Computer Fraud and Security
ISSN (online): 1873-7056

Traditional Chinese Medicine Symptom Text Classification Based on Label Mask

Pengtao Jia[1,*], Qian Ren[1], Jingtao Sha[2]

[1]*College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an, 710054, China*

[2]*Department of Proctology, Xi'an Hospital of Traditional Chinese Medicine, Xi'an, 710021, China*

**Corresponding Author.*

Abstract:
Traditional Chinese Medicine (TCM) symptom text classification refers to the use of computer technology and TCM texts to analyze content and identify different symptom categories, thereby automatically predicting patients' descriptions of their feelings and physical examination results. However, TCM clinical texts are often lengthy and complex, with valuable insights often obscured by noise or redundancy. Additionally, TCM texts contain many obscure words and specialized terms, hindering traditional models from accurately interpreting TCM terminology. To resolve these challenges, this paper first creates a template using label masking, which is incorporated into the input sequence fed into the BERT model. A KAN linear layer is then applied to classify the contextual features extracted by BERT, with the KAN layer adjusting the activation function to better align with the data and enhance classification accuracy. After the original label prediction, a CorNet module is introduced to effectively mine the correlations between labels. F Lastly, the MLM task is included in the training phase, helping the model not only estimate label distributions but also predict labels from masked input positions. This enhances the training process and further improves the model's robustness. Experiments on various datasets show that the proposed model is highly effective and generalizable for real-world clinical data symptom text classification.

Keywords: multi-label text classification, traditional chinese medicine symptom text, bert, masked language model

INTRODUCTION

Medical text categorization is a key application of Natural Language Processing (NLP), particularly for classifying symptoms in clinical records, such as those from traditional Chinese medicine (TCM) doctors treating anorectal diseases. Symptom Text Classification (STC) aims to automatically categorize patient symptom descriptions into predefined categories, helping doctors quickly access relevant information. Traditional classification methods, which assign a single label to each text, are insufficient as patient descriptions often involve multiple, interrelated symptoms. For example, anorectal disease records may describe symptoms like itching, pain, and constipation simultaneously. To address this, the paper frames symptom text classification as a multi-label problem, where each text can be assigned multiple labels. This task involves identifying symptom categories based on patient descriptions, test results, and other clinical data, using natural language processing algorithms. Additionally, while medical entity classification detects and categorizes medical-related phrases, symptom text classification focuses on classifying text at the sentence level. Therefore, this paper proposes the Multi-Label Symptom Text Classification (MSTC) task.

The objective of MSTC is to extract features from symptom texts and assign them to appropriate symptom categories. Unlike single-label models, MSTC allows a symptom text to be assigned multiple labels. However, challenges such as class imbalance, complex label dependencies, and ambiguity in symptom descriptions, particularly in Chinese texts, make this task difficult. Recent improvements in deep learning and Transformer models like BERT have significantly advanced multi-label classification, benefiting clinical research, disease retrieval, and herbal recommendations. In practice, clinical texts are often lengthy and complex, with noise or redundancy obscuring their meaning. Additionally, labels may share subsets of text, creating linguistic correlations that need to be captured. MSTC research focuses on three key areas: effectively capturing semantic patterns, extracting relevant information for each label, and identifying label correlations. Significant progress has already been made in these areas with deep learning techniques.

A standard strategy in multi-label symptom text classification is to treat it as a series of binary classification problems [1]. Methods like CNNs [2] and attention mechanisms [3] also fail to capture these relationships. A label co-occurrence matrix can address this by representing label dependencies, providing valuable insights for datasets with clear co-occurrence patterns [4]. Traditional medical text classification relies on manual feature engineering and machine learning, which struggle with complex texts. Recent deep learning advancements have enabled automatic categorization of symptom texts, improving accuracy in clinical applications. Some methods capture label relationships via label structures, while others treat MSTC as a label generation model [5], or learn label representations [6-7] and model label associations in training data for better predictions [8-9]. However, challenges remain when labels are minimal or absent, and modeling relationships in a complex label space is still difficult.

With the arrival of large-scale pre-trained language models like ELMo [10], BERT [11], and XLNet [12] has greatly advanced text classification and various NLP tasks. BERT, in particular, captures rich linguistic features in its intermediate layers [13], enabling efficient knowledge transfer in NLP. Recent studies show that designing effective prompts for pre-trained models can further enhance their performance [14]. Drawing on the cloze question (CQ) method, this paper introduces the Label Masking Multi-Label Symptom Text Classification model (LM-MSTC), which uncovers latent semantic and associative relationships between labels [15-16]. The model assigns distinct tokens to symptom nouns and creates token prefix templates, which are combined with sentences and input into BERT for classification. During prediction, the model predicts all masked symptom nouns. This approach leverages BERT's ability to capture semantic relationships between symptom nouns and text. To better adapt BERT to medical text structures, this paper proposes a multi-task framework that masks specific label tokens and leverages the Masked Language Model (MLM) to predict them, improving the model's capacity to learn label associations. Our contributions are as follows:

(1) Introducing the LM-MSTC model to uncover latent relationships between symptom nouns and texts, jointly trained with an MLM task for better performance on real-world clinical data.

(2) Using a KAN linear layer to classify high-dimensional contextual features and a CorNet module to capture label correlations, improving prediction accuracy.

(3) Experimental results on different datasets demonstrate the effectiveness of this model in real-world clinical symptom text classification tasks, as well as its strong generalization ability.

RELATED WORK

Multi-Label Text Classification

Multi-label Text Classification (MTC) is a core task in Natural Language Processing (NLP). In recent years, a variety of deep learning methods have been used in research on multi-label classification algorithms, including CNN [2], RNN [17], R-CNN [18], and attention mechanisms. These methods are capable of extracting contextual features from texts. Furthermore, attention mechanisms are frequently employed to extract important features from the text that are associated with the labels. Pre-trained models, such as BERT, have significantly improved the performance of multi-label classification tasks. However, these methods mainly focus on extracting text representations and treat labels as a whole sequence for prediction, without considering the differing contributions of text content and overlooking the correlations between labels. Certain methods approach this by converting multi-label text classification into individual or binary classification problems [1]. The Binary Relevance (BR) method, for instance, considers each label as a distinct binary classification problem, disregarding label dependencies. Other methods leverage pairwise label relationships, such as Pairwise Comparison (RPC), which converts multi-label learning into a label ranking problem through binary preference classification [19]. However, when a label is related to several other labels, utilizing higher-order label dependencies often yields better results. The Classifier Chains (CC) method converts the MTC task into a series of binary classification problems [20], explicitly modeling label correlations by introducing label order. K-labelsets (RAkEL) forms small random label subsets, reinterpreting the MTC task as single-label classification on each subset [21]. Although this method is simple to implement, scalable, and flexible, it still fails to fully leverage label dependencies.

Apart from text modeling, another common method in MTC involves calculating the similarity between feature representations of text and labels through learning label embeddings for classification [22-23]. This type of method is called label-aware methods. It is based on the concept that multi-label documents are treated as mixtures of various label embeddings, with related labels often appearing together in the same or related documents. LW-PT [7] is a powerful label-sensitive approach that trains a label encoder by performing positive and negative document sampling for each label. LSAN [24] is a label-sensitive approach that employs a label attention network, integrating both text and label data, while using an attention mechanism to assess the contribution of each word to the labels. To enhance label semantics, Vu et al. [22] developed a method that incorporates external information from Wikipedia to enhance label embeddings.

Correlations might exist between labels in the MTC task [25]. To capture more abstract label relationships in multi-label classification, some models utilize statistical correlations [26]. However, statistical models often encounter difficulties due to incomplete and noisy label pair co-occurrence patterns in the training data [27]. With the recent advancements in deep learning, some research has utilized sequence learning models to solve the MTC problem, including Sequence Generation Models (SGM) [28]. These models produce a candidate label set using an RNN decoder. However, models based on sequences require finding the best solution in the latent space, and with many labels, the computation becomes time-consuming.

In summary, existing multi-label classification methods focus more on local features and do not fully utilize global information.

Multi-Label Symptom Text Classification (MSTC)

MSTC is designed to automatically categorize the symptom texts provided by patients into predefined symptom categories, thereby helping doctors quickly and accurately obtain relevant information. However, in traditional Chinese medicine (TCM) corpora, there are many rare and specialized terms, making it challenging for traditional models to comprehend the actual semantics of TCM vocabulary. Additionally, the information available in TCM texts is often limited, and traditional methods struggle to make accurate symptom predictions based on the limited textual data. Compared to single-label classification, the challenge of MSTC lies in the potential correlations or dependencies between labels (for example, "itching" and "constipation" often occur together). The model needs to predict multiple labels simultaneously, rather than just a single label.

Some keywords in real-world TCM symptom texts often play a decisive role in the classification results. For example, the sentence "20 years ago, without any obvious cause, I experienced lower abdominal bloating, difficulty defecating, dry stools, and blood streaks in the stool" would be categorized into both constipation and hematochezia. Clearly, terms like "difficulty defecating" are more strongly correlated with constipation than with hematochezia, while "blood streaks" is closely related to hematochezia. To tackle this, Xiao et al. [24] introduced the Label-specific Attention Network (LSAN), a model that integrates both document content and label text, using a self-attention mechanism to assess each word's contribution to the corresponding label. While it achieved good results, it overlooked the correlation between labels. Nam et al. [29], Yang et al. [5], and Qin et al. [30] used Seq2Seq-based methods to establish label correlations and employed attention mechanisms to extract discriminative features from the text. However, traditional single-head attention mechanisms only consider a single layer of semantic information between words and fail to capture comprehensive contextual information. P. Ankit et al. [9] proposed MAGNET, It employs feature and correlation matrices to capture and analyze the key dependencies between labels. However, due to its s strong reliance on building the label correlation matrix, if the correlation matrix between labels cannot effectively capture the correct dependencies, the model may fail to achieve the expected performance.

In this paper, the high-dimensional features from the BERT model are fed into the KAN linear layer. The KAN linear layer, with its learnable activation function, can adaptively adjust during training to accommodate different labels and input patterns. Additionally, the CorNet module is introduced after the original label prediction to effectively mine the correlations between labels and accelerate the model’s convergence. This model can more comprehensively and deeply consider the semantic relationships between text information and labels, thereby improving the classification performance for these labels and further enhancing the prediction accuracy.

METHOD

Based on related research, this paper proposes a MTC model based on label masking, with the overall architecture of the model illustrated in Figure 1.

The model mainly includes two aspects: first, predicting the distribution of multiple labels within the label set; second, predicting the masked label content through the MLM task. Through this multi-task training approach, the model effectively captures the complex interdependencies among labels and significantly enhances both the accuracy and generalization ability of multi-label classification.

First, a template is constructed using label masking (Label Mask, LM), where the positions of the masked labels are marked as MASK, and these templates are incorporated as part of the input sequence and passed to the BERT model. BERT extracts high-dimensional contextual information from the input text, and then the KAN linear layer classifies these features. The KAN linear layer can flexibly adjust the activation function, allowing it to better fit the data and improve classification accuracy. Then, the CorNet module further enhances the correlation between labels, ensuring that label prediction not only depends on individual features but also considers the mutual influence between labels. Finally, the MLM task is incorporated during training, allowing the model to predict the probability distribution of labels as well as the masked labels according to their positions in the input sequence. This strengthens the model's training process and further enhances its generalization ability.

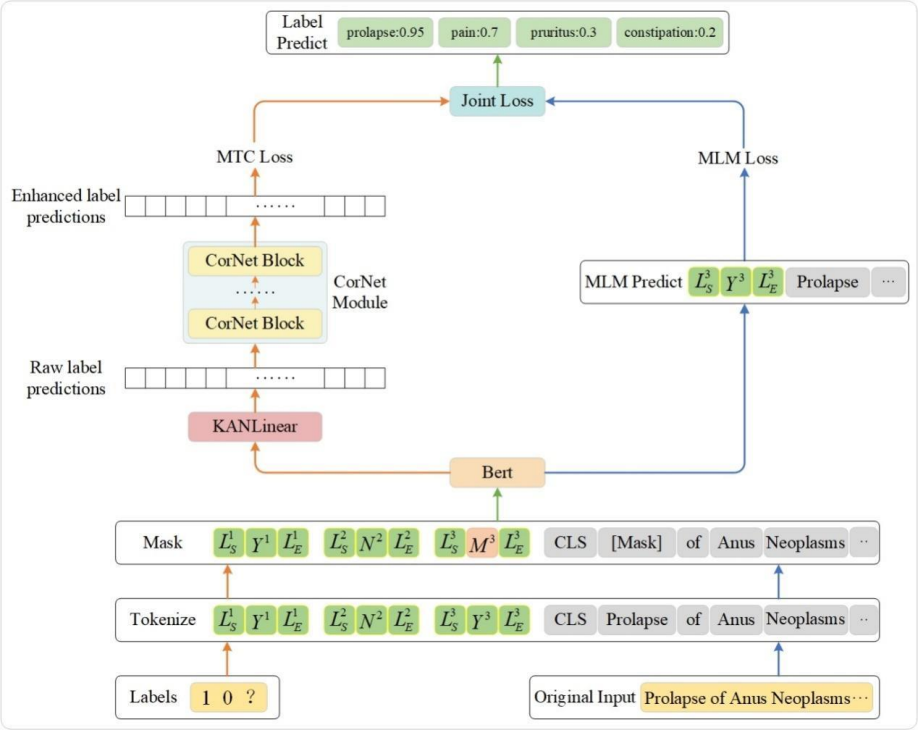


Figure 1. Overall architecture of the model

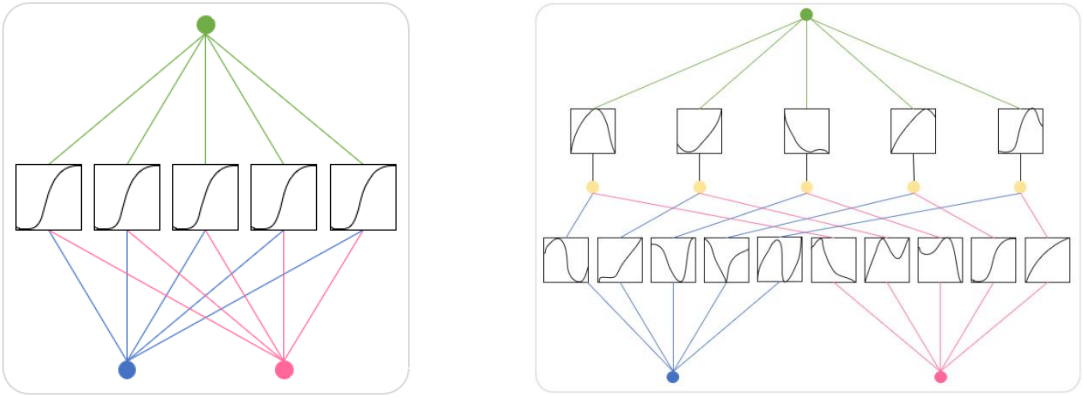
KAN

KAN (Kolmogorov-Arnold Networks) is an innovative neural network architecture, as shown in Figure 2(b). Unlike MLP, KAN does not use linear combination operations; instead, it applies non-linear transformations to each pair of basis elements individually and then combines them into a multi-dimensional space. As shown in Equation (1).

$$f(\mathbf{x}) = f(x_1, \cdots, x_n) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right) \quad (1)$$

Where, $\sum_{p=1}^{n} \phi_{q,p}$ represents the summation of internal functions, $\phi_{q,p}$ denotes the internal function, which is a learnable unary function and represents the activation function on an edge of the network, known as the spline function. $\sum_{q=1}^{2n+1} \Phi_q$ represents the summation of external functions, which are also learnable, and Φ_q is the external function that combines the outputs of the internal functions to generate the final output.

The high-dimensional features obtained from the BERT model are passed into the KAN linear layer, which contains rich contextual information and potential relationships among the labels. In text classification tasks, labels often exhibit complex interdependencies, and there may also be issues such as label imbalance or short descriptions for certain labels. Traditional fully connected layers are often limited in handling these issues because they struggle to flexibly adjust activation functions to better fit these complex patterns, leading to insufficient learning for certain labels and an inability to effectively capture key features. The KAN linear layer, through learnable activation functions, can adaptively adjust during training to accommodate different labels and input patterns, thereby improving the classification performance for these labels and further enhancing prediction accuracy.



(a) MLP (b) KAN

Figure 2. MLP and KAN network architecture diagram

Cornet Module

The structure of the CorNet [31] module is shown in Figure 3, and the function expressions are given in Equations (2) and (3). Here, x represents the original label prediction, y represents the enhanced label prediction, F(x) denotes the output of CorNet, W denotes the weight matrix, and b stands for the bias term, δ and σ stand for the Sigmoid function and the Exponential Linear Unit (ELU), respectively.

$$y = F(x) + x \tag{2}$$

$$F(x) = W_2\delta(W_1\sigma(x) + b_1) + b_2 \tag{3}$$

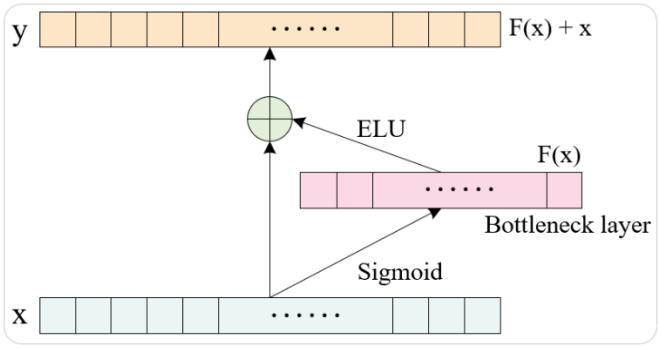


Figure 3. CorNet network structure

Traditional deep learning methods typically map features to the label space through fully connected layers, but they overlook the relationships between labels. By introducing CorNet after the original label prediction, this model effectively captures the correlations between labels and accelerates the convergence process. This is particularly important in multi-label classification tasks, where label dependencies are often strong. CorNet provides a flexible way to model these dependencies, further improving prediction accuracy. Additionally, increasing the number of CorNet layers can alleviate the vanishing gradient problem to some extent, allowing the

network to be effectively trained at deeper layers, thus improving the model's performance and stability in handling complex tasks.

Label Mask

In MTC tasks, documents typically have varying numbers of true labels, making it impractical to construct specific templates for each label. Therefore, a template system has been established to handle the entire label space. In this system, each label position has three possible states: 0, 1, and MASK, representing different label statuses. Additionally, by introducing position-based prompts, the BERT model is able to clearly identify the position of each label. These prompts are used as prefixes for the label sequence and are passed into the BERT model along with the original text during training. After the template is generated, the label sequence is presented in a specific format, where LS represents the label's starting position and LE signifies its ending position. For example, for a label sequence containing [0, 1, MASK], the generated template would clearly mark the status of each label: [LS − 1] [NO − 1] [LE − 1], [LS − 2] [YES − 2] [LE − 2], [LS − 3] [MASK − 3] [LE − 3].

The model's objectives during training include two aspects: first, predicting the probability distribution over different labels within the label space; second, predicting the masked content through the MLM task. After the BERT output is processed by a fully connected layer, the label distribution predictions and masked predictions are obtained. This approach allows the model to capture label correlations effectively, while improving both the accuracy and generalization of multi-label classification. This paper employs Binary Cross-Entropy (BCE) as the loss function for MTC and uses Cross-Entropy for the MLM task. The BCE loss function is shown in Equation (4), and Equation (5) shows the Cross-Entropy loss function.

$$\mathcal{L}_{mtc} = \frac{1}{|L|}\sum_{i=1}^{|L|} (y_{ti} log(\sigma(y_{pi})) + \&(1 - y_{ti}) log(1 - \sigma(y_{pi})) \quad (4)$$

$$L_{mlm} = -\sum_{i=1}^{V} y_i \log (p_i) \quad (5)$$

Here, σ denotes the sigmoid activation function, yt and yi denotes the true labels, and yp and pi denote the predicted results. Equation (6) presents the final joint loss function.

$$\mathcal{L}_{MSTC} = \mathcal{L}_{mtc} + \lambda \mathcal{L}_{mlm} \quad (6)$$

EXPERIMENTS AND RESULTS ANALYSIS

Data Preparation and Preprocessing

The study gathered clinical records from the Department of Colorectal Diseases at Xi'an City Traditional Chinese Medicine Hospital, spanning from March 2021 to March 2024. From these records, 6,575 symptom descriptions were extracted. Preprocessing of the Chinese medical treatment records included the elimination of stop words, the application of the BERT model to identify and remove highly similar texts, and the adjustment of text lengths to ensure uniformity, with an average text length not exceeding 300 words. This process resulted in 5,947 valid symptom descriptions, which were compiled into the dataset named XHTCM. Based on the "Colorectal Diseases Outpatient Electronic Medical Record Template" provided by Xi'an City Traditional Chinese Medicine Hospital, the valid symptom descriptions were categorized into 16 classes: constipation, hematochezia, diarrhea, abdominal distension, abdominal pain, swelling and pain, pain, prolapse, pruritus, abnormal discharge, hyperplasia, erosion, eczema, swelling, a sensation of heaviness, and a feeling of incomplete evacuation.

Dataset

The experiments used the real-world data set of Xi'an Traditional Chinese Medicine Hospital (XHTCM) and Reuters-21578 obtained by desensitization. The dataset's basic information is provided in Table 1.

Table 1. Dataset Statistical Information

Dataset	Total samples	Train	Val	Test	Number of labels
Reuters-21578	10789	5827	1943	3019	90
XHTCM	5947	4757	595	595	16

Evaluation Index

In this paper, Hamming Loss and micro-F1 Score are the main evaluation metrics, with Accuracy and Micro-Jaccard used for further assessment. The calculation of Hamming Loss is given in Equation (7), where a smaller Hamming loss value is better.

$$HL = \frac{1}{|D|}\sum_{i=1}^{D} \frac{xor(x_i, y_i)}{|L|} \tag{7}$$

Where xor is the operation, xi and yi represent the real and predicted values, |D| donates the total count of samples in the dataset, and |L| donates the total count of labels.

The accuracy is calculated as shown in Equation (8), where higher accuracy values are better.

$$Accuracy = \sum_{i}^{|\Gamma|} \frac{\Xi(Y_{ti}, Y_{pi})}{|\Gamma|} \tag{8}$$

Here, |Γ|is the size of the test set, and$\Xi(\cdot)$ is the indicator function. If all elements in Yti and Ypi match at each position, thenΞ(Yti,Ypi)=1, otherwise, Ξ(Yti,Ypi)=0.

Baseline Models

CNN [32]: A CNN-based model is employed for feature extraction from text and to produce a probability distribution across different labels.

CNN-RNN [28]: A combination of CNN and RNN is employed to extract text features and generate the label distribution within the label space.

SGM [5]: The Multi-label Classification Task (MCT) is treated as a sequence generation problem in SGM, where a novel decoder-based sequence generation model is implemented to address the task effectively.

MEGNET [9]: A model leveraging the Graph Attention Network (GAT) framework, which can capture dependencies between labels and model label dependencies using a graph structure and attention mechanism.

LW-LSTM+PT [7], LW-LSTM+FT: Document representations are acquired with label-aware details utilizing pre-trained (PT) models, which are then fine-tuned (FT) to cater to various downstream tasks. PT refers to the initial pre-training phase, while FT signifies the subsequent fine-tuning process for targeted downstream applications.

BERT [11]: A pre-trained language model, that can be adapted and fine-tuned to enhance performance across a variety of downstream tasks.

Experiment

Experimental environment and hyperparameter settings

The particular aspects of the experimental hyperparameters are presented in Table 2, which represent the optimal parameters selected after multiple experiments.

Table 2. Experimental hyperparameter configurations

Hyperparameters	Values
Learning Rate	5e-5
Batch Size	16
Epochs	40
Warm-up Epochs Ratio	0.1
Mask Probability of MLM	0.15
Optimizer	AdamW

Ablation experiment

Ablation experiments were conducted on the Reuters-21578 and Xi'an Traditional Chinese Medicine Hospital datasets to assess the effectiveness of the improvements in the LM-MSTC model. The results are shown in Tables 3 and 4.

Table 3. Ablation experiment results (Reuters-21578)

Bert	LM	MLM	KAN	CorNet	Accuracy (+, %)	Micro-F1(+, %)	Micro-Jaccard (+, %)	HL (-)
√					84.71	88.62	80.3	0.0031
√	√				84.12	89.1	80.83	0.0029
√		√			84.95	89.04	80.72	0.00295
√	√	√			84.9	89.2	81.13	0.0027
√	√	√	√		85.13	89.45	81.24	0.0028
√	√	√	√	√	85.51	89.61	81.65	0.0028

Table 4. Ablation experiment results (XHTCM)

Bert	LM	MLM	KAN	CorNet	Accuracy (+, %)	Micro-F1 (+, %)	Micro-Jaccard (+, %)	HL (-)
√					86.88	96.1	93.35	0.0141
√	√				86.56	96.56	93.82	0.014
√		√			87.1	96.4	93.73	0.0145
√	√	√			87.39	96.89	93.92	0.0116
√	√	√	√		87.52	96.85	93.95	0.0122
√	√	√	√	√	87.75	97.03	93.94	0.0113

From the ablation experiment results, it can be seen that in the Reuters-21578 dataset, the model with the integrated methods, except for the slightly higher Hamming loss compared to the method without the KAN and CorNet modules, outperforms the other methods in all other metrics. In the Xi'an Hospital of Traditional Chinese Medicine dataset, the integrated models show significant improvements across all metrics, thereby validating the performance of each model component.

Comparative experiment

To validate the effectiveness of the proposed model, we compare it with models such as CNN, CNN-RNN, SGM, MEGNET, LW-LSTM+PT, LW-LSTM+FT, and BERT on the Reuters-21578 and Xi'an Traditional Chinese Medicine Hospital datasets. The comparison results are shown in Table 5.

Table 5. Experimental results comparison

Model	Reuters-21578				XHTCM			
	Accuracy (+,%)	F1 (+,%)	Jaccard (+,%)	HL (-)	Accuracy (+,%)	F1 (+,%)	Jaccard (+,%)	HL (-)
CNN	80.5	85.3	77.9	0.0287	81.08	93.9	92.1	0.0192
CNN-RNN	80.6	84.5	77.36	0.0282	80.25	93.51	91.89	0.0201
SGM	83.24	87.1	78.82	0.0277	84.2	94.62	92.35	0.0188
MEGNET	85.32	89.51	81.42	0.0265	87.43	96.75	93.76	0.0163
LW-LSTM+PT	83.41	87.21	78.86	0.0271	84.13	94.52	92.65	0.0185
LW-LSTM+FT	83.46	87.34	78.91	0.0273	85.56	94.83	93.14	0.0151
BERT	84.71	88.62	80.3	0.0031	86.88	96.1	93.35	0.0141
LM-MSTC	85.51	89.61	81.65	0.0028	87.75	97.03	93.94	0.0113

From the table, the model proposed in this paper outperforms others on all metrics—accuracy, F1 score, Jaccard, and Hamming loss—on the Reuters-21578 dataset. Compared to the traditional CNN model, accuracy increased by 5.01%, F1 score improved by 4.31%, and Jaccard rose by 3.75%, and Hamming loss decreased by 90.2%. Compared to the baseline BERT model, accuracy improved by 0.8%, F1 score increased by 0.99%, Jaccard rose by 1.35%, and Hamming loss decreased by 9.7%, and Hamming loss decreased by 9.7%. The model also performed excellently on the Xi'an Traditional Chinese Medicine Hospital dataset, where compared to the baseline BERT model, accuracy increased by 0.87 percentage points, F1 score increased by 0.93%, Jaccard improved by 0.59%, and Hamming loss decreased by 19.9%. These experiments show that the proposed model exhibits good generalization ability and delivers outstanding performance in traditional Chinese medicine text classification.

It can be observed that models like CNN, CNN-RNN, LW-LSTM+PT, LW-LSTM+FT, etc., performed poorly on both datasets, showing a large gap in classification accuracy compared to BERT and other models. The large-scale pre-trained BERT language model, compared to traditional static language models, not only achieves dynamic word vector representations but also provides higher semantic accuracy for the word vectors. Furthermore, the Xi'an Traditional Chinese Medicine Hospital dataset contains many specialized Chinese medicine terms. Language models that have not been pre-trained on large-scale corpora face difficulties in understanding Chinese medicine texts, resulting in inaccurate semantic representations of the texts and thus leading to incorrect predictions by the model.

Additionally, machine learning-based algorithms like SGM also performed poorly. This indicates that for more complex tasks, machine learning methods struggle to extract accurate text semantics based on statistical rules. The field of traditional Chinese medicine has its own unique specialized knowledge and rules, which are difficult for conventional machine learning models to interpret and grasp. The LM-MSTC, utilizing large-scale pre-trained language models, achieved superior prediction results, highlighting the crucial role of pre-training in multi-label text classification for traditional Chinese medicine.

CONCLUSION

This paper examines the limitations of deep learning in TCM clinical text classification and introduces an improved MTC model with label masking, a KAN linear layer, and the CorNet label enhancement module. Experimental results on the Xi'an TCM Hospital dataset indicate that the proposed model excels over other mainstream models in several metrics. Furthermore, the improved model demonstrates good generalization ability on public datasets. However, the proposed model has a large number of hyperparameters, which require extensive experimentation and repeated verification, making the process of determining the optimal hyperparameter combination complex and time-consuming. Future work will focus on researching and designing an adaptive optimization method that enables the model to autonomously choose the optimal hyperparameter combination according to data characteristics and task requirements, thus reducing manual intervention and experimental costs while improving the model's practicality and deployment efficiency.

ACKNOWLEDGMENTS

This research was funded by Scientific and Technology Program Funded by Xi'an City (Program No. 22YXYJ0009).

REFRENCES

[1] Boutell M R, Luo J, Shen X, et al. Learning multi-label scene classification. Pattern recognition, 2004, 37(9): 1757-1771.

[2] Liu J, Chang W C, Wu Y, et al. Deep learning for extreme multi-label text classification//Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval. 2017: 115-124.

[3] Yang Z, Yang D, Dyer C, et al. Hierarchical attention networks for document classification//Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016: 1480-1489.

[4] Yao J, Wang K, Yan J. Incorporating label co-occurrence into neural network-based models for multi-label text classification. IEEE Access, 2019, 7: 183580-183588.

[5] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, H. Wang, Sgm: Sequence generation model for multi-label classification, COLING (2018) 3915-3926.

[6] Pappas N, Henderson J. Gile: A generalized input-label embedding for text classification. Transactions of the Association for Computational Linguistics, 2019, 7: 139-155.

[7] Liu H, Yuan C, Wang X. Label-wise document pre-training for multi-label text classification//Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I 9. Springer International Publishing, 2020: 641-653.

[8] Zhu Y, Kwok J T, Zhou Z H. Multi-label learning with global and local label correlation. IEEE Transactions on Knowledge and Data Engineering, 2017, 30(6): 1081-1094.

[9] Pal A, Selvakumar M, Sankarasubbu M. MAGNET: Multi-Label Text Classification using Attention-based Graph Neural Network.arXiv, 2020. DOI: 10.5220/0008940304940505.
[10] Peters M, Neumann M, Iyyer M, et al. Deep Contextualized Word Representations//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018. DOI: 10.18653/v1/N18-1202.
[11] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 6. long and short papers: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019), 2-7 June 2019, Minneapolis, Minnesota, USA. 2019. DOI: 10.48550/arXiv.1810.04805.
[12] Yang Z, Dai Z, Yang Y, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding//Advances in Neural Information Processing Systems 32, Volume 8 of 20: 32nd Conference on Neural Information Processing Systems (NeurIPS 2019). Vancouver (CA). 8-14 December 2019. 2020.
[13] Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language? //ACL 2019-57th Annual Meeting of the Association for Computational Linguistics. 2019.
[14] Ding N, Chen Y, Han X, et al. Prompt-learning for fine-grained entity typing. Ar Xiv, 2021. DOI: 10.48550/arXiv.2108.10604.
[15] Schick T, Hinrich Schütze. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. 2020. DOI: 10.48550/arXiv.2009.07118.
[16] Schick T, Schuetze H. Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference//16th Conference of the European Chapter of the Association for Computational Linguistics.Association for Computational Linguistics, 2021. DOI: 10.18653/V1/2021.EACL-MAIN.20.
[17] Liu P, Qiu X, Huang X. Recurrent Neural Network for Text Classification with Multi-Task Learning. AAAI Press, 2016. DOI: 10.48550/arXiv.1605.05101.
[18] Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification//Proceedings of the AAAI conference on artificial intelligence. 2015, 29(1).
[19] Hüllermeier E, Fürnkranz J, Cheng W, et al. Label ranking by learning pairwise preferences. Artificial Intelligence, 2008, 172(16-17): 1897-1916.
[20] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification//Springer US. Springer US, 2011: 333-359. DOI: 10.1007/s10994-011-5256-5.
[21] Tsoumakas G, Vlahavas I. Random k-labelsets: An ensemble method for multilabel classification//European conference on machine learning. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007: 406-417.
[22] Vu H T, Nguyen M T, Nguyen V C, et al. Label-representative graph convolutional network for multi-label text classification. Applied Intelligence, 2023, 53(12): 14759-14774.
[23] Chen Z, Liu Y, Cheng B, et al. Integrating label semantic similarity scores into multi-label text classification//International Conference on Artificial Neural Networks. Cham: Springer Nature Switzerland, 2022: 234-245.
[24] Xiao L, Huang X, Chen B, et al. Label-specific document representation for multi-label text classification//Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). 2019: 466-475.
[25] Rastogi R, Mortaza S. Multi-label classification with missing labels using label correlation and robust structural learning. Knowledge-Based Systems, 2021, 229: 107336.
[26] Huang J, Li G, Huang Q, et al. Learning label-specific features and class-dependent labels for multi-label classification. IEEE transactions on knowledge and data engineering, 2016, 28(12): 3309-3323.
[27] Zhang Q W, Zhang X, Yan Z, et al. Correlation-Guided Representation for Multi-Label Text Classification. //International Joint Conference on Artificial Intelligence.International Joint Conferences on Artificial Intelligence Organization, 2021. DOI: 10.24963/IJCAI.2021/463.
[28] Chen G, Ye D, Xing Z, et al. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization//2017 International joint conference on neural networks (IJCNN). IEEE, 2017: 2377-2383.

[29] NAM J, MENC A E, KIM H, et al. Maximizing Subset Accuracy with Recurrent Neural Networks in Multi-label Classification//Advances in Neural Information Processing Systems. 2017: 5419-5429.
[30] Qin K, Li C, Pavlu V, et al. Adapting RNN Sequence Prediction Model to Multi-label Set Prediction//2019. DOI: 10.48550/arXiv.1904.05829.
[31] Xun G, Jha K, Sun J, et al. Correlation networks for extreme multi-label text classification//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 1074-1082.
[32] Y. Kim, Convolutional neural networks for sentence classification, EMNLP (2014): 1746-1751.