Automation and Machine Learning (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2025.060101
ISSN 2516-5003 Vol. 6 Num. 1

Research on Automatic Text Categorization Method Based on Knowledge Enhancement

Wenwen Zhao

Xinjiang Hetian College, Hotan, Xinjiang, China

Keywords: Knowledge enhancement; automatic text categorization

Abstract: In today's information explosion, automatic text categorization has become an important task in text processing, which not only helps users to quickly and effectively distinguish and manage massive amounts of information, but also greatly facilitates the development of fields such as opinion analysis and sentiment recognition. It is worth mentioning that the emergence of deep learning models has brought unprecedented changes to automatic text categorization, especially with the promotion of convolutional neural networks (CNN), long short-term memory networks (LSTM) and BERT models, the accuracy of text categorization has been significantly improved. However, with the advancement of technology, traditional models still face many challenges when dealing with complex text tasks, such as the lack of domain knowledge and insufficient comprehension of long texts. Therefore, automatic text categorization methods based on knowledge enhancement have emerged to make up for the deficiencies of traditional models and improve the classification performance and robustness by introducing an external knowledge base. In this paper, we will explore the application of different knowledge enhancement strategies in text categorization, focusing on the structure of the model framework and its functions of each layer, aiming to provide researchers with new ideas and technical support.

1. Introduction

With the rapid development of internet technology, the sources and varieties of textual information are becoming increasingly rich. How to efficiently and accurately classify these texts has become an urgent problem to be solved. Traditional text classification methods, such as those based on the bag-of-words model and Naive Bayes, still have certain application values in some tasks, but their understanding of text content is limited, making it difficult to cope with complex and ever-changing text scenarios. In recent years, breakthroughs in deep learning technology, particularly Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM), and BERT models, have brought revolutionary changes to text classification. These models improve classification accuracy by learning complex features in text, but they still have some limitations, such as lack of domain knowledge and insufficient understanding of long texts. Knowledge-enhanced text classification methods, by introducing external knowledge bases such as dictionaries, common sense graphs, and domain-specific professional knowledge, can effectively make up for these deficiencies and further improve classification performance [1]. This article will

delve into the application of different types of knowledge-enhanced strategies in text classification and provide a detailed description of the proposed model framework structure, aiming to provide new ideas and technical support for researchers in this field.

2. Research on Text Categorization Algorithms

2.1. Convolutional Neural Network Algorithm

Convolutional Neural Networks (CNNs) have achieved remarkable success in image recognition tasks due to their efficient feature extraction capabilities. In recent years, CNNs have also demonstrated strong performance in text classification tasks. Unlike traditional bag-of-words approaches, CNNs can capture local features and contextual information in text. Through multiple layers of convolution and pooling operations, CNNs progressively extract higher-level abstract features, as illustrated in Figure 1. For text classification tasks, the input to a CNN is typically a matrix of word embeddings, where each word is represented as a vector. Convolutional operations enable the capture of phrase-level features, while pooling layers further compress these features to extract the most important information.

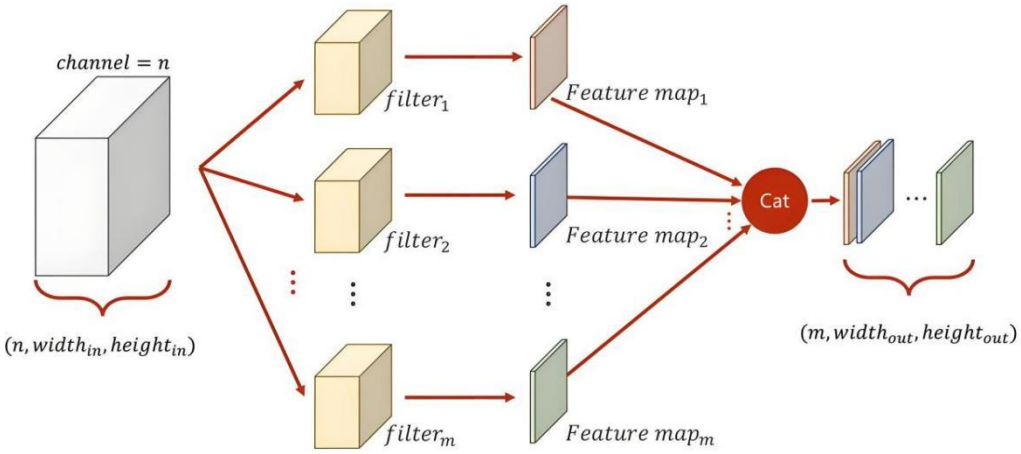

Figure 1. Principle of Convolutional Neural Network (CNN)

Specifically, assume that the input text is represented as a word embedding matrix $X \in R^{d \times n}$, where d is the dimension of the word embedding and n is the length of the text. The convolutional layer performs a convolution operation on the input matrix by means of multiple convolution kernels $W \in R^{d \times k}$ to generate the feature map C:

$$C = \mathrm{Re}\,LU(X * W + b) \tag{1}$$

where * denotes the convolution operation, b is the bias term, and ReLU is the activation function. The pooling layer (usually using maximum pooling) then further compresses the feature map to generate fixed-length feature vectors for classification in the subsequent fully-connected layer, the process of which is shown in Figure 2.

CNN excel in capturing local features and patterns in text classification tasks, particularly for short texts and simpler classification problems. However, their effectiveness may be limited when dealing with longer texts and tasks that require understanding global information, as convolutional operations primarily focus on local features without modeling the overall structure of the text. Therefore, incorporating more contextual information and global features into CNNs has become an important research direction. Studies have shown that techniques such as multi-scale convolutions and dynamic pooling can significantly improve CNN performance in long-text classification tasks.

Additionally, combining CNNs with other deep learning models, such as LSTMs, and integrating external knowledge bases are effective approaches to enhance their classification capabilities [2].

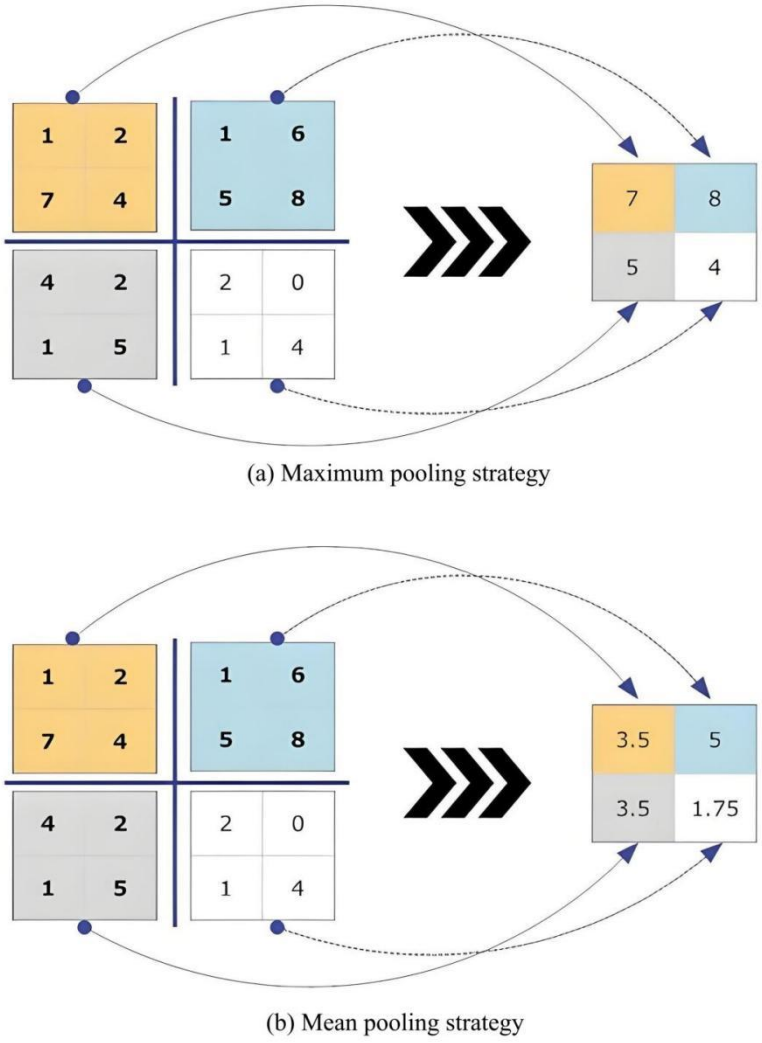

Figure 2. Pooling process

2.2. Long and short-term memory network algorithm

Long Short-Term Memory networks (LSTMs) are a type of neural network specifically designed to handle sequential data, effectively addressing the issues of vanishing and exploding gradients that traditional Recurrent Neural Networks (RNNs) face when dealing with long sequences Figure 3. LSTMs incorporate gating mechanisms that allow them to selectively remember and forget information, thereby better capturing long-term dependencies in text data [3]. Their application is particularly prevalent in text classification tasks, as textual data inherently possesses temporal sequential characteristics, where the relationships between preceding and succeeding words significantly impact the classification outcome.

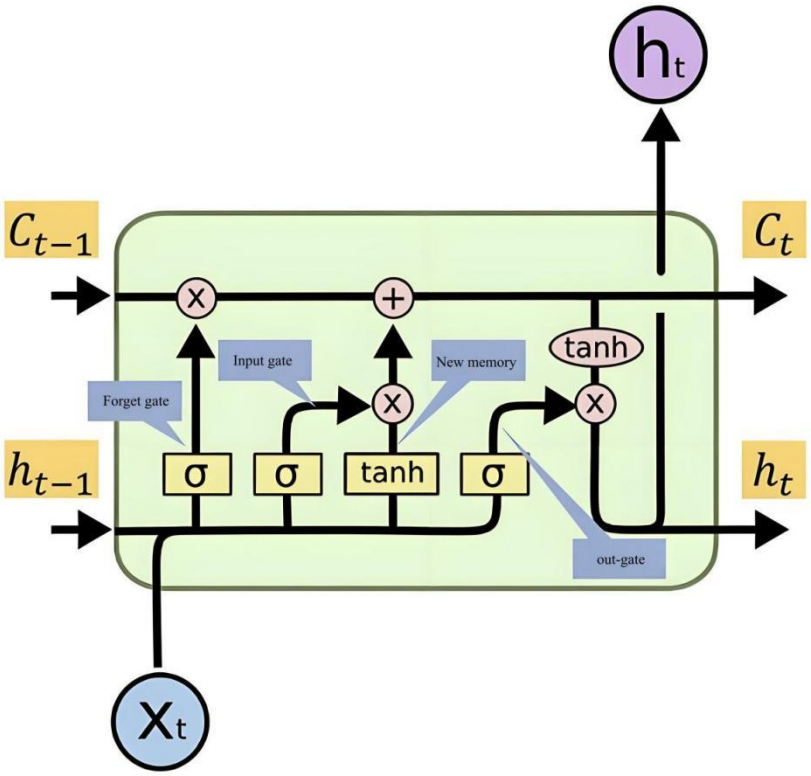

Figure 3. Long and short term memory network (LSTM)

The core structure of the LSTM consists of input gates, forgetting gates, and output gates, and these gating mechanisms allow the LSTM to flexibly control the flow of information at each step. The input gate determines how much of the current input information will be stored into the cell state, the forgetting gate determines which information in the cell state will be forgotten, and the output gate determines which information in the cell state will be output. Specifically, the state update formulas of LSTM are shown in (2)-(7):

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{2}$$

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{3}$$

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \tag{4}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{5}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t \odot \tanh(C_t) \tag{7}$$

where i_t, f_t and ot denote the outputs of the input, forgetting and output gates, respectively, $\tilde{C}_t$ denotes the candidate cell state, C_t denotes the current cell state, h_t denotes the current hidden state, x_t denotes the input at the current moment, and h_{t-1} denotes the hidden state at the previous moment, VVV denotes the Sigmoid activation function, tanh denotes the hyperbolic tangent activation function, and $\odot$ denotes the element-level multiplication. These formulas demonstrate that the LSTM flexibly controls the storage and transfer of information through the gating mechanism, thus effectively capturing long-term dependencies in the text [4].

2.3. BERT model

BERT (Bidirectional Encoder Representations from Transformers) model is a pre-training model that has attracted much attention in the field of natural language processing in recent years, as shown in Figure 4. Unlike traditional models based on word embedding and recursive neural networks, BERT is able to capture complex contextual relationships in text by using the Transformer structure, and the core mechanism of the Transformer model is the self-attention mechanism, which dynamically calculates the importance of each word in a sentence and performs information transfer and feature extraction based on this importance information. BERT's bi-directional encoder effectively solves the problem that traditional models can only transmit information in one direction by considering contextual information in the pre-training phase, thus achieving significant performance improvement in multiple NLP tasks [5].

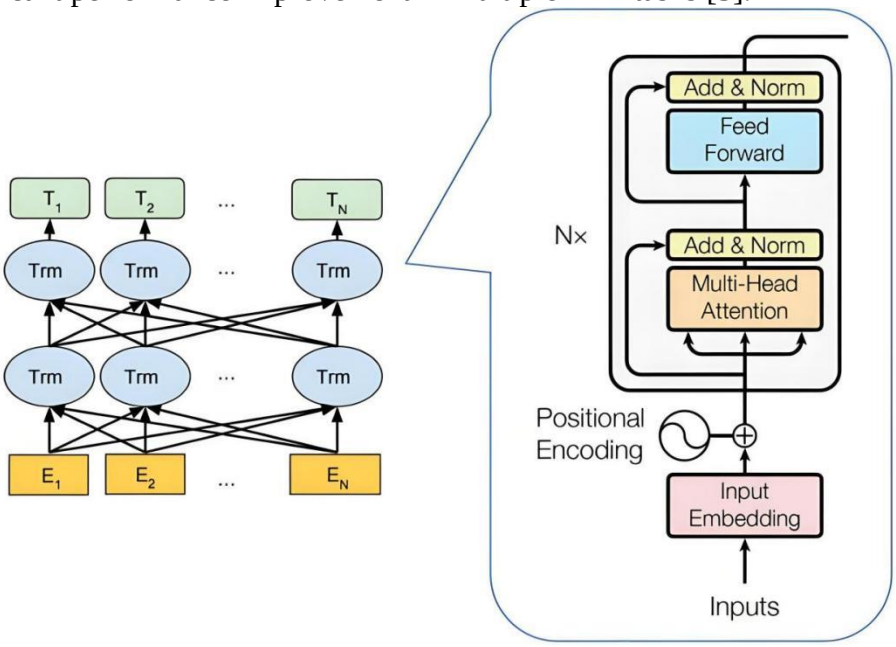

Figure 4. The BERT model

The self-attention mechanism of the BERT model can be expressed as Equation (8):

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{8}$$

Where Q, K and V denote the query matrix, key matrix and value matrix respectively, dk is the dimension of the key, and the softmax function is used to generate the attention weight matrix so as to dynamically assign weights to each word. This mechanism allows BERT to fully take into account the relationships between words during encoding and therefore performs well in understanding the global information of a sentence [6].

The BERT model uses two main tasks in the pre-training phase: masked language modeling (MLM) and next sentence prediction (NSP). The MLM task learns to infer word meanings from contextual information by randomly masking certain words in the input sentence and then predicting these masked words; The NSP task, on the other hand, enables the model to learn to understand the logical associations between sentences by predicting the relationship between two sentences. The design of these tasks greatly enhances the generalization ability of BERT in real-world tasks.

3. Text Classification Model Based on Knowledge Enhancement

3.1. Model framework structure

The classification model constructed herein is comprised of a data augmentation layer, a text embedding layer, a feature extraction layer, a knowledge integration layer, and an output layer. The data augmentation layer enriches the dataset by incorporating external knowledge, thereby generating a novel dataset encompassing the original text, external knowledge, and corresponding labels. This layer serves to augment the volume and diversity of training data, thereby enhancing the model's generalization capabilities. Subsequently, the raw text and external knowledge are processed by the text embedding layer, which transforms them into dense vector representations through word embedding techniques. Employing pre-trained word vector models such as Word2Vec or BERT, the embedding layer generates vectors for each word, capturing semantic information in the process [7]. The feature extraction layer leverages neural network models, including convolutional neural networks or long short-term memory networks, to extract features from both the text and external knowledge. These features encapsulate not only local information but also global contextual insights. Following this, the features are directed into the knowledge integration layer, where gating mechanisms within the neural network regulate the proportion of fusion between the original text features and external knowledge features, yielding a final textual feature representation. These gating mechanisms dynamically adjust the fusion ratio based on the input context, ensuring that the model effectively leverages external knowledge across varying scenarios. Ultimately, the fused data is transmitted to the output layer, where a multilayer perceptron is employed for classification, producing the final classification outcome. This model framework, characterized by its multi-layered feature extraction and knowledge integration, significantly enhances the precision and robustness of text categorization.

3.2. Data enhancement layer

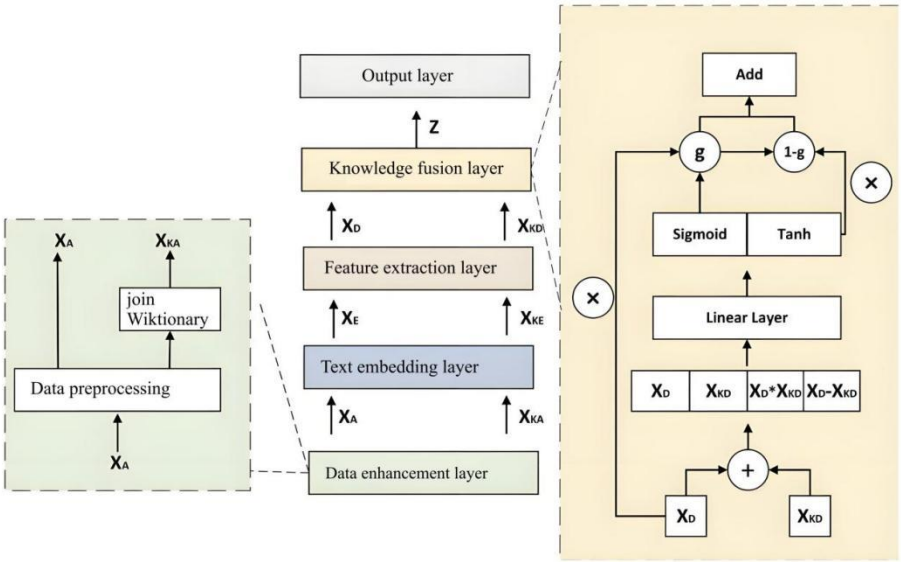

Figure 5. Diagram of text categorization model based on knowledge enhancement

The data enhancement layer is introduced with the aim of improving the quality of the dataset through external knowledge, which in turn enhances the performance of the language model. The methodology consists of searching the words in the dataset using a natural language processing tool

to find the corresponding paraphrase of each word in Wiktionary. These meanings are added to the dataset as external knowledge, and together with the original text and tags, form the new training data.Wiktionary contains more than 999,614 entity descriptions, and in order to ensure the consistency and accuracy of the external knowledge, in this paper, we select the first meaning of each entity in Wiktionary to be added to the dataset. Some of the words may not have appropriate counterparts in Wiktionary, so the lemma form of words is used to process, i.e., retrieve the words closest to their meanings to supplement the missing knowledge (see Figure 5).

This process not only enriches the training data, but also enhances the model's ability to understand lexical diversity and context. By incorporating external knowledge into the dataset, the data enhancement layer plays an important role in improving the model generalization ability and classification effectiveness. In addition, the effectiveness of this method has been validated in several text categorization tasks, proving its feasibility and superiority in practical applications.

3.3. Text Embedding Layer

In this paper, we adopt the GloVe (Global Vectors for Word Representation) model for word embedding when processing Text XA and External Knowledge Text XKA. The GloVe model generates high-quality word vector representations through the combination of global statistical information and local contextual information. Specifically, GloVe utilizes the statistical information in the word co-occurrence matrix to ensure that the distributional properties of word vectors in the global corpus are preserved, while capturing the meanings of words in specific contexts through local information in the context window. This embedding method can effectively transform text and external knowledge into low-dimensional, dense vector representations, providing richer semantic information for subsequent feature extraction layers [8].

3.4. Feature Extraction Layer

The feature extraction layer stands as a pivotal component within a knowledge-enhanced text classification model, tasked with distilling high-level feature representations from both the textual and external knowledge embedding vectors. This study employs three neural network architectures—Convolutional Neural Networks (CNNs), Long Short-term Memory networks (LSTMs), and Bidirectional Encoder Representations from Transformers (BERT)—for feature extraction. Each model boasts its distinct strengths, enabling the capture of textual features from various perspectives.

CNNs, with their multilayered structure, transform raw data into more sophisticated feature representations through a combination of input layers, convolutional layers, pooling layers, and fully connected layers. The convolutional layer employs a sliding window mechanism to perform convolution operations on the output of the text embedding layer, thereby extracting local features. The pooling layer then reduces the dimensionality of the features extracted by the convolutional layer, thereby mitigating computational complexity while preserving crucial information. The fully connected layer consolidates the features extracted by the convolutional and pooling layers, mapping them ultimately to classification outcomes.

LSTM units, as a form of recurrent neural network capable of addressing long-term dependency issues, regulate the flow of information through forget gates, input gates, and output gates. The forget gate determines which information from the previous hidden state should be discarded, the input gate decides which current input information should be retained in the hidden state, and the output gate controls the output of the current hidden state. This gating mechanism allows LSTMs to effectively handle lengthy text and intricate contextual relationships, circumventing the pitfalls of gradient vanishing and exploding. In text classification tasks, LSTM models can capture temporal

information and long-term dependencies within the text, thereby providing richer semantic support for classification results.

The BERT model, a pre-trained language model built upon the Transformer architecture, derives its prowess from a deeply bidirectional encoder framework that extracts a rich tapestry of contextual information from extensive corpora of textual data. The input text is dissected into discrete tokens, augmented with specific markers—[CLS] to denote the commencement of a sentence and [SEP] to signify its termination or to demarcate between sentences. These tokens are then transmuted into vector representations, imbued with positional embeddings to encapsulate the sequential characteristics inherent in the sequence. Leveraging multiple layers of Transformer encoders, BERT calculates the intricate interactions between each token, thereby capturing both global and local contextual nuances at various levels. Ultimately, the vector corresponding to the [CLS] marker serves as the encapsulated feature representation of the entire text, which is harnessed as input for downstream tasks [9].

3.5. Knowledge Fusion Layer

The knowledge fusion layer plays a crucial role in the text classification model based on knowledge enhancement and is responsible for effectively fusing text features and external knowledge features. In this paper, the gating mechanism in neural networks is used to control the fusion ratio of text features X and external knowledge features XK, so as to generate the final text feature representation. The gating mechanism enables the model to flexibly select more textual information or external knowledge information according to the needs of specific tasks by dynamically adjusting the weights. This fusion approach not only improves the richness and accuracy of the feature representation, but also enhances the generalization ability of the model. Specifically, the gating mechanism ensures that the model can better utilize the supplementation of external knowledge when dealing with complex text by weighting the combination of different features, while maintaining high efficiency and accuracy when dealing with regular text.

3.6. Output Layer

In this paper, a multilayer perceptron (MLP) is used to process the fused text feature representations by mapping the feature vectors to different category spaces through multilayer nonlinear transformations. The multilayer perceptron can effectively capture the complex relationship between features and improve the accuracy of classification. The expression of the output layer is shown in Equation (9):

$$y = soft\max(\tanh(ZW + b)) \tag{9}$$

where W and b are both parameters of training and y is the input classification result.

4. Experiments and Results and Analysis

4.1. Experimental data

This article conducts experiments on text classification using the SST-5, SST-2, and IMDB datasets. Among them, the SST dataset (Stanford Sentiment Treebank) is an affective analysis dataset released by Stanford University, comprising sentences covering a myriad of themes, such as movie reviews, music critiques, and product evaluations. The SST-5 dataset assigns each sentence to one of five emotional categories: extremely negative, somewhat negative, neutral, somewhat positive, and extremely positive. The SST-2 dataset is derived from the SST-5 dataset through a

process of refinement. Additionally, the IMDB movie review dataset is a binary sentiment analysis dataset, consisting of 50,000 reviews from the Internet Movie Database, which predominantly express viewers' perceptions and emotional responses to films. Each film is accompanied by no more than 30 reviews, which are categorized into two primary sentiments: positive and negative [10].

4.2. Experimental Results and Analysis of Knowledge Enhancement Algorithms for Classification on Different Models

In order to verify the generality of the knowledge enhancement algorithm proposed in this paper, the RNN model is selected for the experiments for detailed evaluation. The experimental results are shown in Figure. 6, demonstrating the improved classification performance of the knowledge enhancement algorithm on multiple datasets. On the SST-5 dataset, the text classification accuracy of the RNN (KB) model is improved by 8.02 percentage points compared with the traditional RNN model; on the SST-2 dataset, the accuracy of the RNN (KB) model is improved by 2.45 percentage points; on the IMDB dataset, the accuracy of the RNN (KB) model is increased by 2.98 percentage points. These results indicate that the introduction of external knowledge significantly improves the classification performance of the model and enhances the accuracy and generalization ability of the model.

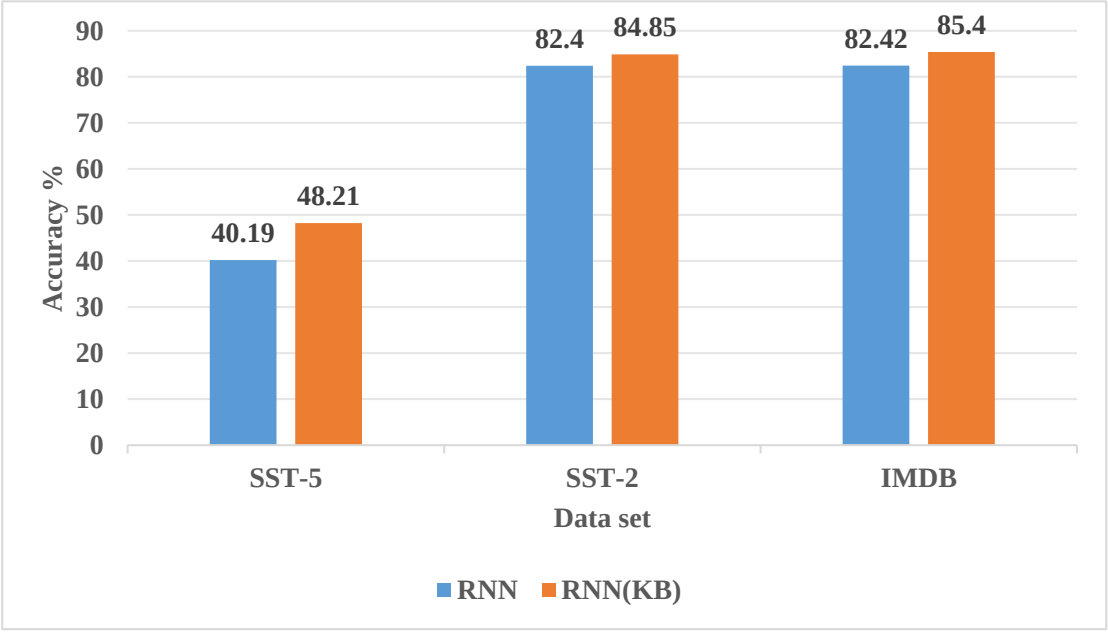

Figure 6. Classification results of knowledge enhancement algorithm on different models

In further experiments, this paper also tests the performance of the knowledge enhancement algorithm on other models, including LSTM and Transformer models. On the SST-5 dataset, the classification accuracy of the LSTM(KB) model improved by 7.58 percentage points over the LSTM model, and the accuracy of the Transformer(KB) model improved by 6.31 percentage points. On the SST-2 dataset, the classification accuracy of the LSTM (KB) model is improved by 2.87 percentage points over the LSTM model, and the accuracy of the Transformer (KB) model is improved by 2.64 percentage points. On the IMDB dataset, the classification accuracy of the LSTM (KB) model improves by 3.15 percentage points and the accuracy of the Transformer (KB) model improves by 2.76 percentage points over the LSTM model.

4.3. Experimental Results and Analysis of Knowledge Enhancement Algorithms for Classification in Different Sized Datasets

In this paper, the BERT model is chosen to verify the classification effect of the proposed knowledge enhancement algorithm on different scale datasets. The experimental results are shown in Figure. 7, demonstrating the improvement of text classification accuracy of the knowledge enhancement algorithm under different training data ratios. Specifically, on the SST-5 dataset, the classification accuracy of the BERT(KB) model is 2.15 percentage points higher than that of the BERT model when using 10% of the training data; 1.70 percentage points higher when using 30% of the training data; 1.69 percentage points higher when using 50% of the training data; 1.50 percentage points higher when using 80% of the training data; and 1.50 percentage points higher when using 80% of the training data. 1.50 percentage points with 80% training data, and 1.50 percentage points with 100% training data. These results show that the introduction of external knowledge can effectively improve the text classification performance of the BERT model regardless of the size of the training data. It is especially worth noting that the knowledge enhancement algorithm is more effective when the training data sizes are small (e.g., 10% and 30%), which indicates that the external knowledge gains the model's performance more obviously in the case of data scarcity.

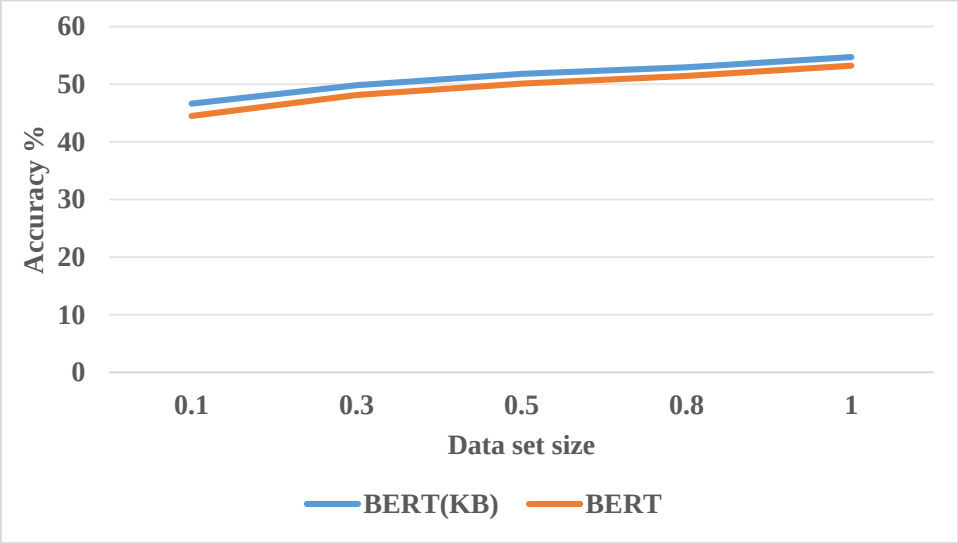

Figure 7. Classification results of the knowledge enhancement algorithm on different sizes of SST-5 dataset

5. Conclusion

In summary, automatic text categorization methods based on knowledge enhancement show great potential in improving classification accuracy and robustness. By introducing an external knowledge base, the model not only understands the text content better, but also shows stronger capability in dealing with long text and complex tasks. The structure of the model framework proposed in this paper, including data enhancement layer, text embedding layer, feature extraction layer, knowledge fusion layer and output layer, provides a systematic reference for researchers. The experimental results show that the knowledge enhancement strategy has significant performance improvement on different models and datasets of different sizes, which further validates its effectiveness in the field of text categorization. Future research directions will focus on more efficient knowledge fusion methods and larger-scale knowledge base applications to further

enhance the comprehensive performance of the model. It is hoped that the research results in this paper can contribute to the development of text categorization technology and stimulate more innovative ideas.

References

[1] Kannampallil T G, Farrell R G. Automatic Learning Object Categorization For Instruction Using An Enhanced Linear Text Classifier [M]// Knowledge Management: Nurturing Culture, Innovation, and Technology. 2005: 299-304.

[2] Hou Y, Wang Y, et al. Chinese text classification based on attention mechanism and feature-enhanced fusion neural network [J]. Computing, 2020, 102: 683-700.

[3] Shehata S, Karray F, Kamel M. An efficient concept-based mining model for enhancing text clustering [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1360-1371.

[4] Yu W, Zhu C, Li Z, et al. A survey of knowledge-enhanced text generation [J]. ACM Computing Surveys, 2022, 54(11s): 36-38.

[5] Sebastiani F. Machine learning in automated text categorization [J]. ACM computing surveys (CSUR), 2002, 34(1): 44-47.

[6] Hernandez Barrera A O, Montero Valverde J A, Hernández Hernández J L, et al. Automated creation of a repository for learning words in the area of computer science by keyword extraction methods and text classification [C]//International Conference on Technologies and Innovation. Cham: Springer Nature Switzerland, 2023: 186-203.

[7] Hawalah A. Semantic ontology-based approach to enhance Arabic text classification [J]. Big Data and Cognitive Computing, 2019, 3(4): 53.

[8] Glazkova A. Using automatic text categorization technologies in the modern educational process [J]. Faculty of Educational Sciences, 2016: 23-26.

[9] Kadhim A I. Survey on supervised machine learning techniques for automatic text classification [J]. Artificial intelligence review, 2019, 52(1): 273-292.

[10] Apté C, Damerau F, Weiss S M. Automated learning of decision rules for text categorization [J]. ACM Transactions on Information Systems (TOIS), 1994, 12(3): 233-251.