correct part of speech, that is, each word is a noun, verb, adjective or other part. The part-of-speech tagging method based on statistics is adopted in the process of part-of-speech tagging [28].

The parser retains nouns, verbs, and adjectives, and it masks all the results of word segmentation other than the above parts of speech with the [mask] tag to get X'. The purpose of masking other parts of speech words is to keep the words that are helpful to the classification accuracy (that is, the words with the high TF-IDF values) and keep their original positions unchanged, so as to avoid the dislocation phenomenon in the subsequent weight addition, and improve the efficiency and accuracy of model training. The operations performed by the KP are shown in Fig. 2.

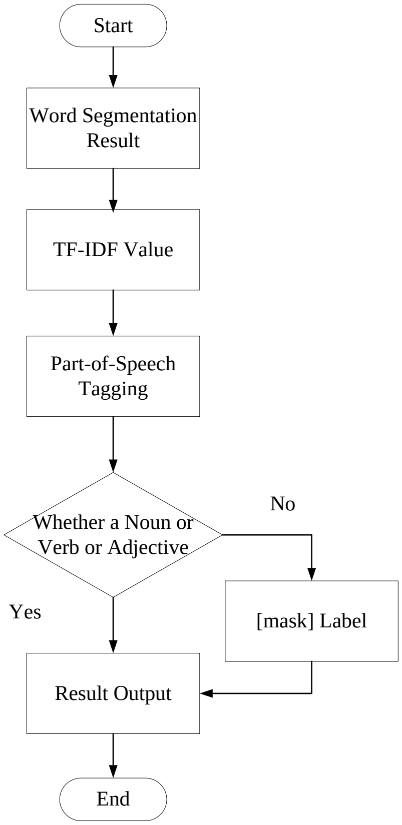

Fig. 2. Operations performed by KP.

B. *Word Embedding Layer*

The main task of embedding layer is to receive the text information X' which is the output by the KP and the unprocessed word segmentation result X. The text information of two incoming models make vectorial operation through two embedding layers respectively, and the data from the text information would be transformed to the computer recognizable vector information, meanwhile the relationship vector between different words obtained.

Compared with the traditional high-dimensional sparse feature matrix, the embedding can represent a word with low-dimensional vector and calculate the distance between this one and other words, so as to determine whether the two words are semantically similar.

In this paper, the two embedding layers are used to receive the text information X without the KP information and the text information X' with KP information respectively. The word embedding layer received X information will directly reduce the dimension of the information through the transformation of high-dimensional image and low-dimensional. The other word embedding layer converts *X*' into vector form after random initialization, and passes the results into the fusion layer [29] for concatenation, difference and dot multiplication operations with the vectorization result of *X*. The obtained result and the X result are respectively connected by residual to improve the information weight value.

There are a large number of [mask] labels in *X*', which will be directly set to 0 in the vector during the operation. The difference between keywords and non-keywords has been increased by the operations above, which is conducive to the subsequent model training. The structure of the word embedding layer is shown in Fig. 3.

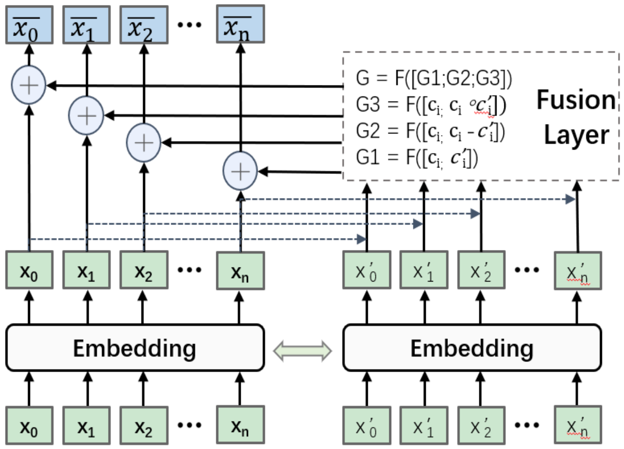

Fig. 3. Structure of the embedding layer.

C. Bi-LSTM+Attention Layer

Bi-LSTM is an improved model based on LSTM. In fact, two LSTM process the sequence from the forward direction and the reverse direction respectively, and the results of the two directions are combined to obtain a new vectorized result which will pass to the attention layer to assign different attention values [34]. Finally, the loss function in the fully connected layer is used to calculate the category. The structure of BiLSTM+Attention layer is shown in Fig. 4.

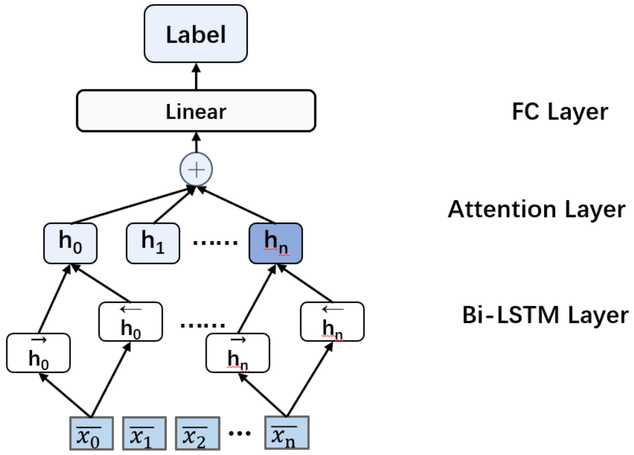

Fig. 4. Structure of BiLSTM+Attention layer.

Bidirectional LSTM is the LSTM Neural Network structure that combines the vector information obtained in the forward direction and the reverse direction on the basis of the LSTM processing sequence information in the reverse direction, and computes the new vector results. Fig. 5 shows the structure of bidirectional LSTM [35].

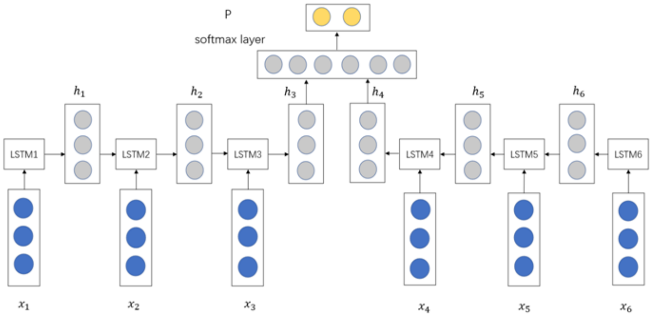

Fig. 5. Structure of bidirectional LSTM.

The new vector obtained by bidirectional LSTM is passed into the attention layer to calculate the attention weight of each word vector through the self-attention mechanism. The self-attention mechanism not only helps the current node focus on the current word, but also obtains contextual semantic information.

Self-attention will calculate three new vectors: Query, Key and Value, which are obtained by multiplying the word embedding vector and the random initialization matrix. Then, Query and Key will be dot-multiplied to get the weight, which is the score of self-attention. When we encode a word, the score of the self-attention determines how much attention is paid to the input sequence:

$$f(Q,K_i)=Q^{\mathrm{T}}K_i \quad (1)$$

The correlation size of each word for the current position will get by a softmax calculation:

$$\alpha_i = soft\max(f(Q,K_i)) = \frac{\exp(f(Q,K_i))}{\sum_j f(Q,K_j)} \quad 2)$$

Value and softmax are multiplied and added together to obtain the attention value of the current node. The units in each sequence are made attention calculating with all the units in the sequence to obtain the self-attention value.

$$Attention(Q,K,V)=softmax(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V \quad (3)$$

where QK^{T} is the attention matrix, and $\sqrt{d_k}$ is the conversion from the attention matrix to a standard normal distribution. Multi-Head Attention mechanism is to carry out multiple attention operations, so that the model has multiple attention values with the same structure but different weights.

$$head_i = A(Q,K,V) \quad (4)$$

$$MultiHead(Q,K,V)=Concat(head_1,head_2,\ldots\ldots,head_n)W^{Q} \quad (5)$$

The vector obtained after vector operation by bidirectional LSTM Neural Network and self-attention mechanism can remember the semantic information of context well, and the accuracy of classification results is higher.

D. Focal Loss Based on Class Balance

Due to the unbalanced distribution of legal problems, Focal Loss [30] is introduced as an effective method to solve the problem of unbalanced quantity of different samples in data sets. Focal Loss can reduce the weight of easily classified samples to make the model pay more attention to difficult classified samples in training. The traditional cross entropy loss should be improved by this approach.

As an example of taking binary classification, the loss function of cross entropy is shown as (6).

$$CE(p,y)=\begin{cases}-\log(p) & \text{if y=1}\\ -\log(1-p) & \text{otherwise}\end{cases} \tag{6}$$

where $p\in[0,1]$ is the classification probability. p_t is the probability that the sample belongs to the true class.

$$p_t=\begin{cases}p & \text{if y=1}\\ 1-p & \text{otherwise}\end{cases} \tag{7}$$

The size of p_t can reflect the degree of difficulty of sample classification. In the training process, the model should pay more attention to hard-to-classify samples, so the proportion of these kinds of samples should be increased.

Focal Loss adds an item to the binary classification cross-entropy loss function to attenuate the original cross-entropy loss, which reduces the weight of easily classified samples to focus on the training of difficult samples. At the same time, in order to solve the imbalance of positive and negative samples, a weight factor $\alpha\in[0\ \ 1]$ will be added to each category of the loss function to coordinate the class imbalance. The improved Focal Loss function formula is shown in (8).

$$FL(p_t)=-\alpha(1-p_t)^{\gamma}\log(p_t) \tag{8}$$

where $\gamma\geq 0$ is an adjustable focusing parameter, and the larger the value of γ, the smaller the loss of easily classified samples, A larger γ will expand the samples range which have small loss. When γ>1, Focal Loss can reduce the loss of easy to classify samples, but not much for difficult to classify samples. When γ=0, the Focal Loss formula becomes a cross-entropy loss function.

E. FGM and R-Drop Based on Confrontation Training

1) FGM: Because of the linear characteristics of Neural Network, it is easily attacked by linear disturbance. The concept of Adversarial training [31] is proposed to improve the robustness of the model. Adversarial training is to add a disturbance r_{adv} to the original sample x, and then train the model with the adversarial sample. The disturbance is increasing in the direction of increasing the loss.

The perturbation definition formula is shown as (9).

$$\mathrm{r}_{adv}=\varepsilon\cdot\mathrm{sgn}(\nabla_x L(\theta,x,y)) \tag{9}$$

where sgn() is the sign function and *L* is the loss function. If the input sample is further moved in the direction of rising loss, the resulting adversarial sample can cause greater loss and improve the error rate of the model, so as to meet the requirement that adding small disturbance to the adversarial sample can make the model judgment wrong.

Madry [32] redefined the problem as a saddle point finding problem from the perspective of optimization:

$$\min_{\theta} E(x,y)\sim D[\max_{r_{adv}\in S} L(\theta,x+r_{adv},y)] \tag{10}$$

where *S* is the range of perturbations. θ is the internal parameter of the model, *D* is the distribution of input samples, and x and y correspond to the input and output respectively. There are two parts of this formula: the internal loss function maximization and the external empirical risk minimization. The internal maximization is to obtain the disturbance parameter in the most serious case of model misjudgment, and external risk minimization is to find the parameters that make the model the best robust for internal attacks. The disturbance increased by FGM is:

$$r_{adv}=\varepsilon\cdot\frac{g}{\|g\|_2} \tag{11}$$

$$g=\nabla_x L(\theta,x,y) \tag{12}$$

The new adversarial sample is:

$$x_{adv}=x+r_{adv} \tag{13}$$

2) R-drop: Another way to improve model robustness and generalization is R-drop. Although the traditional Dropout is often used to regulate the training of Deep Neural Networks, its practice of randomly dropping some neurons lacks explanation, resulting in inconsistencies between training and reasoning.

R-drop construes the output consistency of the random submodel due to Dropout through KL divergence. The formula for the R-drop method to apply regular constraints on the output prediction is:

$$L_{KL}^{i}=\frac{1}{2}(D_{KL}(P_i(y_i\mid x_i)\,\|\,P_2(y_i\mid x_i))+D_{KL}(P_2(y_i\mid x_i)\,\|\,P_1(y_i\mid x_i))) \tag{14}$$

Since Dropout randomly drops a few neurons at one time, both probability values of $P_1(y_i\mid x_i)$ and $P_2(y_i\mid x_i)$ are predicted probabilities derived from different submodels of the same model. The difference between these two prediction probabilities is constrained with uses the symmetric KL divergence.

$$L_{NLL}^{i}=-\log P_1(y_i\mid x_i)-\log P_2(y_i\mid x_i) \tag{15}$$

The final loss function is:

$$L_{NLL}^{i}=L_{NLL}^{i}+\alpha\cdot L_{KL}^{i} \tag{16}$$

where α is the coefficient used for control L_{KL}^{i}. Compared with the traditional Dropout, a regular constraint term is introduced to improve the robustness of the model and reduce the inconsistency of the model output.

F. GHM Loss and Dice Loss

Although Focal Loss introduced is considered to be more suitable as a loss function to solve the problem of unbalanced sample data sets, the experimental results of the task in this paper are slightly short of the cross-entropy loss function, which may be due to the phenomenon of category imbalance in the data set. However, in the classification process, it is difficult to classify many samples, which leads to a large loss value of Focal Loss, and the small number of samples in individual categories led to the failure of Focal Loss to give full play to its performance.

GHM Loss is an improved approach based on Focal Loss. Focal Loss is mainly the unbalanced distribution of difficult and easy samples [33], while GHM Loss believes that not all difficult samples are worthy of attention. It adopts a gradient harmonic mechanism and abandons some outliers by formula calculation.

A new gradient norm is introduced by binary cross entropy loss function.

$$L_{CE}(p,p^{*})=\begin{cases}-\log(p) & \text{if } p^{*}=1\\ -\log(1-p) & \text{if } p^{*}=0\end{cases} \tag{17}$$

where p=sigmoid(x) is the category probability of the model prediction sample and p* is the label information. The gradient with respect to x can be calculated by the formula (18).

$$\frac{\partial L_{CE}}{\partial x}=\begin{cases}p-1 & \text{if } p^{*}=1\\ p & \text{if } p^{*}=0\end{cases}=p-p^{*} \tag{18}$$

The gradient norm is defined as:

$$g=|p-p^{*}|=\begin{cases}1-p & \text{if } p^{*}=1\\ p & \text{if } p^{*}=0\end{cases} \tag{19}$$

Intuitively, g represents the distance between the real values and the predicted ones of the sample. The gradient mode norm distribution after convergence of the binary classification model is shown in Fig. 6.

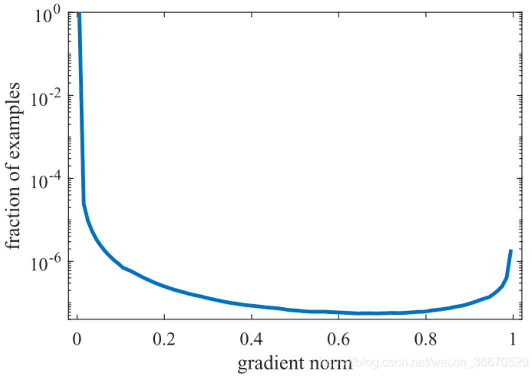

Fig. 6. Gradient norm distribution of the binary classification model.

After logarithmic scaling, the samples close to the y axis are ‘easy to classify’, while those close to the x=1 axis are ‘very difficult to classify’, and the middle parts represent ‘difficult to classify’.

It is standardized according to the proportion of the gradient mode norm of the samples, so that all kinds of samples have a more balanced contribution to the model parameters. Since gradient equalization essentially changes the contribution by weighting the gradients generated by different samples, and the contribution is added to the loss value can also achieve the same effect. The gradient density is defined as the number of samples distributed within the unit value region.

$$GD(g)=\frac{1}{l_{\varepsilon}(g)}\sum_{k=1}^{N}\delta_{\varepsilon}(g_{k},g) \tag{20}$$

where g_k represents the gradient of the k sample, and:

$$\delta_{\varepsilon}(x,y)=\begin{cases}1 & \text{if y-}\frac{\varepsilon}{2}\le x\le y+\frac{\varepsilon}{2}\\ 0 & \text{otherwise}\end{cases} \tag{21}$$

$$l_{\varepsilon}(g)=min(g+\frac{\varepsilon}{2},1)-max(g-\frac{\varepsilon}{2},0) \tag{22}$$

Define the density coordination parameter β:

$$\beta_{i}=\frac{N}{GD(g_{i})} \tag{23}$$

where N represents the number of samples. It can ensure that the weight value is 1 when the distribution is uniform or only one unit area is divided, that is, the loss is unchanged. It can be seen that the weights of samples with high gradient density will decrease. The definition of loss function obtained by applying GHM idea to classification problem is shown as (24).

$$L_{GHM}=\frac{1}{N}\sum_{i=1}^{N}\beta_{i}L_{CE}(p_{i},p_{i}^{*})=\sum_{i=1}^{N}\frac{L_{CE}(p_{i},p_{i}^{*})}{GD(g_{i})} \tag{24}$$

According to (24), the weights of the easily classified negative samples and the extremely difficult classified samples in the candidate samples will be reduced, and the loss value will also be reduced and the impact on model training will be reduced. The weight of the normal hard-to-classify samples will be increased, so that the model will pay more attention to the normal hard-to-classify samples to improve the model performance.

However, GHM Loss is more used in binary classification problem of target detection, and there may be errors in judging extremely difficult samples in multiple classification tasks. At the same time, there might exist a problem that too many hard-to-classify samples are discarded as outliers; will lead to inadequate training. Therefore, the training effect of the model is still inferior to that of the cross-entropy Loss function, and even the experimental effect is lower than that of Focal Loss due to the operation of directly discarding outliers.

Dice Loss, inspired by Dice Coefficient (DSC), was first proposed to apply loss function in the field of image segmentation [33] to solve the problem of pixel sample imbalance in image segmentation.

In this paper, the loss function is transferred to the text classification task to verify whether it can be applied to the field of NLP through experiments. The DSC coefficient is used to evaluate the similarity of two sets.

$$DSC(A,B)=\frac{2|A\bigcap B|}{|A|+|B|} \quad (25)$$

where A represents the set predicted by the model as a positive category, and B represents the set with a true label as a positive class. Dice Loss directly optimizes F1-score:

$$DSC=\frac{2TP}{2TP+FN+FP}=\frac{2\Pr e\times \operatorname{Re}c}{\Pr e+\operatorname{Re}c}=F1 \quad (26)$$

where TP, FP, FN are discrete values, and the DSC coefficient is based on discrete values. For a sample, the continuous form of DSC coefficient can be defined as follows:

$$DSC(x)=\frac{2p_1y_1}{p_1+y_1} \quad (27)$$

where y_1 represents the label value of a positive sample, and p_1 represents the prediction probability of a positive model sample. If the sample is positive, the higher the prediction probability is, the higher the DSC value is. When the sample is a negative class and the molecule is 0. In order to preserve the value of the DSC coefficient of the negative class, the smoothing term γ is added.

$$DSC(x)=\frac{2p_1y_1+\gamma}{p_1+y_1+\gamma} \quad (28)$$

The greater the DSC value is, the more accurate the prediction is, and we convert the formula to a loss.

$$DSC(x)=1-\frac{2p_1y_1+\gamma}{p_1+y_1+\gamma} \quad (29)$$

Since DSC coefficient is a metric function to measure the similarity of two samples, the larger the positive sample p is, the larger the DSC value is, indicating that the more accurate the model prediction is, the smaller the loss should be at this time. Therefore, the final form of Dice Loss is shown as (30).

$$DL=1-\frac{2p_1y_1+\gamma}{p_1^2+y_1^2} \quad (30)$$

Dice Loss has a good performance for the scene where the positive and negative samples are seriously unbalanced, and pays more attention to the prediction of the foreground region in the training process. But the value of training losses is highly volatile.

The reason Dice Loss can solve the problem of sample imbalance is mainly because it is a loss related area , that is, the loss of the current pixel is related not only to the predicted value of the current pixel, but also to other points. The intersection form of Dice Loss can be understood as the operation of mask.

Therefore, no matter how large the picture is, the loss value calculated for a positive sample area of fixed size is the same, and the supervision contribution to the network will not change with the picture size.

Experimental results show that Dice Loss is not as good as cross entropy loss function, which may be due to the fact that the loss value calculated by the loss function is not stable enough. When the loss function is used in the case of only foreground and background, the loss value of a small target will change drastically when there is a partial prediction error, resulting in a drastic change in gradient. For example, considering that there is only one positive sample in the extreme case, as long as the prediction is correct, the loss value will be close to 0, while the loss of the prediction error will be close to 1. The cross entropy loss function is equal in dealing with positive and negative samples and averaging the population, so it is more stable than Dice Loss.

In addition, the above three loss functions are first proposed in the field of image processing, and the difference between image information and text information is also the reason for the unsatisfactory results. Therefore, through experiments, it is believed that the cross-entropy loss function combined with the model can achieve the best results of the experiment in this paper.

IV. Example Analysis

A. Subjects

The subjects of our experiments come from the open-source data set on github website. There are a total of 200,000 legal Q&A pairs of 13 types of questions, which are put forward and recorded by customers who need legal consulting services. We mainly conduct classification experiments through two aspects of problem description and problem classification. After deleting some categories with fewer cases in the data set, a total of 190,000 data are retained in the experiment, corresponding to 11 different categories.

The names of 11 categories and their distribution in the training set and test one are shown in Table I. There are a total of 152,000 descriptions of legal issues. The average length is 23.918, among which the descriptions of legal problems with text length of 28 characters is the most, the descriptions with text length of 32 characters account for 81.377%, the descriptions with 50 characters account for 98.204%, and the ones with 150 characters account for 99.926%.

There are 19,000 descriptions of legal problems in the test set, with an average length of 23.723, among which the descriptions of legal problems with text length of 28 characters are the most, the descriptions with 32 characters account for 81.311%, the descriptions with 50 characters account for 98.437% and the ones within 150 characters account for 99.921%. It can be seen that most legal advice texts are within 32 characters.

TABLE I. DISTRIBUTION OF DATA SET SAMPLES

Category	Set		
	Training set	*Test set*	*Validation set*
Real estate disputes	9272	1159	1159
Traffic accidents	17504	2188	2188
Creditor's rights and debts	16888	2111	2111
Tort	8176	1022	1022
Marriage and family	29624	3703	3703
Company law	7760	970	970
Contract disputes	10608	1326	1326
Demolition and Resettlement	5432	679	679
Medical disputes Labor disputes	5632	704	704
Criminal defense	27072	3384	3384
Real estate disputes	14032	1754	1754

B. Evaluation Index

In this paper, the classification task of legal consulting texts is modeled as a multi-classification problem. In order to compare the performance of the KP-BiLSTM-Att model method proposed and the baseline model approaches, a unified measurement standard is needed. Therefore, we select the evaluation indexes commonly used in Machine Learning to deal with text classification: Accuracy, Precision, Recall and F1-score.

Accuracy $A(i)$ is the proportion of the number of correctly predicted samples to the total number of samples in the dataset:

$$A(i) = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \quad (31)$$

The accuracy rate $P(i)$ is the proportion of correctly predicted positive samples to the predicted positive samples in the dataset.

$$P(i) = \frac{TP_i}{TP_i + FP_i} \quad (32)$$

The recall rate R(i) is the proportion of the sample in which the positive sample was correctly predicted.

$$R(i) = \frac{TP_i}{TP_i + FN_i} \quad (33)$$

The $F1(i)$ value is the weighted harmonic average of the $P(i)$ and $R(i)$.

$$F1(i) = \frac{2 \times \mathrm{Re}call(i) \times \mathrm{Pr}ecision(i)}{\mathrm{Re}call(i) + \mathrm{Pr}ecision(i)} \quad (34)$$

C. Analysis of Experimental Results

From four traditional ML methods, we choose Term Frequency–Inverse Document Frequency (TF-IDF), Naive Bayes model (NB), Support Vector Machine (SVM) and Random Forest (RF) as the baseline methods, and three other models of TextCNN, TextRCNN and Transformer models from Deep Neural Networks.

For the implementation of the benchmark approaches, the classification method based on ML is repeated through sklearn library and XGBoost library. In order to make a fair comparison of their performances, the same word segmentation methods were used to divide the data set. The results of various indicators were shown in Table II, Table III, Table IV and Table V respectively.

Table II shows the comparison of the Accuracy results of the KP-BiLSTM-Att approach proposed in this paper with other benchmark methods on the test set. According to the experimental results, the Accuracy of KP-BiLSTM-Att is higher than others in terms. Ours test result is improved by 4.3% of the performance of TF-IDF algorithm, which has the best performance among Machine Learning text classification algorithms. Ours is improved by 1.2% of TextRCNN which has the best performance among DL text classification algorithms.

In addition, we also tried to use TextRNN model in the experiment, but because the problem description content in the data set was too long, TextRNN could not train long-distance text well, resulting in low accuracy. It also confirmed that the effect of BiLSTM Neural Network is better than the commonly used TextCNN and TextRNN Neural Network structure. According to the Accuracy index alone, KP-BiLSTM-Att approach performs better than other benchmark methods in the task of legal consulting text classification

Table III, Table IV, and Table V show the Precision values, Recall values, and F1-score values of KP-BiLSTM-Att approach and other baseline methods in each category.

Since there is a pair of contradictory indicators for Accuracy rate and Recall rate, it is difficult to reach high values at the same time. Therefore F1-score is finally selected as the model performance evaluation index.

TABLE II. EVALUATION RESULTS BASED ON ACCURACY INDEX

	Approaches							
	TF-IDF	*NB*	*SVM*	*RF*	*Text CNN*	*Text RCNN*	*Transformer*	*Ours*
Accuracy	0.755	0.682	0.721	0.751	0.768	0.787	0.783	**0.798**

TABLE III. EVALUATION RESULTS BASED ON PRECISION INDEX

Category	**Approaches**							
	TF-IDF	***NB***	***SVM***	***RF***	***Text CNN***	***Text RCNN***	***Transformer***	***Ours***
Real estate disputes	0.824	0.573	0.633	0.624	0.650	0.635	0.630	0.654
Traffic accidents	0.658	0.625	0.629	0.550	0.901	0.888	0.895	0.908
Creditor's rights and debts	0.726	0.865	0.825	0.843	0.843	0.808	0.852	0.846
Tort	0.767	0.944	0.944	0.709	0.513	0.540	0.615	0.551
Marriage and family	0.688	0.723	0.723	0.864	0.899	0.906	0.880	0.891
Company law	0.769	0.876	0.876	0.270	0.704	0.630	0.543	0.851
Contract disputes	0.742	0.812	0.912	0.627	0.583	0.657	0.632	0.862
Demolition and Resettlement	0.630	0.655	0.722	0.774	0.856	0.801	0.811	0.839
Medical disputes Labor disputes	0.833	0.713	0.652	0.672	0.705	0.641	0.647	0.804
Criminal defense	0.703	0.867	0.671	0.721	0.821	0.829	0.822	0.869
Real estate disputes	0.821	0.580	0.725	0.588	0.670	0.697	0.733	0.731

TABLE IV. EVALUATION RESULTS BASED ON RECALL INDEX

Category	**Approaches**							
	TF-IDF	***NB***	***SVM***	***RF***	***Text CNN***	***Text RCNN***	***Transformer***	***Ours***
Real estate disputes	0.907	0.347	0.594	0.624	0.638	0.633	0.630	0.626
Traffic accidents	0.383	0.952	0.829	0.550	0.906	0.912	0.895	0.930
Creditor's rights and debts	0.795	0.359	0.813	0.843	0.803	0.827	0.852	0.853
Tort	0.309	0.073	0.693	0.709	0.539	0.520	0.615	0.535
Marriage and family	0.712	0.607	0.668	0.864	0.536	0.917	0.880	0.943
Company law	0.910	0.115	0.574	0.270	0.358	0.402	0.543	0.854
Contract disputes	0.405	0.817	0.912	0.627	0.652	0.580	0.632	0.864
Demolition and Resettlement	0.594	0.394	0.653	0.774	0.789	0.820	0.811	0.837
Medical disputes Labor disputes	0.959	0.785	0.713	0.672	0.493	0.565	0.647	0.802
Criminal defense	0.541	0.962	0.887	0.721	0.903	0.903	0.822	0.865
Real estate disputes	0.810	0.186	0.580	0.588	0.757	0.732	0.736	0.698

According to the experimental results of F1-score, KP-BiLSTM-Att approach proposed in this paper reaches the optimal values in 8 out of 11 categories. In the category of traffic accidents, KP-BiLSTM-Att has increased by 0.8% compared with the baseline model with the best performance. In the category of bond debts, ours has increased by 0.8% compared with the best one; and in the category of marriage and family, this method has increased by 1.4% compared with the best one. In the category of company law, KP-BiLSTM-Att has increased by 3.4% compared with the baseline model with the best performance. Compared with the best-performing baseline model in the contract disputes category, KP-BiLSTM-Att has increased by 2.3%, and compared with the best-performing model in the relocation and resettlement category, ours has increased by 0.3%. In the category of medical dispute, KP-BiLSTM-Att has increased by 13.1%, and it has increased by 0.3% in the category of labor dispute.

In the other three categories that do not reach the optimal value, KP-BiLSTM-Att approach is also stronger than most of the baseline methods, and the weakness stems from an imbalance in the number of samples in different categories. It can be considered that combined with the comprehensive evaluation of Accuracy and F1-score, the superiority of KP-BiLSTM-Att method in the classification task of legal consulting texts is preliminarily confirmed.

Then, weighted average processing is performed for all indicators of the proposed method, and the results are shown in Table VI.

TABLE V. EVALUATION RESULTS BASED ON F1-SCORE INDEX

Category	**Approaches**							
	TF-IDF	***NB***	***SVM***	***RF***	***Text CNN***	***Text RCNN***	***Transformer***	***Ours***
Real estate disputes	**0.864**	0.438	0.612	0.624	0.634	0.634	0.616	0.739
Traffic accidents	0.484	0.754	0.851	0.550	0.899	0.904	0.904	**0.912**
Creditor's rights and debts	0.760	0.508	0.793	0.841	0.825	0.822	0.804	**0.849**
Tort	0.441	0.135	0.627	**0.709**	0.525	0.520	0.443	0.543
Marriage and family	0.700	0.660	0.699	0.864	0.913	0.902	0.918	**0.932**
Company law	0.817	0.205	0.526	0.270	0.474	0.491	0.465	**0.852**
Contract disputes	0.526	0.510	0.839	0.611	0.601	0.622	0.609	**0.862**
Demolition and Resettlement	0.616	0.834	0.640	0.774	0.826	0.811	0.836	**0.839**
Medical disputes Labor disputes	0.672	0.612	0.618	0.611	0.550	0.601	0.557	**0.803**
Criminal defense	0.615	0.793	0.852	0.721	0.864	0.855	0.822	**0.867**
Real estate disputes	0.695	0.296	0.531	0.588	0.711	0.704	**0.735**	0.719

TABLE VI. EVALUATION INDEX BASED ON WEIGHTED AVERAGE

Approaches	**Evaluation Index**			
	Accuracy	**Precision**	**Recall**	**F1-core**
TF-IDF	0.755	0.750	0.755	0.751
NB	0.682	0.691	0.682	0.637
SVM	0.751	0.743	0.751	0.734
RF	0.721	0.712	0.721	0.725
TextCNN	0.768	0.758	0.769	0.769
TextRCNN	0.787	0.781	0.783	0.782
Transformer	0.783	0.777	0.783	0.779
Ours	**0.798**	**0.782**	**0.798**	**0.794**

Combined with the weighted average evaluation indicators, all evaluation indicators of KP-BiLSTM-Att have reached the optimal values, among which Accuracy index has increased by 1.1%, Precision index by 0.1%, Recall index by 1.5%, and F1-score index by 1.2%. The experimental results prove that KP-BiLSTM-Att approach can achieve better classification performance than the traditional ML text classification method and the common reference method of DL text classification Neural Network model.

V. CONCLUSIONS

In this paper, KP-BiLSTM-Att model is proposed and applied to the task of classifying legal consulting texts. The preprocessed data set text is passed into the Keywords Parser, and all the results of word segmentation are weighted by TF-IDF and the result is defined as *X*. Meanwhile, all the parts of speech of the results of word segmentation are marked, processed by [mask], and the result is output as *X*'. *X* and *X*' are passed into the two word embedding layers respectively, and the adversarial training FGM algorithm is added to add disturbance to the vector quantization results to improve the robustness of the model. The R-drop method was introduced into the BiLSTM structure of deep learning to accelerate model training and improve model generalization ability. Then, the Focal Loss function method based on class balance was introduced to solve the unbalanced data set samples in multiple classification tasks and improve the classification accuracy.

The performance of KP-BiLSTM-Att method is compared with seven popular machine learning and deep learning methods. Experimental results show that the Accuracy index of KP-BiLSTM-Att method is 4.7%, 11.6%, 7.7% and 4.3% higher than that of TF-IDF, NB, SVM and RF algorithm in traditional ML classification methods, respectively. Compared to common DL baseline models, the KP-BiLSTM-Att approach offers a 3.0%, 1.1%, and 1.5% improvement over TextCNN, TextRCNN, and Transformer respectively. The test results show that KP+BiLSTM+Attention model has significantly improved the accuracy and F1 value of the best benchmark method for text classification tasks of legal consulting. It can be seen that KP-BiLSTM-Att method has better performance in the classification of legal consulting texts.

There is still some follow-up work for improvement in this research. 1) Different data sets of legal consulting texts and other commonly used data sets in the field of text classification will be presented to verify the feasibility of this model. At the same time, it is considered whether the model can be applied to other related tasks in the legal field through experiment to test whether better results can be obtained. 2) More advanced deep learning models will be used to improve the overall structure and further improve the classification accuracy and operation efficiency of the model. 3) Other methods will be explored to improve the performance of unbalanced data samples and improve the overall classification accuracy of the model.

ACKNOWLEDGMENT

This work is supported by the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province (No. UNPYSCT-2020212), Natural Science Foundation of Heilongjiang Province of China (No. YQ2020G002), and Science Foundation of Harbin Commerce University (No. 2019CX22, No. 18XN064).

REFERENCES

[1] K. He, Prediction model of juvenile football players' sports injury based on text classification technology of ML, Mobile Information Systems, 12 (2021) 1-10.

[2] S. Shah, H. Ge, S.A. Haider, et al, A quantum spatial graph convolutional network for text classification, Computer Systems Science and Engineering, 36 (2021) 369-382.

[3] E.K. Anoual, I Zeroual, The effects of pre-processing techniques on arabic text classification, International Journal of Advanced Trends in Computer Science and Engineering, 10 (2021) 41-48.

[4] J. Atwan , M. Wedyan, Q. Bsoul, et al, The effect of using light stemming for arabic text classification, International Journal of Advanced Computer Science and Applications, 12 (2021) 768-773.

[5] H Amazal, Kissi M, A new big data feature selection approach for text classification, Scientific Programming, 2 (2021) 1-10.

[6] Q. Wang , W. Li, Z. Jin, Review of text classification in deep learning, Open Access Library Journal, 8 (2021) 1-8.

[7] X. Luo, Efficient english text classification using selected ML techniques, AEJ-Alexandria Engineering Journal, 60 (2021) 3401-3409.

[8] C. P. Bara, M. Papakostas, R. Mihalcea, A deep learning approach towards multimodal stress detection, in: AffCon@ AAAI, 2020, pp.67-81.

[9] N. Jaouedi, N. Boujnah, M.S. Bouhlel, A new hybrid deep learning model for human action recognition, Journal of King Saud University-Computer and Information Sciences, 32 (2020) 447-453.

[10] D. Liciotti, M. Bernardini, L. Romeo, E. Frontoni, A sequential deep learning application for recognising human activities in smart homes, Neurocomputing, 396 (2020) 501-513.

[11] G. Diraco, A. Leone, A. Caroppo, P. Siciliano, Deep Learning and Machine Learning Techniques for Change Detection in Behavior Monitoring, in: AI* AAL@ AI* IA, 2019, pp. 38-50.

[12] S. Mirjalili, H. Faris, I. Aljarah, Introduction to evolutionary machine learning techniques, in: Evolutionary Machine Learning Techniques, Springer, 2020, pp. 1-7.

[13] A. Saif, Z.R. Mahayuddin, Moving Object Detection Using Semantic Convolutional Features, Journal of Information System and Technology Management, (2022) 24-41.

[14] A. Saif, Z.R. Mahayuddin, Crowd Density Estimation from Autonomous Drones Using Deep Learning: Challenges and Applications, Journal of Engineering and Science Research, (2021), pp.01-06.

[15] A. Saif, Z.R. Mahayuddin, An Efficient Method for Hand Gesture Recognition using Robust Features Vector, Journal Information System and Technology Management (JISTM), (2021), pp.25-35.

[16] S. Zebhi, S. Almodarresi, V. Abootalebi, Human activity recognition by using MHIs of frame sequences, Turkish Journal of Electrical Engineering & Computer Sciences, 28 (2020) 1716-1730.

[17] A. Saif, Z.R. Mahayuddin, Vision based 3D Object Detection using Deep Learning: Methods with Challenges and Applications towards Future Directions, International Journal of Advanced Computer Science and Applications, (2022).

[18] G. Salton and A. Wong and C.S. Yang, A vector space model for automatic indexing, Communications of the ACM, 18 (1975) 613-620.

[19] W. Pu and H. Wang, A review of question answering systems based on natural language processing, Technology Innovation and Application, 22 (2021) 77-79.

[20] C. Zhang, Research and implementation of tourism question and answer system based on knowledge graph, Guilin University of Electronic technology, 2019, pp. 25-37.

[21] X. Li, Research and application of knowledge question answering syster in education based on knowledge graph, Gilin University, 2019, pp. 18-38.

[22] W. Huang, Deep neural networks for legal question answering based on knowledge graph, University of Chinese Academy of Sciences, 2020, pp. 27-30.

[23] Y. Yu, Y. Fu and X. Wu, Summary of text classification methods, Chinese Journal of Network and Information Security, 5 (2019) 1-8.

[24] D. Liang, Multi classification of Chinese text based on LSTM, Journal of Shanghai University of Electric Power, 36 (2020) 598-602.

[25] Y. Jia, Research and application of text classification technology based on deep learning, Hebei University of Engineering, 2021 pp. 27-35.

[26] N. Sager, C. Friedman and M S. Lyman, Review of medical language processing computer management of narrative data, Boston:Addison-Wesley Longman Publishing Co, Inc, 15 (1987) 195-198.

[27] X. T. Liang and L. Gu, Study on word segmentation and part-of-speech tagging, Computer Technology and Development, 25 (2015) pp. 195-198.

[28] R. Yang, J. Zhang, X. Gao, et al, Simple and effective text matching with richer alignment features, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4699-4709.

[29] T. Y. Lin, P. Goyal, R. Girshick, et al, Focal loss for dense object detection, Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980-2988.

[30] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and Xplaning and Harnessing Adversarial Examples, stat, 2015, pp. 1050: 20.

[31] A. Madry, A. Makelov, L. Schmidt, et al, Towards deep learning models resistant to adversarial attacks, International Conference on Learning Representations, 2017, pp. 1050.

[32] B. Li , Y. Liu , X. Wang. Gradient harmonized single-stage detector, AAAI2019, 2019, pp. 8577-8584.

[33] F. Milletari ,N. Navab ,S A. Ahmadi, V-Net: fully convolutional neural networks for volumetric medical image segmentation, 2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016, pp. 565-571.

[34] A.F.M. Saifuddin Saif, E.D. Wollega, S.A. Kalevela, Spatio-temporal features based human action recognition using convolutional long short-term deep neural network, International Journal of Advanced Computer Science and Applications, 14 (2023) 1-15.

[35] T. Jain, V.K. Verma, A. K. Sharma, et al, Sentiment analysis on COVID-19 vaccine tweets using machine learning and deep learning algorithms, International Journal of Advanced Computer Science and Applications, 14 (2023) 32-41.