

Received 7 September 2023, accepted 21 September 2023, date of publication 26 September 2023,
date of current version 3 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3319384

RESEARCH ARTICLE

A Multilabel Learning-Based Automatic Annotation Method for Semantic Roles in English Text

LI LEI¹ AND HAO WANG^{1,2}

¹School of General Education, Hunan University of Information Technology, Changsha 410151, China

²School of Foreign Language, Changsha Normal University, Changsha 410100, China

Corresponding author: Hao Wang (wanghao@csnu.edu.cn)

This research work was supported by Scientific Research Fund for Outstanding Youth of year 2022 granted by Education Department of Hunan Province, China (Grant no. 22B1027).

ABSTRACT With the increasing amount of textual information in the Internet, smart semantic comprehension is a practical demand. Among, automatic annotation for semantic roles remains the fundamental part for effective semantic comprehension. Although machine learning-based methods had received much attention in recent years, they mostly divided each sentences into separable parts for calculation. To deal with such challenge, this paper introduces multilabel learning to propose a novel automatic annotation method for semantic roles in English text. In the semantic representation of words, the method uses convolutional neural networks to extract local feature information of words from the character level. Such design can alleviate the problem of inconspicuous semantic features caused by random initialization of unregistered words. Secondly, in the process of implication recognition, by combining the interactive attention mechanism to construct a capsule for each implication relation separately, the recognition of the final implication relation is completed in the way of categorical learning. At last, some experiments are conducted on real-world data to verify the proposed method with being compared with several typical relevant methods. The obtained results show that the proposal achieves better Macro-F1 results on eight datasets compared to seven algorithms. Besides, the proposal also performs better than others in the sensitivity testing, as its performance can remain stable with the increase of noise input. In summary, the proposal can achieve good results and show strong capability in semantic role labeling tasks.


INDEX TERMS Multi-label learning, semantic comprehension, automatic annotation, deep neural networks.

I. INTRODUCTION

In the era of big data, with the continuous improvement of science, the data collected has grown exponentially in terms of dimensionality and quantity [1]. The development of the Internet, the Internet of Things, mobile networks, and various social networks are surfacing, and the scale of data is exploding while the complexity and multiplicity of data are also increasing dramatically [2]. In many fields such as computer vision, bioinformatics, natural language processing, and information security, how to effectively mine valuable information from high-dimensional multi-sense data to help decision-makers make scientific decisions and achieve accurate management has become an important

problem to be solved [3]. Textual implication recognition, also known as natural language inference, is a fundamental yet challenging task in the field of natural language processing [4]. The goal of this task is to determine the directed semantic relationship between two consecutive texts, where the embodied antecedent is noted as the text T and the embodied consequent is noted as hypothesis H [5].

A text T is said to imply hypothesis H if the semantics of hypothesis H can be inferred from the semantics of text T when interpreted in the context of text T placed in the context of text T , notated as $T \Rightarrow H$. The implication relation between texts is directed and can be classified as forward inference, reverse inference, bidirectional inference, contradiction, and neutrality according to the implication relation between two texts [6]. The growing set of candidate tags for multi-label data poses a challenge to accurately label

The associate editor coordinating the review of this manuscript and approving it for publication was Alicia Fornés .

multi-label data. For example, the website MULAN, which provides multi-label datasets, includes 26 datasets, half of which have a candidate tag set size higher than 100, up to 3993. Therefore, verifying whether the candidate tags are relevant to the current sample one by one is time-consuming and laborious.

Textual implication relations, as a kind of directed semantic relations, are widely distributed in natural language texts [7]. When the semantics of texts and the implication relations between texts are obtained, these texts can be linked to form an interrelated semantic network, which enables computers to further understand and apply the semantic information of texts [8]. Text implication recognition aims to promote the semantic study of text and improve the computer's understanding of natural language [9]. And it can assist other natural language processing tasks with rich application scenarios [10]. From a tagging perspective, the semantics of things are increasingly diverse and granular [11].

English text proofreading tasks include many components, such as grammatical error correction, spelling correction, sentence simplification, fact-checking, sentence compression, and sentence paraphrasing. Among the many tasks of English text proofreading, grammatical error proofreading and fact-checking are two of the most important tasks to ensure the grammatical and semantic correctness of texts, and therefore have received much attention from academics. The grammatical error proofreading task aims at making grammatical corrections to ensure that the text conforms to the grammatical rules and that the text is fluent. For factual verification, it is more about verifying the semantic information of the text and determining whether it satisfies the established factual knowledge.

The data consists of a small number of fully labeled samples and a large number of unlabeled samples, also known as "small sample" data. For this scenario, the semi-supervised learning model can effectively learn and utilize unlabeled samples [12]. Semi-supervised learning means that the learner can independently and automatically utilize unlabeled data to improve learning performance without relying on external interactions. Therefore, combining semi-supervised learning with multi-label feature selection has become a hot research topic in recent years. The label space of the data is incomplete, i.e., only partial labels are given for each sample. When there are few missing labels, supervised multi-label feature selection methods can achieve some recognition performance of selected features by ignoring the processing of missing labels; when label information is severely missing, supervised multi-label feature selection methods will fail [13].

Therefore, to select the optimal feature subset in weak label learning scenarios, the combination strategy of label missing learning and multi-label feature selection becomes open research. In addition to grammatical relations in English, the lexical properties of words also contain information, which defines the usage and function of words. Machine learning models can extract information from many aspects, but if

a word has been labeled with lexically, it can not only disambiguate and strengthen word-based features, making it more accurate as features for models, but also effectively remove deactivated words. Therefore, lexical annotation is also widely used in tasks such as text classification, machine translation, and text summarization.

II. RELATED WORK

As the field of machine learning has expanded with other disciplines, a large number of new ideas and theories have emerged to support the further development of classification tasks and multi-label learning has been widely discussed and studied as a result. In particular, with the growing interest in deep learning algorithms, this learning paradigm also brings a lot of opportunities and challenges for classification tasks. Some progress has also been made in classification tasks based on the collision of traditional ideas with deep learning networks, and this progress facilitates the extraction of feature representations from large amounts of data to fit predicted dataset distributions [14]. Both the problem transformation method and algorithmic adaptive method are the core ideas of multi-label learning.

While multi-label learning is based on the idea of problem transformation, this method generally converts the multi-label problem into multiple single-label problems, which are then solved by traditional classification algorithms. This method is mainly applicable to the simple computational principle, faster operation speed, and is more suitable for online learning, but at the same time, with the increase of the number of labels, the time complexity is large, and the problem of slow convergence speed does exist. In contrast, the algorithmic adaptive method is an extension of the traditional multi-class algorithm to a multi-label algorithm without the need for problem transformation, but the computational complexity may be higher than that of the problem transformation method.

He et al. [15] introduces the multi-label learning algorithm for sample association relations (ML-K-Nearest Neighbor, MLKNN), i.e., the k-nearest neighbor method is used to measure the degree of similarity of samples to reason about their multi-label sets. Ameer et al. [16] analyzed two types of label dependencies using contextual information in a multi-label dataset. The literature [14] investigated a multi-label conditional random field model, which is implemented by directly parameterizing the labels. Decision trees are also a class of methods based on algorithmic adaptive solutions for multi-label learning, and Zhang et al. [17] modifies the entropy formula in decision trees to allow multiple labels to be included in the leaf nodes. In addition, methods such as deep learning can also be used to handle multi-label learning tasks, and Xiong et al. [18] gives an efficient means of feature extraction for multi-label learning algorithms based on deep self-encoder and label projection methods (Canonical Correlated Autoencoder, C2AE).

The literature [19] analyzes the label dependency and partial multi-label dependency problems based on extracting

sample relations from input features based on positive and negative labels respectively and obtaining label information from the output space, which provides a broad idea for the introduction of multi-label association relations. There are also many approaches to introduce sample relations and label correlations in multi-label learning. The literature [20] proposed Glove (Global Vectors for Word Representation), which combines the advantages of two mainstream models, global matrix decomposition and local context window, and the model only trains the non-zero elements in the word co-occurrence matrix instead of the whole sparse matrix or the context window in a large corpus, making full use of the global statistical information of the corpus while improving the training speed of word vectors on large corpora.

With the increasing application of word representation learning in the field of natural language processing, representation learning methods for sentences have also attracted the attention of many researchers, and Yu et al. [21] proposes skip-thought vectors, which adopt the idea of word2vec to encode sentences and obtain the semantic representation of sentences, and experiments have proved that it has achieved good results in many tasks. Experimentally, it has been shown to achieve good results in many tasks. In the literature [22], each word is represented as a dense low-dimensional real vector by training the language model, and these vectors form a word vector space, and each vector can be regarded as a point in this space, on which the similarity between words can be calculated using distance or angle, which effectively preserves the semantic relevance between words. The existing grammar correction models tend to consider the grammar correction task as a low-resource translation task, and different methods have been proposed for corpus augmentation to improve the effectiveness of the grammar correction model.

The literature [23] further integrates various grammar correction data augmentation strategies such as random substitution and reverse translation to improve the pre-training effect of the grammar correction model. The literature [24] uses BERT to encode input sentences and fuses the BERT-encoded word vector representation into the grammar reformation model to enhance the grammar reformation model. However, some of the latest pre-trained language models for natural language generation tasks, e.g., GPT2 as well as T5, whose model effects have not been evaluated on grammar correction tasks.

III. METHODOLOGY

A. MULTI-LABEL LEARNING ALGORITHM BASED ON CLASS ATTRIBUTES

Natural language processing is one of the core research areas of artificial intelligence, which plays an important role in processing natural language using computer technology and has given rise to numerous applications, such as machine translation, information retrieval, automatic question and answer, automatic text proofreading, and so on. With the development of deep neural networks and pre-trained

language models, the ability of natural language processing techniques to understand text and model language models has been further improved. Automatic text proofreading methods aim to help people implement automatic text proofreading systems, which consist of two parts, grammatical error proofreading and fact verification, to ensure the correctness and authenticity of the text.

The multi-label learning algorithm based on class attributes is the first method to construct unique attributes for class labels. Instead of using the same feature space when constructing classification models for different labels, it uses clustering techniques to construct its attribute features for the labels. The algorithm is divided into two main parts: constructing class attributes and training the classification model. First, the training samples are divided into two sets, and cluster analysis is performed on these two sets to generate cluster centers, and the cluster centers are used to construct class attributes; then, the generated class attributes and the binary learning algorithm are used to construct a classification model for each label. For the test samples, the label relevance is predicted using the classifier, and the relevant labels are combined to obtain the relevant label set of the samples.

First, the training set can be divided into two sets, namely the positive sample set and the negative sample set, based on the correlation between the samples and the labels. For the label $l_n \in \phi$, its positive sample set P_n and negative sample set M_n can be expressed as follows:

$$P_n = \{x_i \mid (x_i, Y_i) \in C, l_n \in Y_i\} \quad (1)$$

$$M_n = \{x_i \mid (x_i, Y_i) \in C, l_n \notin Y_i\} \quad (2)$$

LIFT uses clustering techniques that can explore the underlying properties of the data to generate discriminative attributes that can capture the characteristics specific to each label. Specifically, clustering analysis is performed on the two generated sets separately, and LIFT employs a simple and efficient kmeans algorithm that uses clustering techniques to partition the set M into m^+ clusters and the set P into n^- clusters. To alleviate the category imbalance problem, LIFT sets the number of clusters in both sets to be equal, i.e., $m^+ = n^- = m_k$, which also allows the information obtained from both sets to be treated equally. The number of clusters in the sets is set as follows:

$$m_k = p \cdot \min(|P_k|, |M_k|) \quad (3)$$

Instead of using the same feature space for all labels, LIFT uses clustering techniques to construct feature spaces for different labels that match their semantic information, which is different from many algorithms. In addition, LIFT ignores the correlation between labels and is an algorithm that uses a first-order strategy. Two shortcomings of the algorithm can be found. First, because the randomness of clustering may lead to unstable clustering results, the class attributes of class labels may not reflect the structural information of the feature space well if the clustering process is executed only once and constructed based on the generated results; second, LIFT

adopts a first-order strategy, which means it does not consider the correlation between labels [25]. The multilabel learning algorithm based on the integration of clusters for class attributes is improved to address the above two shortcomings. In the process of constructing class attributes, LIFTACE adopts the cluster integration technique, i.e., it performs another clustering on top of the original clustering process and merges the two processes, which not only makes the clustering results more stable but also takes into account the correlation between labels, thus compensating for the shortcomings of LIFT.

Based on the correlation between samples and class labels, LIFTACE divides the training set into a positive sample set and a negative sample set. It is assumed in LIFTACE that if two labels are similar, then their corresponding clustering results should also be similar. That is, the instance similarity matrices corresponding to similar labels should be similar to each other. Therefore, the clustering results of a specific label can be updated by combining the clustering results of all labels, and the label correlation is considered in the process of combination. According to the above assumptions, for the label $l_k \in \mathcal{Y}$, its instance similarity matrix can be updated by combining the instance similarity matrices of other labels. The updated instance similarity matrix W_1 is as follows:

$$W_l = \varphi \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \left[\frac{p(x, y)}{\ln p(x, y)} - Ax - Cy \right] \quad (4)$$

The iterative label propagation process needs to reduce the confidence of irrelevant labels in the candidate label set on the one hand and increase the confidence of relevant labels in the non-candidate labels on the other hand. Therefore, C2LP-IML adopts an untruncated label propagation approach, which considers both the positive influence of the relevant labels of the nearest neighboring samples on the label confidence and the negative influence of the irrelevant labels of the nearest neighboring samples on the label confidence. The tenth iteration label propagation process is expressed as:

$$F(t) = \left(\frac{m + \gamma}{n} \right) \cdot \sum (x^m - t^{m-1}) f(t) \quad (5)$$

where $\gamma \in (0, 1)$ is a coordination parameter in the propagation process to control the weight of the dependence of the value of the confidence matrix W_{\downarrow} on the results of the previous iteration. At the end of the iterative propagation, to avoid the scale imbalance problem, the confidence matrix W_1 is normalized to obtain the final confidence matrix W_1^* , and the normalization equation is expressed as:

$$W_l^* = \iint g(t) dt = \left(\frac{1 + \gamma}{n} \right) \cdot \sum (x - 1) f(t) \quad (6)$$

After obtaining the final label confidence matrix W_1^* , to avoid the phenomenon of overfitting in the training set due to unbalanced data division, the similarity between the test sample and its k nearest neighbors in the training sample space is considered as the confidence weight from

the perspective of matrix complementation, and the weighted label confidence matrix W_{final} is obtained as follows:

$$W_{final} = P g W_1^* \quad (7)$$

$$p = \begin{cases} 1 - \frac{t_j * \sum_{i=1}^n x_{ij}}{\sum_{j=1}^m T_j}, x_{ij} \in T \\ 0 \end{cases} \quad (8)$$

After obtaining the weighted label confidence matrix W_{final} , the label confidence levels in the original candidate label set and the non-candidate label set are determined separately. First, by setting the confidence threshold of the candidate label set, when the value of the label confidence exceeds the threshold, it will be identified as a trusted label. The selection of trusted labels in the candidate label set can be expressed as:

$$w = \frac{kx + \delta}{g^2(x)} + C \quad (9)$$

The final replacement set of trusted tags consists of two parts, the one hand from the tags exceeding the threshold in the original candidate tag set, and the other hand from the tags exceeding the threshold in the original non-candidate tag set. Thus, the purpose of filtering noisy tags in the candidate tag set while recovering reliable tags in the non-candidate tag set is achieved [26]. The previous grammar error quality assessment models as well as grammar error checking models ignore the evidence of high-quality grammar corrections among multiple grammar correction results provided by grammar error checking models, thus limiting the effectiveness of grammar error checking models as well as grammar error quality assessment models. Therefore, we hope to make full use of the multiple correction results provided by the grammar error checking model to suggest possible grammatical errors and correction results to further improve the effectiveness of grammar error checking and grammar correction quality assessment, and to improve the effectiveness of grammar error checking by reordering the column search results.

To improve the coverage of the pre-trained word vector table in the vectorization process of the short textbook, the improved similarity is used to find possible spelling errors in the short textbook and thus more accurately match the corresponding words in the corpus. Immediately afterward, to address the problem of limited semantic information that can be provided by the short text, an external knowledge base is introduced to conceptualize the short text and its related words, which extends the semantics of the short text. Convolutional neural networks also have obvious drawbacks in the feature extraction process. The mechanism that requires setting the convolutional window size makes it impossible to mine the information of long-distance text data and ignores the dependency between contexts in long-sequence text data. Recurrent neural networks add self-connection and interconnection operations in the hidden layer to better preserve and remember textual information.

However, the RNN model still suffers from important information loss and gradient disappearance as the

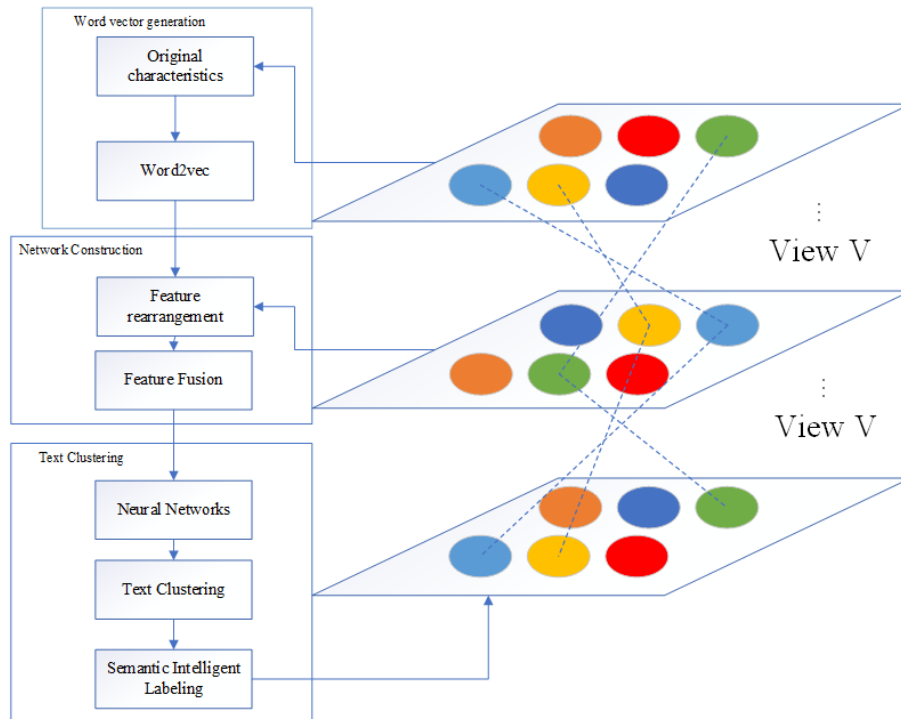


FIGURE 1. Flow chart of text feature extraction algorithm based on multi-label learning.

information weight decreases in information transmission. Starting from improving the feature extraction model, an LSTM combined with the CNN network algorithm based on the ECA attention mechanism is proposed for the feature extraction of text data. The algorithm uses LSTM networks to capture long-range features while alleviating the gradient disappearance problem in RNN networks; in addition, the ECA attention mechanism is introduced to emphasize the regions of interest in the utterance and suppress irrelevant background regions by dynamic weighting; then the second feature extraction is performed by convolutional neural networks to obtain the required text feature vectors, and finally, the K-means clustering with RWMD distance similarity. The feasibility of the proposed algorithm is verified by combining the K-means clustering algorithm with the RWMD distance similarity function. The flow framework of the LSTM-CNN text feature extraction algorithm based on the attention mechanism is shown in Figure 1.

The attention mechanism is a core technique commonly applied to image detection, natural language processing, and other fields proposed by research scholars during the development of deep learning based on human attention characteristics. The core idea draws on the fact that when capturing relevant and important information in the visual field, humans tend to pay more attention to those regions that match the features to focus on thought processing, while selectively ignoring those regions that do not match the feature expression. The introduction of the attention mechanism in deep learning enables giving higher weights to the significant influential features in the process of feature extraction to obtain more information and set lower weights

to discard the irrelevant information to avoid being influenced by them. Finally, the purpose of rational allocation of limited resources is achieved. The attention mechanism can be divided into two modules: the hard attention mechanism and the soft attention mechanism [27].

The hard-attention mechanism restricts and selects the regions of interest as input to the model by filtering them according to the extended attention present in each point of the image or text. The hard attention mechanism is a stochastic prediction process that emphasizes the importance of dynamic changes. Although it can achieve good results in machine learning algorithms, its non-differentiable nature makes it difficult to implement and less widespread in the training process. In contrast, the soft attention mechanism is deterministic attention, which utilizes the weights obtained from neural network training to combine with the input features in the channel or space to generate the corresponding weighted input features, ultimately achieving the purpose of focusing on the channel and space regions. Most specifically, soft attention is everywhere microscopic, which allows the neural network to obtain the weights of attention based on the gradient algorithm for bi-directional propagation. Therefore, soft attention is increasingly practical and concise in the learning and implementation process.

B. AUTOMATIC ANNOTATION METHOD FOR ENGLISH TEXT ROLES BASED ON ATTENTION MECHANISM

In the SE-NET attention mechanism, two main types of channel attention modules, SE-Var2 and SE-Var3, are used, both of which learn the weights of each channel

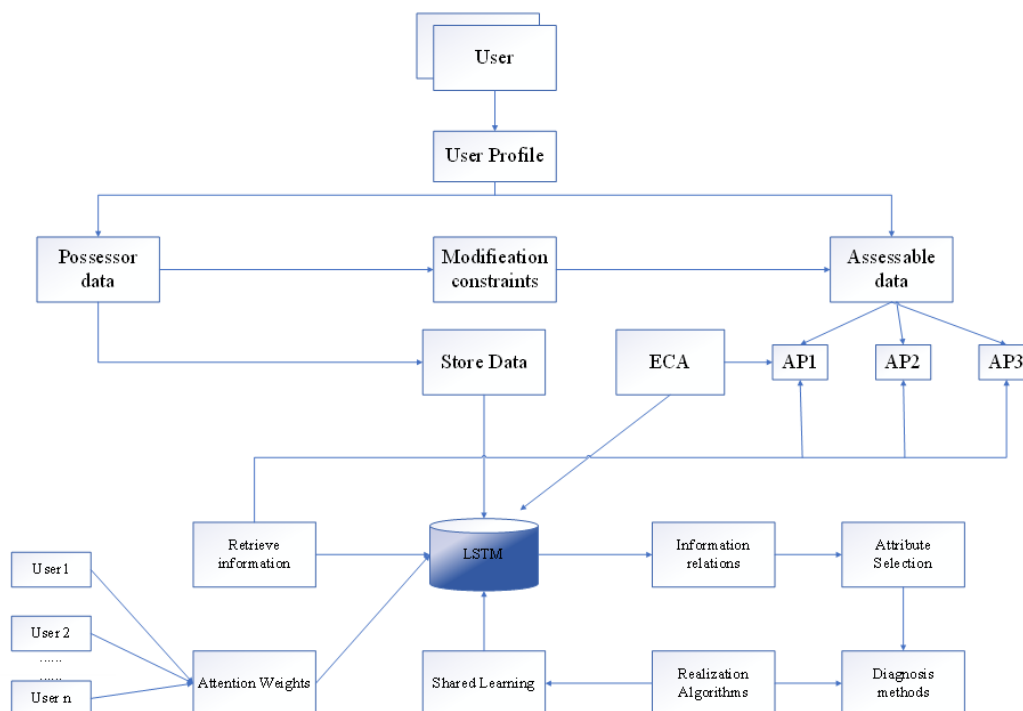


FIGURE 2. Schematic diagram of the model structure of the ECA channel attention mechanism.

independently through a fully connected network to establish a direct mapping between channels and their weights. Moreover, since the weights of SE-Var2 are a symmetric matrix and the weights of SE-Var3 are a full matrix, they can be considered as deeply separable convolution and fully connected layers with grouped convolution, respectively, allowing the attention mechanism to capture the interaction between channels by dividing the feature map into multiple groups while performing independent linear transformations in each group. However, SE-NET has obvious drawbacks, as the fully-connected dimensionality reduction approach not only leads to inefficient capturing of dependencies between channels but also causes the problem of losing dependencies within different groups. In contrast, a more advanced ECA attentional decimal combined with the LSTM network is proposed in this section.

This method uses a one-dimensional sparse convolution operation to optimize the fully connected operations involved in the SE module and thus compares the hidden node states with the input feature vector corresponding to the hidden node states to obtain the attention assignment probability distribution values, i.e., attention weights. After adding the attention weights to the output of the LSTM network, the extracted contextual information relations are weighted one by one. Such an operation allows the information at any position in the text to obtain different degrees of attention while effectively avoiding the negative effects of dimensionality reduction and the independence problem between groups on the neural network. In addition to maintaining the model performance, the cross-channel interaction mechanism

reduces the complexity of the model using the feature of shared learning parameters and improves the risk of feature loss caused during the training process. The model structure of the ECA channel attention mechanism is schematically shown in Figure 2.

After the text feature vector is globally averaged and pooled by the GAP layer without dimensionality reduction, a one-dimensional sparse convolution of size k in the ECA attention mechanism is used instead of the fully-connected layer in the traditional attention module to learn the interactions across channels. where the size of k indicates the size of the convolution kernel, i.e., the range of action covered by the channel interactions, i.e., how many nearest neighboring channels are involved in the attention prediction of a channel [28]. Moreover, the choice of k value still varies for different channel dimensions C and different neural network structures. After determining the added attention mechanism, the text feature vectors from the LSTM network-based feature extraction model are fused and spliced with the ECA attention mechanism, and the representative feature vectors are further mined from the many text feature vectors. The original text data is converted into continuous real word vectors by the Word2vec model after the steps of word separation, deactivation, and stem, and the input word vector matrix is formed by combining.

After feature extraction by the LSTM network model based on the ECA attention mechanism, although the dependencies existing between the distance context information are effectively captured, the single neural network model is still inadequate in solving the text feature extraction problem.

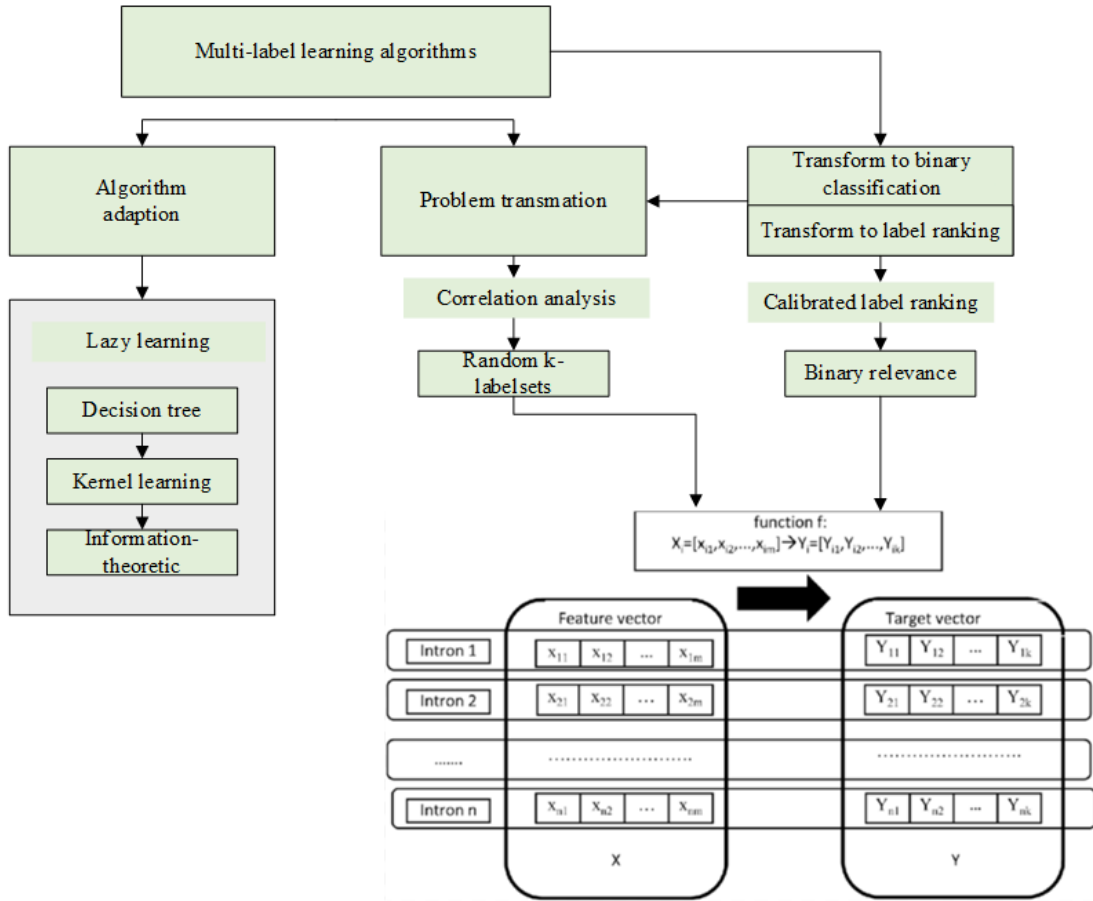


FIGURE 3. Inaccurately labeled multi-label learning problem and other related limited supervised information.

Therefore, this section proposes to use the feature vectors output from the above model as the input of the convolutional neural network for secondary feature extraction [29]. This process makes up for the neglect of the attention to the importance of local features of text during the training process of LSTM networks by mining the semantic relationships between adjacent words, and finally, outputs feature vectors that combine both contextual feature information and local features and are relevant to the current text topic. The traditional distance function calculates the similarity of word vectors, which sometimes only correlates with words but ignores the correlation with sentences or even the whole text, thus leading to poor clustering results.

To address this problem, RWMD distance similarity is used to improve the accuracy and efficiency of semantic similarity calculation between texts, which is based on the WMD distance similarity algorithm by restricting the correlation conditions and reducing the complexity of the algorithm. The function measures similarity by converting the similarity problem between data variables into a cost-minimization problem between everyday transportation items. When applied to text research, the minimum sum of the transformations of all feature vectors in a given text into the corresponding feature vectors of another text is

used as the evaluation criterion [30]. When the sum of the transformations is smaller, the more similar the two texts are. RWMD distance mainly uses the Euclidean distance between word vectors and weights to measure the similarity between texts. The formula for calculating the Euclidean distance between word vectors is defined by the following equation:

$$D(x) = \sum_{j=1}^J \sum_{i=1}^I (b_{ij} - c_{ij}) x_{ij} \alpha_j \beta_j \quad (10)$$

Then, the RWMD distance is used to calculate the similarity between text d and d', if each feature word vector in text d can be converted into the corresponding feature word vector in text d' by adding weight coefficients through the Euclidean distance, the RWMD distance calculation formula is defined as:

$$RWMD(x) = \sum_{k=1}^K \sum_{i=1}^I \left(x_j^k(t) - \sum_{j=1}^J x_{ij}^k(t) \right) d_{ij} \quad (11)$$

In the calculation of the reward function, we use the grammar correction model to rank the column search results obtained from the baseline grammar correction model, and the first ranked grammar correction result is used to calculate the corresponding grammar correction evaluation index F_t ,

to obtain the score F_t for the grammar correction model corresponding to the current moment t , which is used to introduce the grammar correction evaluation index F_t into the training of the model, and according to $t_t - 1$ and t as the reward function at the current moment:

$$R_t = I_{i=1}^t F - I_{i=1}^t U_i + c \quad (12)$$

There are many ways to generate bugs in adversarial attacks, but since we require the generated adversarial sentences to be visually and semantically similar to the original human understanding, we want the changes to the original words to be as small as possible. Therefore, we consider two kinds of perturbations, namely character-level perturbations and word-level perturbations. For character-level perturbation, a key observation is that words are composed of alphabetic symbols, and deep learning-based text classification systems usually use dictionaries to represent a limited set of possible words. And the size of a typical word dictionary is much smaller than the possible combinations of characters of similar length. In the deep learning model, all unknown words are mapped uniformly to an “unknown” word embedding vector [29]. Our results show that this simple strategy can effectively force the text classification model to make incorrect decisions. For word-level perturbation, we expect to trick the classifier by performing a nearest-neighbor search in the word embedding space to find some approximate words to replace the significant words in the original text without changing their original meaning.

It is assumed that the number of samples for multi-label data is n , the total number of features is d , and the number of labels in the set of labels is q . The time complexity of mutual information and conditional mutual information is $O(n)$ since all samples need to be visited for probability estimation:

$$O(n) = \frac{\delta d}{\delta t} \left(\frac{n!}{r!(n-r)!} d^r + q \right) \quad (13)$$

The redundant information is not obtained from any candidate features for both labels, i.e., it does not provide new information for multi-label classification. Therefore, this part of redundant information should not be added to influence the judgment when measuring the amount of information provided by features for the labels. Therefore, when the labels are interdependent, it is inaccurate to measure feature relevance using the sum of candidate features and the mutual information of each label. When the data labels are both incomplete and contain noise, it can be seen as a scenario where both missing labels and biased labeling problems exist. For example, some social media ask users to select several contents of interest when they sign up for the first time. It is likely that new users will not select all the contents of interest among many categories, and may select some contents that they do not know much about but are not interested in.

Figure 3 depicts the multi-label learning problem with inaccurate labeling and other related limited supervised information problem examples. In general, argument recognition and labeling are treated as classification problems.

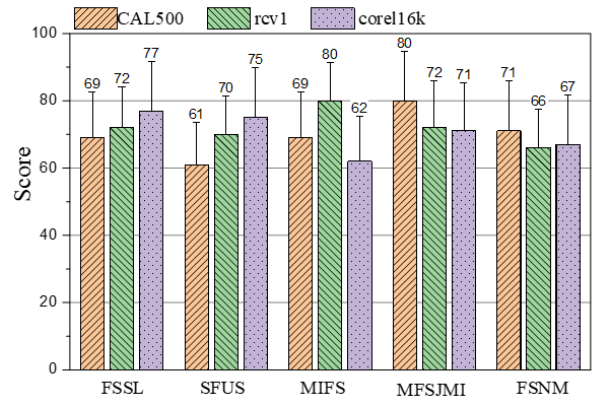


FIGURE 4. Performance comparison of sparsity learning methods.

The recognition stage is regarded as a binary classification problem, and the real argument is identified from the candidate terms after pruning. The labeling phase is regarded as a multi-valued classification problem, whose class set is all the semantic role labels. At first, people build a rule knowledge base based on the rule knowledge summarized by experts, and then use these rules for annotation. However, this approach not only requires a lot of expert knowledge, but also conflicts between rules. Later, people began to use statistical learning methods to establish effective conflict resolution mechanisms with statistical knowledge. At the same time, it is only necessary for people with certain professional knowledge to label according to the task objectives and construct corresponding statistical models for the training data.

Inaccurate labeling data is often easy to obtain and relatively inexpensive to acquire. However, few studies have addressed the challenges posed by inaccurate labeling. Existing multi-label learning methods for inaccurate labels usually require a small set of accurately labeled training samples, or supervised information from multiple perspectives. The additional supervised information in most inter-application scenarios requires experts or significant labor costs, which is difficult to apply to practical tasks. As mentioned in the previous section, the label confidence reflects the correlation between the label and the corresponding sample, so how to remove the noise while enriching the missing labels to restore a more accurate label confidence matrix is a core research problem in this task, and the probability P of the grammar correction quality assessment label y is:

$$P(y | \omega_p^k) = L(O) \cdot H(n) \quad (14)$$

We averaged the scores (i.e., probability $(y = 1 | \omega_p)$) of the number of English grammar proofs incorporating multiple grammar correction evidence to obtain the input sentence and the correction result sentence pair (s, c) of grammatical correction quality assessment scores we further evaluate the quality of all words in the grammatical correction results in

the k th node S_F as:

$$S_f = \frac{n! \sum_{i=1}^n H(i)}{(n-r)!} \quad (15)$$

We further train our VERNet model under word-level labels. Here we use both the word-level training labels for the input sentences and the correction results as supervised signals to guide the learning of the model in the sense of labeling the grammatical errors of the input sentences and the accuracy of the grammatical correction results, respectively [30]. The assumption of spatial sparsity affects the robustness of the algorithm. Semi-supervised learning has been widely used in many practical tasks due to the weakened dependence on the amount of data. However, even for labeling a small number of samples, it is difficult to obtain completely accurate true labels for data with large label space and complex data relationships. That is, among these obtained supervised information, there may still be several supervised information limitations mentioned above, which in turn make the multi-label learning task more complex and difficult. For example, when the amount of acquired data is too large, it is difficult to label all training samples even based on crowdsourcing techniques, and the labeled data may have biased labeling problems.

C. ILLUSTRATION DEMO

To make readers easier to understand operation points of the proposal, we give an illustration example for clarification. We use the Word2Vec toolkit to train the word vector and preprocess the input data. Among them, the word vector training is obtained by using all the original English texts as corpus and pre-processing, using the Skip-Gram model provided by Word2Vec. Experimental corpus is a text with marked information, which must be vectorized in order to enable the computer to process it. We regard each sentence as a sequence input by the network layer, and each word in the sentence as the input data at every moment. The corpus used in this paper is labeled data. Therefore, the back propagation algorithm is used in the training of the model, and the connection weights of the network layer are changed according to the difference between the output value of the model and the target label until the model converges to obtain the optimal solution of the model.

In the model training stage, the training text is preprocessed first, and the obtained vector is used as the input of the LSTM network layer. After calculation, the output value is sent to the Softmax layer for transformation, and after post-processing, the semantic role label of each word is obtained. The loss function value is then calculated from the original label. Finally, BP algorithm is used to update the connection weights of each layer until the model is trained. LSTM can make full use of the information of the whole text sequence, and can mine the information of the relationship between words and words, and apply the information to the processing of each word feature expression.

The training of the network starts after all the data is processed in the first sequence. First, the original label is compared with the output results of the network layer to calculate the value of the loss function, and then the gradient learning is carried out according to the decline direction of the loss function to update the connection weight between each layer and each gate. After the input data is processed by the LSTM network layer, the semantic role feature vector related to each word is obtained, and the input data at each moment can be used to the input information at all previous moments. Then, the obtained feature vector is sent to the Softmax layer for normalization processing.

IV. EXPERIMENTAL ANALYSIS AND CONCLUSION

A. EXPERIMENTAL DATA SET AND SETUP

To build the training, development, and test sets for training VERNet, we use ERRANT, an automated grammar error information annotation tool, to annotate the input sentences and the grammar correction result sentences generated by the grammar correction model with the grammar correction results given by the human annotator, and to annotate the areas that need to be modified. ERRANT, the automated grammar correction information annotation toolkit, performs various editing operations on the given sentence, such as deletion, insertion, and replacement, to further obtain the desired grammar correction result. Therefore, we label the input sentence and the grammar correction results generated by the grammar correction model with the manual annotation results using ERRANT to obtain sequential annotation labels, which indicate the grammatical correctness of the input sentence words and the accuracy of the grammar correction results provided by the grammar correction model, respectively. Each of these words is marked as correct (i.e., marker label is 1) or incorrect (i.e., marker label is 0).

In our experiments, we use the large-scale generic fact-validation dataset FEVER, the division of which remains the same as that of the FEVER shared task. The annotation of FEVER data is divided into two phases, the generation of the text to be validated and the fact labeling. In the first stage, the annotator rewrites randomly selected sentences from Wikipedia to form the text to be verified. The second phase is the annotation of the text to be verified, which requires the annotator to label each sentence as either supported, rejected, or insufficient information. The experiments use three publicly available multi-label datasets with a high number of labels or average labels CAL500, rcv1(subset1), and corel16k. The data sets rcv1(subset1) and corel16k will be abbreviated as rcv1 and corel16k in this section. To generate data that match the problem set, a certain percentage of true labels are first randomly removed for each training sample, while an equal percentage of irrelevant labels are added, where the percentage varies in the range of 10%, 20%, and 50%.

To fully review the performance of the C2LP-IML algorithm, further statistical tests were conducted to check its statistical significance at the Friedman level of 95%

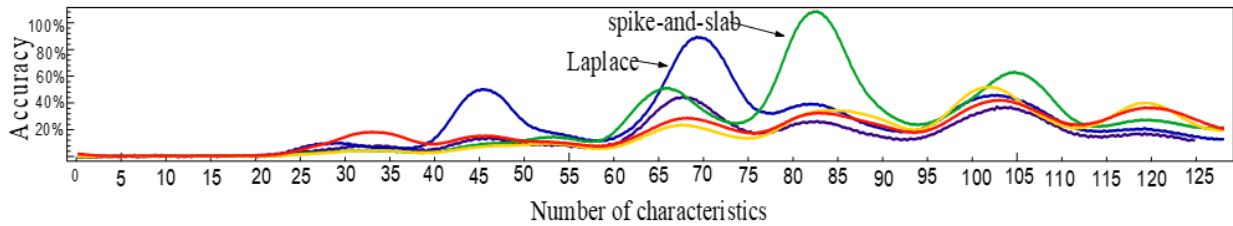


FIGURE 5. Effect of different parameters on the learning performance of the MFSJMI algorithm.

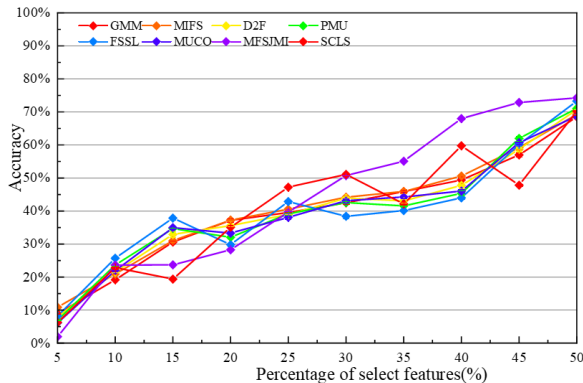


FIGURE 6. Classification results on the Macro-F1 indicator.

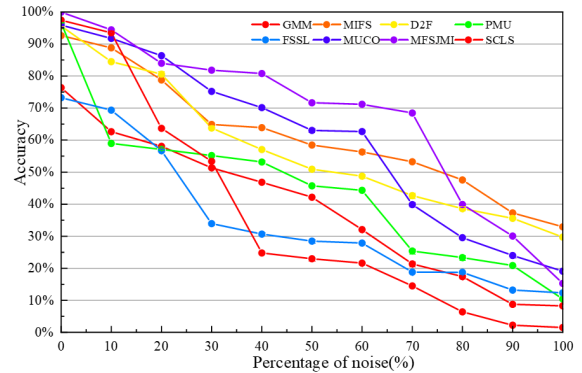


FIGURE 7. Test results of feature selection algorithm under different noise levels.

significance, as Figure 4 shows. Specific observations and analyses are as follows: the MFSJMI algorithm outperforms the comparison method in most learning scenarios, which indicates that label correlation based on reliable and accurate positive and negative label extraction is beneficial to improve the recognition ability and classification performance of the selected feature subset. As the size of missing labels grows, the selection performance of the MFSJMI algorithm gradually decreases as the information of reasonable labels decreases, further weakening the ability to guide feature selection. Sparse feature selection methods (i.e., FSNM algorithm and SFUS algorithm) perform relatively poorly, especially when the proportion of missing labels is relatively high, because these methods are less capable of handling missing labels that interfere with the feature selection process, and the limitation based on a single sparse regularization has a higher probability of missing features that discriminate sparse labels, which suggests that selective sparse mechanisms for feature selection tasks oriented to weakly labeled data are extremely important for feature selection tasks for weakly labeled data. At the same time, it also points out that it is very meaningful to properly identify and utilize missing labels to guide feature selection.

The impact of different sparsity learning mechanisms on feature selection is evaluated by comparing the MFSJMI algorithm using the spike-and-slab prior with the MFSJMI algorithm using the Laplace prior (i.e., l_1 regularization, with the regularization parameters determined using cross-validation). This section uses the Emotions dataset as a benchmark and sets the percentage of missing labels to 20%.

As the selection of features grows from 18 to 127 (the number of original features), Figure 5 shows the classification performance of selecting a subset of features using two different priors of the MFSJMI algorithm. The specific observations and analysis are as follows: in the learning scenarios of this section, the classification performance of the MFSJMI algorithm using the spike-and-slab before selecting a subset of features is generally better than that of the MFSJMI algorithm using the Laplace prior, because the selective decay mechanism helps to select informative features, especially those discriminated from sparsely labeled features. Also using a single l_1 regularization may lose some relevant features because it attenuates the weights equally for all features. The classification performance of the MFSJMI algorithm using the spike-and-slab before selecting features is improved when the size of the selected feature subset is extended from 12 to 36, due to more relevant features being selected, and the classification performance peaks when the selected feature subset is 36.

B. FEATURE RELEVANCE VALIDATION BASED ON JOINT MUTUAL INFORMATION AND INTERACTION WEIGHTS

Macro-F1 averages the precision and recall of all classes, and then calculates the F1 value as macro-F1. macro-F1 does not take into account the amount of data, so it treats each category equally. Because precision and recall of each class are between 0 and 1, they will be relatively affected by the high precision and high recall classes. Thus, we utilize the macro-F1 and accuracy as evaluation metrics.

The MIFS algorithm obtains the best Macro-F1 performance on the Science data set. In addition, the proposed algorithm MFSJMI obtains the best average Macro-F1 performance from the statistics in the ‘‘Average’’ row. In Fig. 6, the proposed algorithm MFSJMI achieves better Macro-F1 performance on 8 datasets compared to 7 multi-label feature selection algorithms, while MIFS and D2F algorithms obtain the best Macro-F1 results on computer and Social datasets, respectively. The results in the table show that the MFSJMI algorithm has the best average Macro-F1 performance, followed by the GMM, D2F, MIFS, MUCO, FSSL, SCLS, and PMU algorithms, respectively. Figure 6 shows the Macro-F1 performance results of the proposed algorithm MFSJMI and 7 comparative multi-label feature selection algorithms. As observed from the results in the figure, MFSJMI obtains better Macro-F1 performance on four multi-label datasets. In particular, MFSJMI significantly outperforms the other comparative algorithms when the number of selected features on these datasets is greater than 30% of the total number of features.

The experimental results of the Macro-F1 performance show that the Macro-F1 performance of the MFSJMI algorithm outperforms the other comparative algorithms as the number of selected features increases. The results indicate that the MFSJMI algorithm has better Hamming Loss performance on these data sets. Overall, the quality of features extracted by the MFSJMI algorithm is better than the other seven feature selection algorithms. From Figure 7, it can be seen that the overall trend decreases as the noise level increases, and the MFSJMI algorithm outperforms the other algorithms in terms of noise resistance. With the increase of noise level, MFSJMI outperforms than others. it can be seen from Figure 7 that MFSJMI is not as effective as others when the noise level is 80%, but its anti-noise ability is still better than other algorithms when the noise level is 0-70%. Meanwhile, it can be found that the slope of the curve of MFSJMI is smaller, which indicates that the overall anti-noise ability of MFSJMI is better than other algorithms.

However, in general, the anti-noise ability of multi-label learning based on the hybrid processing of the two algorithms has better performance than that of the single algorithm, but the computational complexity is high, and at the same time, because the algorithm itself has a high dependence on the label, so when the noise level has reached a certain level, about 30% of the noisy label, the trend of the algorithm’s prediction performance decreases obviously. In order to verify the stability of the MFSJMI algorithm and the comparison algorithm under the same index, radar plots were used, and the stability differences existing between each algorithm were reflected visually by the graphs. The experimental results of each algorithm under different evaluation indexes as well as data sets differ significantly from each other, and if stability analysis is done on the original values, some prominent values affect the overall analysis results. So in order to deal with the possible bias caused by the data, the original experimental results need to be normalized.

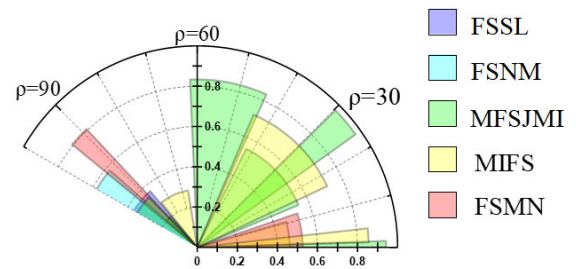


FIGURE 8. Sample analysis of factual verification results.

Figure 8 shows the sample analysis of factual verification results. When $\rho = 30$, the experimental results of the algorithm in this chapter are optimal, so the data at $\rho = 30$ are mapped onto the interval $[0,1]$, and the normalized values are used to represent the stability index values of the algorithm. The effectiveness of the corresponding model is further enhanced with the inclusion of the domain-oriented pre-trained language model, thus demonstrating the importance of the enhanced language model for the semantic understanding of the domain terms in the domain-oriented fact-checking task. Compared to the RP model, the mask-based language model continues to be trained in a way that benefits from its corpus size to more significantly improve the effectiveness of the language model for specialized domain textual reasoning, thus further enhancing the performance of the model for fact-checking tasks.

V. CONCLUSION

As the Internet continues to grow and the number of online texts increases, it is a very important task to be able to automate text proofreading. Automated text proofreading methods are designed to help people implement automated text proofreading systems, which consist of two parts, grammatical error proofreading and factual verification, to ensure the correctness and authenticity of the text. To identify grammatical and factual errors in text, we can integrate rich information such as linguistic knowledge, world knowledge, and domain knowledge to proofread text, and rely on the language modeling and reasoning capabilities of pre-trained language models to further realize an efficient automatic text proofreading tool. In this paper, we propose a fine-grained joint inference algorithm for the automatic annotation of semantic roles of English texts incorporating multi-label learning. Since the relevant factual evidence retrieval is done by information retrieval models during the fact verification process, additional noise is inevitably introduced. Moreover, only a fraction of the retrieved sentences is useful for verifying the semantic truth as well as the integrity of the current text. Therefore, this work hopes to further enhance the inference capability of the model at the fact-verification level by using multiple fact-verification evidence for fine-grained joint inference. The method outperforms other baseline models on the fact-verification generic dataset, proving its good inference ability and fact-verification effectiveness.

In this paper, the proposed method for automatic annotation of English semantic text involves a large number of word vector distance calculations when acquiring the related words of the text, and thus the time consumption needs to be further studied and improved. In the next study, the time complexity of the algorithm will be taken into account and the method of acquiring text-related words will be optimized in order to improve the classification efficiency of short texts.

REFERENCES

- [1] Q.-H. Kha, Q.-T. Ho, and N. Q. K. Le, "Identifying SNARE proteins using an alignment-free method based on multiscale convolutional neural network and PSSM profiles," *J. Chem. Inf. Model.*, vol. 62, no. 19, pp. 4820–4826, Oct. 2022, doi: [10.1021/acs.jcim.2c01034](https://doi.org/10.1021/acs.jcim.2c01034).
- [2] Z. Zhao, J. Gui, A. Yao, N. Q. K. Le, and M. C. H. Chua, "Improved prediction model of protein and peptide toxicity by integrating channel attention into a convolutional neural network and gated recurrent units," *ACS Omega*, vol. 7, no. 44, pp. 40569–40577, Nov. 2022, doi: [10.1021/acsomega.2c05881](https://doi.org/10.1021/acsomega.2c05881).
- [3] Z. Guo, K. Yu, N. Kumar, W. Wei, S. Mumtaz, and M. Guizani, "Deep-distributed-learning-based POI recommendation under mobile-edge networks," *IEEE Internet Things J.*, vol. 10, no. 1, pp. 303–317, Jan. 2023.
- [4] Z. Guo, K. Yu, A. Jolfaei, G. Li, F. Ding, and A. Beheshti, "Mixed graph neural network-based fake news detection for sustainable vehicular social networks," *IEEE Trans. Intell. Transp. Syst.*, early access, Jul. 7, 2022, doi: [10.1109/TVTITS.2022.3185013](https://doi.org/10.1109/TVTITS.2022.3185013).
- [5] W. Huang, C. Su, and Y. Wang, "An intelligent work order classification model for government service based on multi-label neural network," *Comput. Commun.*, vol. 172, pp. 19–24, Apr. 2021.
- [6] Q. Li, L. Liu, Z. Guo, P. Vijayakumar, F. Taghizadeh-Hesary, and K. Yu, "Smart assessment and forecasting framework for healthy development index in urban cities," *Cities*, vol. 131, Dec. 2022, Art. no. 103971.
- [7] Q. Zhang, Z. Guo, Y. Zhu, P. Vijayakumar, A. Castiglione, and B. B. Gupta, "A deep learning-based fast fake news detection model for cyber-physical social services," *Pattern Recognit. Lett.*, vol. 168, pp. 31–38, Apr. 2023.
- [8] Z. Guo, D. Meng, C. Chakraborty, X.-R. Fan, A. Bhardwaj, and K. Yu, "Autonomous behavioral decision for vehicular agents based on cyber-physical social intelligence," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 4, pp. 2111–2122, Aug. 2023.
- [9] L. Zhao, Z. Yin, K. Yu, X. Tang, L. Xu, Z. Guo, and P. Nehra, "A fuzzy logic-based intelligent multiattribute routing scheme for two-layered SDVNs," *IEEE Trans. Netw. Service Manage.*, vol. 19, no. 4, pp. 4189–4200, Dec. 2022.
- [10] B. Žitko and H. Ljubić, "Automatic question generation using semantic role labeling for morphologically rich languages," *Tehnički vjesnik*, vol. 28, no. 3, pp. 739–745, 2021.
- [11] Z. Guo, K. Yu, A. Jolfaei, F. Ding, and N. Zhang, "Fuz-spam: Label smoothing-based fuzzy detection of spammers in Internet of Things," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 11, pp. 4543–4554, Nov. 2022.
- [12] J. Liu, Y. Shen, Y. Zhang, and S. Krishnamoorthy, "Resume parsing based on multi-label classification using neural network models," in *Proc. 6th Int. Conf. Big Data Comput.*, May 2021, pp. 177–185.
- [13] E. Deniz, H. Erbay, and M. Coşar, "Multi-label classification of e-commerce customer reviews via machine learning," *Axioms*, vol. 11, no. 9, p. 436, 2022.
- [14] N. K. Rajput and B. A. Grover, "A multi-label movie genre classification scheme based on the movie's subtitles," *Multimedia Tools Appl.*, vol. 81, no. 22, pp. 32469–32490, Sep. 2022.
- [15] Z. He, H. Wu, and G. Wu, "Spectral-spatial classification of hyperspectral images using label dependence," *IEEE Access*, vol. 9, pp. 119219–119231, 2021.
- [16] I. Ameer, N. Ashraf, G. Sidorov, and H. G. Adorno, "Multi-label emotion classification using content-based features in Twitter," *Computación y Sistemas*, vol. 24, no. 3, pp. 1159–1164, Sep. 2020.
- [17] S.-Z. Zhang, J. Wang, L.-B. Zhu, S. Toufeeq, X. Xu, L.-L. You, B. Li, P. Hu, and J.-P. Xu, "Quantitative label-free proteomic analysis reveals differentially expressed proteins in the digestive juice of resistant versus susceptible silkworm strains and their predicted impacts on BmNPV infection," *J. Proteomics*, vol. 210, Jan. 2020, Art. no. 103527.
- [18] J. Xiong, L. Yu, X. Niu, and Y. Leng, "XRR: Extreme multi-label text classification with candidate retrieving and deep ranking," *Inf. Sci.*, vol. 622, pp. 115–132, Apr. 2023.
- [19] H. Bouziane and A. Chouarfia, "Use of Chou's 5-steps rule to predict the subcellular localization of gram-negative and gram-positive bacterial proteins by multi-label learning based on gene ontology annotation and profile alignment," *J. Integrative Bioinf.*, vol. 18, no. 1, pp. 51–79, Mar. 2021.
- [20] J. Kim, H. Kim, T. Kim, N. Kim, and Y. Choi, "MLPD: Multi-label pedestrian detector in multispectral domain," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7846–7853, Oct. 2021.
- [21] X. Yu, J. Sun, W. Wang, L. Jiang, B. Cheng, and J. Fan, "Assessment of the fusion tags on increasing soluble production of the active TEV protease variant and other target proteins in *E. coli*," *Appl. Biochem. Biotechnol.*, vol. 182, no. 2, pp. 769–781, Jun. 2017.
- [22] L. Humphreys, G. Boella, L. van der Torre, L. Robaldo, L. Di Caro, S. Ghanavati, and R. Muthuri, "Populating legal ontologies using semantic role labeling," *Artif. Intell. Law*, vol. 29, no. 2, pp. 171–211, Jun. 2021.
- [23] Á. Aldunate, S. Maldonado, C. Vairetti, and G. Armelini, "Understanding customer satisfaction via deep learning and natural language processing," *Expert Syst. Appl.*, vol. 209, Dec. 2022, Art. no. 118309.
- [24] R. Widyaningrum, I. Candradewi, N. R. A. S. Aji, and R. Aulianisa, "Comparison of multi-label U-Net and mask R-CNN for panoramic radiograph segmentation to detect periodontitis," *Imag. Sci. Dentistry*, vol. 52, no. 4, pp. 383–391, 2022.
- [25] P. Schrempf, H. Watson, E. Park, M. Pajak, H. MacKinnon, K. W. Muir, D. Harris-Birtill, and A. Q. O'Neil, "Templated text synthesis for expert-guided multi-label extraction from radiology reports," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 2, pp. 299–317, Mar. 2021.
- [26] W. Kaur, V. Balakrishnan, and K.-S. Wong, "Improving multi-label text classification using weighted information gain and co-trained multinomial naive Bayes classifier," *Malaysian J. Comput. Sci.*, vol. 35, no. 1, pp. 21–36, Jan. 2022.
- [27] L. Nie, T. Chen, Z. Wang, W. Kang, and L. Lin, "Multi-label image recognition with attentive transformer-localizer module," *Multimedia Tools Appl.*, vol. 81, no. 6, pp. 7917–7940, Mar. 2022.
- [28] T. Matsui, K. Suzuki, K. Ando, Y. Kitai, C. Haga, N. Masuhara, and S. Kawakubo, "A natural language processing model for supporting sustainable development goals: Translating semantics, visualizing nexus, and connecting stakeholders," *Sustainability Sci.*, vol. 17, no. 3, pp. 969–985, May 2022.
- [29] A. Blanco, A. Pérez, and A. Casillas, "Exploiting ICD hierarchy for classification of EHRs in Spanish through multi-task transformers," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 3, pp. 1374–1383, Mar. 2022.
- [30] A. Brandsen and M. Koole, "Labelling the past: Data set creation and multi-label classification of Dutch archaeological excavation reports," *Lang. Resour. Eval.*, vol. 56, no. 2, pp. 543–572, Jun. 2022.



LI LEI was born in Changde, Hunan, China, in 1983. She received the master's degree from Central South University, China. She is currently works with the School of General Education, Hunan University of Information Technology. Her research interests include applied linguistics, contrastive analysis between english, chinese, and foreign language teaching.



HAO WANG was born in Changde, Hunan, China, in 1981. He received the master's degree from Central South University, China. He is currently works with the School of Foreign Language, Changsha Normal University. His research interest includes Translation Theories and Practice.

• • •