HighTech and Innovation Journal

ISSN: 2723-9535

Available online at www.HighTechJournal.org

HighTech and Innovation Journal

Vol. 5, No. 1, March, 2024

Innovative Label Embedding for Food Safety Comment Classification: Fusion of Self-Semantic and Self-Knowledge Features

Yiming Zhang [1*], Haozheng Liu [1], Jiaming Feng [1], Xu Zhang [1]

[1] *Department of Computer Science and Technology, Henan Agricultural University, Zhengzhou 450002, China.*

Received 19 October 2023; Revised 01 February 2024; Accepted 08 February 2024; Published 01 March 2024

Abstract

Food safety comment classification represents a specialized task within the realm of text classification. The objective is to efficiently identify a large volume of food safety comments, aiding relevant authorities in timely food analysis and safety alerts. Traditional methods typically employ one-hot encoding for label processing. However, in real-world situations, classified labels often convey valuable semantic information and guidance. This paper introduces an innovative approach to enhance the classification performance of food safety comments by embedding label information. Initially, we extracted generic sentiment pivot words from various classification labels as label description information. Subsequently, we employ a joint embedding approach to integrate this label description information into the text. This process will pool the expressions of the pivot word into the corresponding sentiment labels in the known domains after averaging to get the embedded expression. This aims to acquire highly detailed self-semantic feature vectors and self-knowledge feature vectors that are integrated with labeled descriptive information. Then, feed the semantic representation of comments and the word-embedded representation of labeled description information into a time-step-based multilayer Bi-LSTM and a step-based multilayer CNN, respectively. Ultimately, we concatenate these two feature vectors to facilitate matching, thereby fusing the self-semantic and self-knowledge features of labeled description information to train a classification model for food safety comments. Experimental results on the food safety comment dataset showcase a noteworthy improvement of 1.74% and 1.27% in Macro_Precision and Macro_F1 metrics, respectively, compared to BERT, BERT-RNN, and BERT-CNN. Through extensive ablation experiments and additional studies, our method effectively embeds labeling information, demonstrating a clear advantage over traditional methods in the task of classifying food safety comments.

Keywords: BERT; Label Embedding; Siamese Network; Pre-trained Models; Short Text Classification; Food Safety Lead Discovery.

1. Introduction

Given the rapid advancement of social sciences, technology, and the economy, especially in recent years, the mobile internet has undergone remarkable expansion. This rapid growth has facilitated the swift emergence and widespread adoption of Social Media Platforms. These platforms offer individuals convenience and openness, allowing them to express their opinions and comments on social media at any time [1]. Within this context, a new phenomenon has surfaced: food safety comments [2]. This refers to internet-based food safety information originating from various catering industry merchants, known for its rapid dissemination and high level of attention [3]. With the internet boasting a broad user base and food safety being inherently tied to the country and our daily lives, the issue of food safety is particularly pressing and significant.

* Corresponding author: yimingzhang@stu.henau.edu.cn

http://dx.doi.org/10.28991/HIJ-2024-05-01-013

The objective of classifying comments on food safety is to efficiently recognize food safety concerns and offer analysis and early warning information to relevant regulatory authorities [4, 5]. This enables the timely resolution of food safety issues. Nevertheless, faced with the substantial volume of daily-generated food safety information, the key challenge lies in accurately classifying this information according to specific rules and assigning suitable labels [6, 7].

The sentiments expressed in food safety comments are fine-grained and sentence-level, including comments on specific issues like bugs and cold cakes. However, the task of food safety lead discovery goes beyond simple sentiment analysis. For example, words like "bad" and "super bad" may convey negative sentiments in everyday language, but they are represented positively in food safety comment labels. This presents a unique challenge compared to classifying everyday sentiments [8]. Additionally, comments within the food safety domain exhibit diverse characteristics, such as colloquialism, the prevalent use of emotional vocabulary, relatively concise text, and direct or indirect portrayal of textual information. These domain-specific features pose significant challenges for the classification of comments in the realm of food safety.

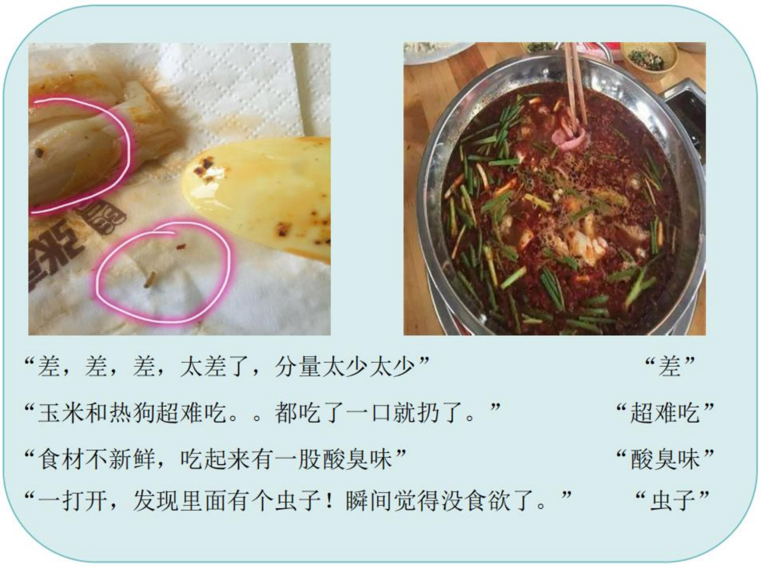
For example, take into account the subsequent comment sentences: "Not that flavor, not very good,", "Poor, poor, poor, so bad, portions are too small, too small.", "Corn and hot dogs were super bad.", "Both threw them away after one bite." Sentiment words such as "not good", "hard to eat" and "bad" directly express the sentimental direction of the overall sentence, in fact, the sentimental content is not actually related to the food safety issue. Conversely, the comment sentences "The flavor and portion size are very good, I was very happy to eat it yesterday, but today I woke up and started to have diarrhea, and I had diarrhea until I was dehydrated ......", "The ingredients are not fresh, and it tastes like a sour smell. ", "As soon as I opened it, I found a worm inside! Instantly feel no appetite.". In these sentences, expressions such as "diarrhea" indirectly point to food safety issues, while domain-specific words such as "sour smell" and "bugs" indirectly relate to food safety issues. (The detailed Chinese translations of the text and the corresponding pivot words are shown in Figure 1). Furthermore, a large number of sentences utilize the negative rule and the adverbial rule of degree, which invert and intensify the sentiment of the sentence. If we can link these domain-specific vocabularies with the corresponding labeling information, we will be better equipped to handle the intricacy of domain sentence vocabularies as well as the direct and indirect representation of textual information.

The classification of food safety lead comments typically involves traditional machine learning and deep learning methods. Traditional machine learning methods require intricate feature engineering, such as sparse lexical features (e.g., bag-of-words models, N-grams) [9], and depend on large amounts of labeled data. Currently, most of the research on text classification has shifted towards deep learning methods, including convolutional neural networks (CNN) [10–12] and recurrent neural networks based on long short-term memory (LSTM) [13]. These models utilize sliding windows to extract syntactic and semantic information from N-grams, capturing the hidden information of each token through autoregressive modeling. They also adjust word embedding representations based on the semantics of contextual words, ensuring that words have distinct embedding representations in different contexts. However, all of these methods achieve excellent results only in terms of feature extraction and do not consider the embedded representation of word meanings.

Recently, pretrained language models (PLMs) such as BERT [14] have demonstrated remarkable performance in the field of lexical representation embedding. They have achieved state-of-the-art results in all 11 natural language processing tasks. However, these pretrained models are trained using extensive general-purpose corpora, and there is a significant lack of fine-grained sentiment analysis in specific food safety domains. Miyazaki et al. [15] recently employed various types of tweet label descriptions based on a pretrained model to enhance the fine-grained classification performance of tweet data. Nevertheless, task-specific label descriptions are often insufficient in real-world scenarios. Subsequently, Zhang et al. [16] proposed a description-enhanced label embedding comparative learning method, which integrates external knowledge to obtain label description information. However, they acknowledged that the introduction of external knowledge may lead to unexpected fine-grained noise issues and therefore designed an interaction module to filter out the noise. Additionally, they developed a novel self-supervised relational (R^2) classification task to effectively represent and classify the labels.

Although these deep learning methods have had some success in text classification performance, they tend to either ignore the labeled information in the text or directly use external information to represent the labeled information through filtering in training. This limitation makes it challenging for them to represent domain-specific vocabulary and fully capture sentence meaning [17].

Based on the aforementioned problems, this paper proposes an innovative approach based on the label information of known domains to extract pivot words as label information, then get the self-semantic knowledge feature expression of the label embedding by means of joint embedding, and enhance the performance of text classification by fusing the self-semantic knowledge features. And the main contributions are as follows:



“差，差，差，太差了，分量太少太少”	“差”
“玉米和热狗超难吃。。都吃了一口就扔了。”	“超难吃”
“食材不新鲜，吃起来有一股酸臭味”	“酸臭味”
“一打开，发现里面有个虫子！瞬间觉得没食欲了。”	“虫子”

Figure 1. No Food Safety Accidents and Food Safety Accidents classes

- Proposing a method for classifying food safety comments with embedded label information. The method achieves text by selecting the pivot words of various kinds of labels in the text as the label description information, embedding the information through the joint learning method, using multilayer Bi-LSTM and multilayer CNN to extract the semantic features of the sentences and words, respectively, and employing feature fusion to concatenate the classification without losing the original information.
- Proposing an innovative label description and feature extraction scheme that utilizes a pivot word-based approach to represent label information through joint learning. This involves obtaining word embedding representations by avg-pooling using text corresponding to pivot words of various labeling classes in known domains and utilizing multilayer CNN with different step sizes to extract disjoint word-level features and acquire self-knowledge features at the word level.
- Introducing a novel method for fine-tuning the adaptation domain that quickly adapts to the domain and dynamically adjusts based on specific tasks.
- The experimental findings demonstrate that the approach enhances the classification of food safety comments and yields favorable outcomes in the experiment.

2. Related Works

Label embedding techniques have been widely utilized in the field of natural language processing (NLP), especially for tasks like text classification and multi-task learning within diverse networks. These techniques capture the semantic features of labels in the embedding network, significantly improving prediction performance, particularly for less discernible classes. In computer vision, label embedding techniques have been extensively used for image classification [18, 19], text recognition in images [20], and multi-modal learning involving both text and images [21, 22]. They have shown notable successes in zero-shot learning tasks [23, 24], contributing to predicting classes that have insufficient training samples.

In the domain of natural language processing (NLP), research has validated the efficacy of label embedding for tasks like text classification [25] and multi-task learning, particularly within the context of diverse networks. However, there is a gap in the literature concerning the effective implementation of label embedding and the complete utilization of label information to create text sequence representations that can enhance text classification performance.

Short text classification involves assigning concise textual content to predefined classes. Traditionally, machine learning algorithms are utilized in short text categorization to autonomously learn text features and conduct classification. The origins of short text classification can be traced back to the 1960s, initially relied heavily on knowledge engineering and manually defined rules for classification. However, with the increase in online text volumes and the development of machine learning in the 1990s, researchers shifted their focus towards addressing the challenges of deep learning text classification [26].

Currently, prevalent algorithms for short text classification include CNN, recurrent neural networks (RNN) [27], and attention mechanisms. Among these, RNN-based models exhibit distinct advantages in NLP tasks due to their ability to process input vectors sequentially in time steps, resembling the way humans comprehend textual information. Nonetheless, when dealing with long texts, RNN may face issues such as gradient vanishing or explosion, hindering their effectiveness in learning long-distance dependencies. In response to these challenges, LSTM has been introduced. LSTM incorporates mechanisms such as forgetting gates, memory gates, and output gates to better handle long texts, yielding significant results in text classification tasks.

Nonetheless, standard models may face constraints when addressing the unique characteristics of short texts, including sparsity, singularity, dynamics, and crossover. The Siamese neural networks, introduced by Zagoruyko et al. [28], constitute an architecture consisting of two interconnected neural networks. This model is primarily employed within the framework of supervised learning, aimed at maximizing the differences in representation between various labels and minimizing the representation of the same label. In this paper, we adopt a Siamese-like network architecture, incorporating distinct neural network structures for different text types. This approach seeks to better capture text features and improve text classification performance.

The Siamese neural network, introduced by Zagoruyko, is a connected architecture comprising two artificial neural networks primarily employed in supervised learning tasks. When the weights of the two networks are not shared but consist of two distinct neural network layers, it is referred to as a pseudo-Siamese neural network. In the case of pseudo-Siamese neural networks, the two networks can differ, such as one being an LSTM while the other is a CNN.

Originally primarily used in face recognition tasks for extracting facial features and determining their ownership by the same individual, Siamese neural networks have also been widely utilized in tracking tasks. Within the field of text classification tasks, Siamese neural networks are viewed as a potential solution technique. However, due to the distinct nature of text data, effectively adjusting and employing Siamese neural networks to tackle challenges in text classification remains an area deserving of exploration.

In this paper, we employ a Siamese-like network architecture inspired by Siamese neural networks. We make specific adjustments to the neural network structure based on the diverse characteristics of different levels of text. For continuous sentence-level text, we utilize a multilayer Bi-LSTM network to extract features by considering successive time steps, aiming to capture complex semantic relationships and contextual information. For discontinuous word-level label information, we employ a CNN based on different sliding windows to effectively extract label description information. Subsequently, we integrate the crucial information from both labels and sentences into the classifier for prediction, leveraging the strengths of specific network architectures based on the nature of the text data to enhance the performance of our classification model.

3. Material and Methods

For word selection, we utilize the Weighted Log Likelihood Ratio (WLLR) metric proposed by Yu & Jiang [29] to judiciously select pivot words as label description information. Following that, we learn pivot word expressions in specific domains through joint learning: various classes of pivot words obtain embedded expressions in known domains and subsequently acquire high-quality induced expressions for the pivot words through avg-pooling lexicalization concepts. In the context of neural network models, this paper adopts a fusion model that integrates label embedding techniques and Siamese-like network tuning techniques to improve the accuracy of text classification. Specifically, we adopt a fusion model with a three-layer, two-class structure, including a text Look-up layer, a label Look-up layer, a text Learning layer, a label Learning layer, and a Fusion layers and Classifiers. Figure 2 illustrates the architecture of the model. Below is an explanation of every layer:

In the text Look-up layer, we analyze sentence-level text, denoted as $X^{(k)} = \{x_1, x_2, x_3, \cdots, x_T\}$ as input to the BERT encoder. This encoder has been fine-tuned to extract detailed feature information, resulting in excellent representations of text word embeddings.

In the label Look-up layer, adopting an embedding learning approach to represent pivot words $Y^{(k)}_{pos\_i/neg\_i} = \{y_1, y_2, y_3, \cdots, y_T\}$. We extracted the representations of the embedded pivot words in the known domain text. Subsequently, we employed word-level avg-pooling to generate a high-dimensional spatial representation for each pivot word.

In the text Learning layer, the word embedding representation from the encoder is input into a multilayer Bi-LSTM. This network captures intricate time-step-based feature information, and the resulting output is utilized to concatenate the vector sequences of the final layer of the forward and the initial layer of the reverse, yielding a deep-level semantic feature vector encompassing both the forward and reverse information of the entire sequence.

In the label Learning layer, convolutional layers with varying step sizes are utilized to extract features for positive and negative label description information. Subsequently, max-pooling is applied to obtain the feature vectors of fused labeled information knowledge. In order to address the potential negative effects that may be associated with the quality of the selected pivot words, we included a word-level attention layer in order to filter out the noise.

In the Fusion layers and Classifiers, fixed-length vectors (representations of sentence text, label description information) are concatenated and fused in the last dimension. These vectors are fed into the classifier to obtain the classification probability of the embedded labeling information through the Softmax activation function.

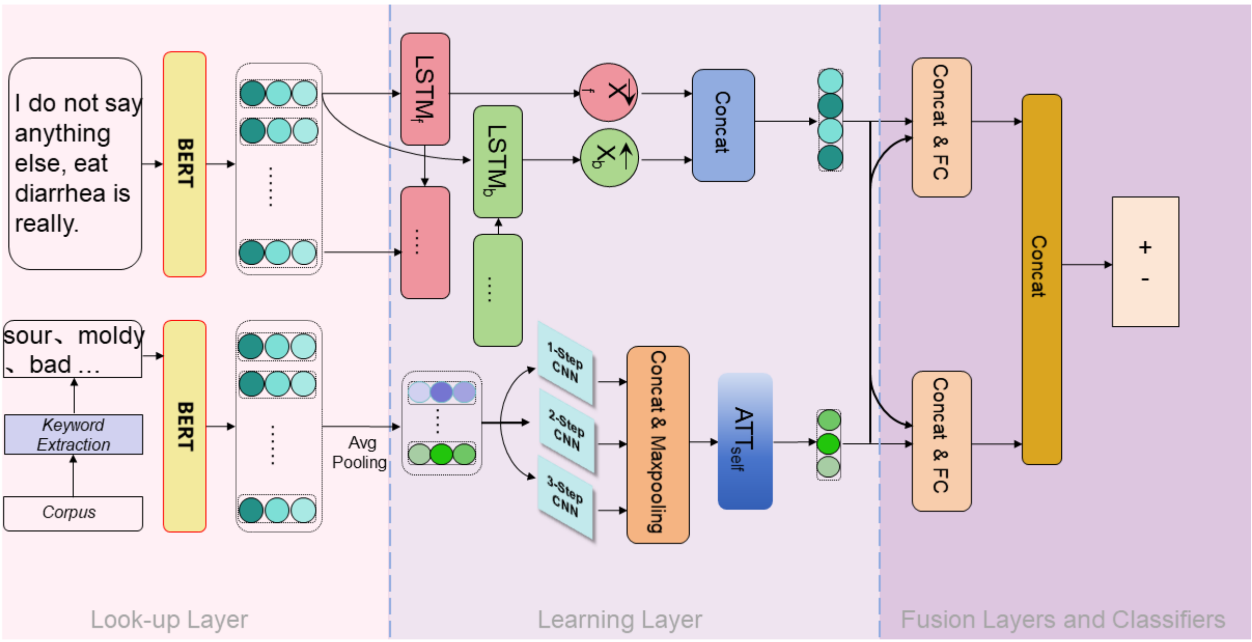


Figure 2. Fusion model

3.1. Look-up Layer

In the domain of text classification tasks, the intermediate step of text representation is of utmost importance. When compared to the original Word2Vec [30] and GloVe [31] word-embedding models, the BERT model, pre-trained on a large amount of open corpus., excels at extracting relational features among words in sentences. It captures these relational features at multiple levels, offering a more comprehensive reflection of sentence semantics. It is important to note that in the BERT-Base-Chinese model, Chinese text is segmented at the character level, diverging from traditional Chinese segmentation practices. To address this, we adopt the Chinese-BERT-WWM model, as proposed by the Harbin Institute of Technology (HIT) [32]. This model incorporates the Whole Word Masking (WWM) technique during training, utilizing Chinese Wikipedia data (both Simplified and Traditional Chinese).

The WWM technique offers an advantage by masking an entire word's composition with all the characters that make up that word. This compensates for the partial use of WordPiece tokenization in the pre-trained BERT model. Utilizing these benefits, we opt for the BERT pre-trained model as our encoder. In relation to pivot word expressions, we employ a word embedding approach. This involves obtaining sentence embedding expressions corresponding to pivot words with known domain label polarity and extracting the word embeddings of the pivot words to create the embedding expressions by avg-pooling. Equations 1 to 3 clarify the operations of the Look-up layer:

$$\mathrm{X}_1^{(\mathrm{k})} = \mathrm{BERT\_encoder}(\mathrm{X}^{(\mathrm{k})}) \tag{1}$$

$$\mathrm{Y}_{\mathrm{pos\_j1}}^{(\mathrm{k})} = \mathrm{AvgPooling}(\mathrm{BERT\_encoder}(\mathrm{Y}_{\mathrm{pos\_j}}^{(\mathrm{k})})) \tag{2}$$

$$\mathrm{Y}_{\mathrm{neg\_j1}}^{(\mathrm{k})} = \mathrm{AvgPooling}(\mathrm{BERT\_encoder}(\mathrm{Y}_{\mathrm{neg\_j}}^{(\mathrm{k})})) \tag{3}$$

Here, $X_1^{(k)}$ represents the embedding result of the sentence, where we make use of BERT's “Last_Hidden_State” as the representation of the text. Additionally, $Y_{pos\_j1}^{(k)}$ and $Y_{neg\_j1}^{(k)}$ denote the encoded representations of positive and negative generic pivot words, respectively. It is important to note that a single pivot word may have different representations in the training set. To ensure a consistent representation of a single keyword, we utilize a pooling layer. After conducting a series of experimental validations, we have opted for avg-pooling as the pooling operation due to its superior performance in the comparison between avg-pooling, max-pooling and so on.

3.2. Learning Layer

For different types of text, we employ diverse feature extraction strategies. At the sentence level, our main focus is to extract text information to create an overall semantic representation. A robust semantic representation offers potential advantages in the fusion learning of neural networks. Due to the long-term dependency challenges and susceptibility to gradient vanishing in traditional RNN models when dealing with lengthy sequences, we have chosen to use a model based on LSTM. LSTM excels at capturing long-term dependencies in sequential data through cellular states and gating mechanisms, making it well-suited for learning patterns and features in time-series data. As a result, we have implemented a multilayer Bi-LSTM model, processing both forward and reverse sequences and combining the information. This approach preserves the original feature information while obtaining an excellent semantic representation.

At the word level, discontinuous label description information like pivot words often lacks correlation between words. Methods for extracting continuous features may not be effective in such cases. On the other hand, Multilayer CNN(MCNN) tend to be more effective in processing word-level text by using various sizes of convolutional kernels to capture different aspect features. To address this, we introduce multiple convolutional layers, each using a different convolutional kernel size. Then, we apply the Tanh activation function and a max-pooling layer to obtain a comprehensive word-level representation. After this, we calculate "Label Attention" where in Label Attention weights $A_{pos/neg}$ and label attention vectors Y_{pos_j3/neg_j3} are are determined. Experimental results indicate that CNN outperform LSTM in extracting word-level features. The operations in the Learning layer are represented by the following Equations 4 to 7:

$$X_2^{(k)} = \mathrm{BiLSTM}(X_1^{(k)}) \tag{4}$$

$$Y_{pos\_j2/neg\_j2}^{(k)} = \text{Max-pooling}(\mathrm{MCNN}(Y_{pos\_j1/neg\_j1}^{(k)})) \tag{5}$$

$$A_{pos/neg} = \mathrm{Softmax}(H'\mathrm{ReLu}(HY_{pos\_j2/neg\_j2}^{(k)})) \tag{6}$$

$$Y_{pos\_j3/neg\_j3}^{(k)} = \sum_{d=1}^{D} A_{pos/neg} \cdot Y_{pos\_j2/neg\_j2}^{(k)} \tag{7}$$

Here, $X_2^{(k)}$ represents the semantic meaning representation of the embedded text. Additionally, $Y_{pos\_j2}^{(k)}$ and $Y_{neg_j2}^{(k)}$denote the combined information representation of positive and negative general pivot words. $H' \in \mathbb{R}^{24\times1}$ and $H \in \mathbb{R}^{dm\times24}$ signifies a fully-connected layer of the combined information representation of labeled attention. $D \in \mathbb{R}^{dm}$, $\cdot$ represents dot product. The features of these pivot words are extracted using different step-size MCNN based on varying step sizes combined with different sizes of convolutional kernels. Subsequently, the Tanh activation function is applied, followed by a max-pooling layer to obtain this composite information representation. The feature vectors of weighted labeled attention are then acquired using dot-products.

3.3. Fusion Layers and Classifiers

In the fusion layer, we utilize a concatenation strategy to maximize the utilization of information from both the text and label description representations while preserving the integrity of the original information. Specifically, we extract information from the overall semantic representation of the text and the self-semantic representations of label information. These are combined to form a joint representation through a concatenation operation on their last dimension. Subsequently, we input this joint representation into the classifier to obtain scores for the positive and negative labeling categories. The mathematical representation of the fusion layer operation is depicted in Equations 8 and 9:

$$S_j^{(k)} = \sigma(MP_{3m\times1}(X_2^{(k)} \oplus Y_{pos\_j3}^{(k)}) \oplus MN_{3m\times1}(X_2^{(k)} \oplus Y_{neg\_j3}^{(k)})) \tag{8}$$

$$L^{(k)} = -\sum_{j=1}^{n} \hat{y}_j^{(k)} \log(S_j^{(k)}) \tag{9}$$

The symbols used here are as follows: σ denotes the Softmax activation function, ⊕denotes the concatenation of vectors, $MP_{3m\times1} \in \mathbb{R}^{3*dm\times1}$ and $MN_{3m\times1} \in \mathbb{R}^{3*dm\times1}$ denote the linear layers, $S_j^{(k)}$denotes the probability of positive and negative classes, and $\hat{y}_j^{(k)}$ is the one-hot encoding of the true labels corresponding to $X^{(k)}$. The overall training objective is to minimize the weighted linear combinations from all classes.

4. Experimentation

4.1. Data Set

The experimental dataset originates from a real dataset in O2O Food-Safety-Review-master (CCF Big Data & Computing Intelligence Contest, *https://www.datafountain.cn/competitions/370/datasets*), encompassing a diverse collection of food safety comments. Our evaluation aims to examine the performance and robustness of our approach within the food safety domain. These food safety comments we divide them into two classes: "No FSA" (No Food Safety Accidents) and "FSA" (Food Safety Accidents), as illustrated in Table 1.

Table 1. Food Safety Review dataset

	Food Safety Review	No FSA	FSA
Train	10000	8346	1654
Test	2000	1563	437

4.2. Evaluation Metrics

In this study, due to the limited size of the dataset, we opted for a k-fold experiment rather than a random split. We utilized k-fold cross-validation to evaluate the model's performance, as illustrated in Table 2. The dataset was divided into 10 subsets, and a cyclic approach was used for model training and evaluation. Within each iteration, 9 subsets were dedicated to training and 1 subset for testing. This process was repeated 10 times, and the aggregated results were used to calculate the final performance metric by taking the mean value. We selected k=10 based on its suitability for our dataset in prior studies. Accuracy, precision, recall, and F1 score were employed as essential metrics for assessing the model's performance. Equations 10 to 13 outline the formulas for these metrics.

Table 2. Statistics

Fold	1..3	4..7	8..10
Train Set Size	10800	10800	10800
Test Set Size	1200	1200	1200

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{10}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{FP+FN}} \tag{12}$$

$$\text{F1} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \tag{13}$$

In these equations, TP (true positive) represents the count of positive classes correctly predicted as positive, FN (false negative) indicates the count of positive classes predicted as negative, FP (false positive) denotes the count of negative classes predicted as positive, and TN (true negative) signifies the count of negative classes correctly predicted as negative. The sum of these values provides the total number of tested samples.

Furthermore, we utilize macro-averages and weighted averages as evaluation metrics. Equations 14 to 16 outline the formulas for these metrics.

$$\text{Macro\_ Precision} = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{14}$$

$$\text{Macro\_ Recall} = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{15}$$

$$\text{Macro\_ F1} = \frac{1}{n}\sum_{i=1}^{n} F1_i \tag{16}$$

Here, n represents the number of classes. These metrics provide a broader perspective on the model's overall performance across different classes.

4.3. Use of Pre-trained Models

This paper utilizes the pre-trained word vector model from the Chinese-BERT-WWM, developed by HIT, for all experiments involving neural network models aimed at constructing input word embeddings. The model assigns a word vector to each individual word and is trained on Chinese Wikipedia, encompassing both Simplified and Traditional Chinese. It features a 12-layer, 768 hidden state Transformer model with 12 attentions and a vocabulary size of 21,128 words. Additionally, the pre-trained word vectors are continually learned and updated during the training process of the neural network model.

4.4. Baseline Model

To evaluate the effectiveness of our model, we executed a series of comparative analyses to assess its performance against various standard text classification models. Our objective is to thoroughly evaluate the performance and robustness of our model through comparisons with other established models. We selected a group of widely utilized text classification models and subjected them to testing on the same dataset. Subsequently, we utilized metrics like accuracy, precision, recall, and F1 score to comprehensively evaluate the performance of these models.

- ***BERT***: BERT (Bidirectional Encoder Representations from Transformers) is a groundbreaking pre-training model in the field of NLP, attracting significant attention. What sets BERT apart is its unique ability to understand linguistic context bidirectionally, leading to exceptional performance across a variety of NLP tasks. The model takes character-level raw word vectors as input and produces a word vector representation as output, encapsulating the entire semantic information of the text. This showcasing BERT's ability to capture nuanced contextual relationships within language. In this study, we employ a fine-tuned BERT-WWM based on HIT.
- ***BERT-CNN***: The collaborative BERT-CNN model seamlessly integrates the semantic understanding capabilities of BERT with the precise local feature extraction abilities of CNN, enhancing text classification performance. BERT grasps the semantics of the entire text, while CNN processes each text segment for more detailed semantic

features. This approach achieves a comprehensive understanding of textual information by combining effective semantic understanding with precise local feature extraction.

- ***BERT-RNN***: The BERT-RNN model, similar to BERT-CNN, leverages the advantages of BERT but differs by integrating RNN instead of CNN. The inclusion of the RNN module aims to capture semantic information within the text over longer distances. Unlike CNN, which is proficient in local feature extraction, the recurrent nature of RNN enables it to grasp long-term dependencies in sequential data, making it suitable for tasks that require understanding across broader contexts. Consequently, BERT-RNN combines the global context comprehension of BERT with RNN's ability to capture semantic subtleties across extensive portions of the text.
- ***TEXT-RNN***: The TEXT-RNN model standardizes all sentences to a uniform length, using pre-trained word vectors or random reordering of the vectors. Word2Vec is utilized for pre-training. In the standard configuration, the hidden states of forward/backward LSTM are obtained and concatenated at the final time step. A Softmax layer with Softmax activation function is then used for multi-class classification. An alternative method involves obtaining the hidden states of the forward/backward LSTM at each time step, concatenating them, and then averaging the concatenated hidden states across all time steps. This is followed by a Softmax layer to obtain the final classification result. The TEXT-RNN model aims to efficiently handle sentences of varying lengths for multi-class classification tasks.
- ***TEXT-CNN***: This model is similar to the aforementioned TEXT-RNN, but the difference lies in replacing RNN with CNN. TEXT-CNN excels in tasks such as text classification and sentiment analysis (SA), especially in handling shorter and more succinct texts efficiently. It improves model training efficiency and semantic understanding, enabling it to excel in various NLP tasks.
- ***ERNIE***: ERNIE (Enhanced Representation through Knowledge Integration) [33] model is firmly rooted in the continuous learning semantic understanding framework ERNIE and its corresponding pre-trained ERNIE model, which is established on the PaddlePaddle open-source platform. Remarkably, ERNIE outperforms BERT in performance across a total of 16 Chinese and English tasks, achieving State-of-the-Art results. Unlike BERT, ERNIE goes beyond mere semantic understanding; it explores the lexical structure, syntactic structure, and semantic information present in the training data. This synthesis approach substantially enhances the model's ability to generate general semantic representations. ERNIE proves to be a formidable pre-trained model for Chinese NLP, with robust text classification capabilities. In this study, we directly use open corpus pre-trained ERNIE.

4.5. Hyper-parameterization

For deep learning models, our method initializes word embeddings and label embeddings with 768-dimensional BERT word embeddings. These embeddings are derived through fine-tuning the BERT-WWM from HIT. Furthermore, the baseline model uses the same approach. Training involves utilizing the Adam Optimizer [34], with an initial learning rate set to 1e-5 and a mini-batch size of 64. The model is implemented using PyTorch and trained on an NVIDIA GPU 3060Ti. Table 3 presents the parameters for each network layer.

Table 3. Parameters

Parameters	Setting
optimizer	Adam
Learning_rate	1e-5
Num_layer	2
Hidden_size	256
Convolutional filtering window size k	1,2,3

4.6. Analysis of Experimental Results

In this sub-section, we focus on validating the following inquiries:

- Is the superiority of fusion models over traditional text classification models significant?
- Does the fusion model demonstrate an edge over the transformer-based variations of the approach, which have recently delivered noteworthy outcomes?
- Can our model demonstrate the capacity for effective domain adaptation?

To validate our approach, we conducted a comparative analysis employing four metrics against various baseline models. The results represented reflect the average value over 10 iterations. The outcomes are depicted in Table 4 and Figure 3.

Table 4. Classification results of the existing model and fusion model

	Accuracy	Precision	Recall	F1
BERT	0.9306	0.9695	0.9476	0.9584
BERT-RNN	0.9306	0.9674	0.9494	0.9583
BERT-CNN	0.9314	0.9685	0.9495	0.9589
TEXT-RNN	0.9167	0.9759	0.9256	0.9500
TEXT-CNN	0.9142	0.9648	0.9329	0.9486
ERNIE	0.9262	0.9632	0.9482	0.9556
Fusion model	**0.9366**	**0.9695**	**0.9545**	**0.9619**

Table 5. Macro averages of existing model and fusion model

	Macor_Precision	Macro_Recall	Macro_F1
BERT	0.8579	0.8928	0.8739
BERT-RNN	0.8618	0.8900	0.8750
BERT-CNN	0.8623	0.8923	0.8763
TEXT-RNN	0.8090	0.8896	0.8415
TEXT-CNN	0.8241	0.8713	0.8449
ERNIE	0.8572	0.8800	0.8680
Fusion model	**0.8753**	**0.8993**	**0.8866**

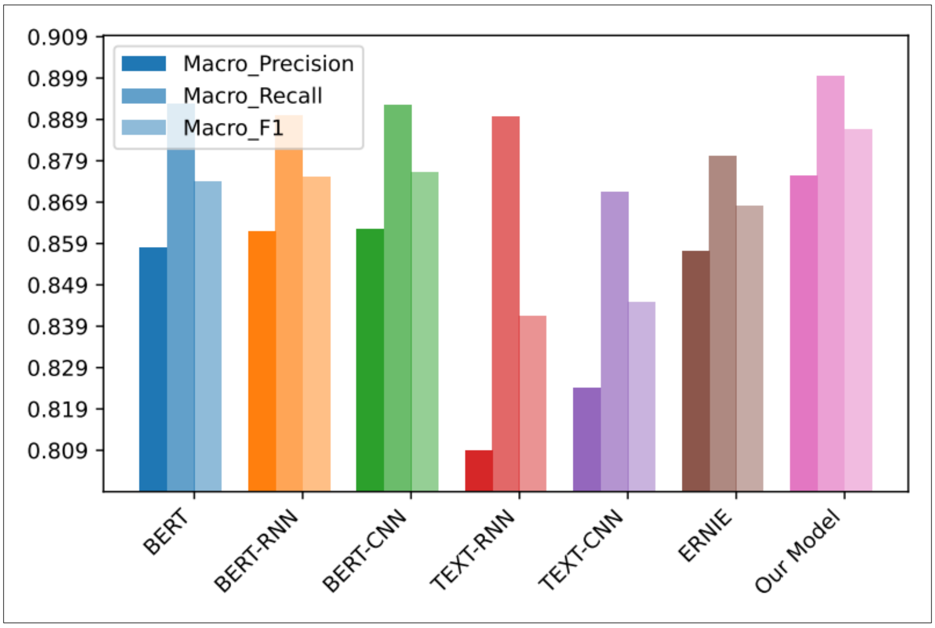


Figure 3. Macro_Precision, Macro_recall, Macro_F1 for the respective models

1. Is the advantage of fusion models over traditional text classification models significant?

Although traditional TEXT_RNN and TEXT_CNN exhibit commendable results, there is a noticeable gap when compared to other deep learning models. The fusion model demonstrates superior and more effective performance compared to traditional text classification models, which can be observed from the following two aspects.

Firstly, as shown in Table 4, the traditional TEXT-RNN and TEXT-CNN have an accuracy of 91.67% and 91.42%, respectively, while the fusion model achieves 93.66%. Clearly, the accuracy is approximately 2% higher. The accuracy directly validates that the fusion model is more effective than traditional methods. Additionally, from Table 1, it is evident that there is a large volume of No FSA, which indirectly reflects the model's accuracy in identifying No FSA. This greatly enhances our confidence in the effectiveness of classifying real food safety comments.

Secondly, based on Table 5, Macro_F1 which measures the combined model precision and recall, reached 84.15% and 84.49% for the traditional classification methods TEXT-RNN and TEXT-CNN, and 88.66% for the fusion model, which is also higher than about 2%. The macro average clearly reflects that the fusion model is superior compared to the traditional methods.

2. ***Does the fusion model have an advantage over the transformer-based variants of the approach that have recently achieved impressive results?***

The fusion model demonstrates continued effectiveness despite the recent success of transformer-based approaches. On the one hand, upon analyzing Table 4, it is evident that BERT, based on a 12-layer transformer encoder architecture, has shown impressive results across all 11 NLP tasks. Additionally, BERT-RNN and BERT-CNN, which incorporate extracted features using networks of RNN and CNN, still demonstrate have a 0.6% improvement in accuracy in recognizing No FSA. However, although not great news, it is important to note that despite the relatively high accuracy of BERT in the SA tasks, further improvements for BERT and its superior variants are challenging to achieve.

On the other hand, an analysis of Table 5 reveals a substantial 1.74% improvement in Macro_F1 of the fusion model compared to BERT, as well as BERT-RNN and BERT-CNN. Macro_F1 serves as a crucial metric for the SA task, and while the improvement in accuracy of the fusion model may not be substantial, Macro_F1 reflects the robustness and effectiveness of the model. This underscores the effectiveness of our work.

3. ***Can our model possess the capability of effective domain adaptation?***

Involving the domain adaptation problem, we train the Chinese NLP pre-trained model ERNIE without fine-tuning by utilizing various heterogeneous corpora, including dialogue data, news data, Wikipedia data, and other general-purpose corpora. As demonstrated in Table 4 and Table 5, its accuracy and Macro_F1 are 92.62% and 86.80%, respectively. The fusion model shows an average improvement of 1.5% on both metrics. However, considering that this is not enough to reflect the fine-grained problem of domain adaptation. Therefore, the subsequent case studies in Section 4.7 will further delve into this issue.

4.7. Case Study

To further illustrate the effectiveness of our models and the ability of domain adaptation to facilitate fine-grained identification, we present the prediction results for four test cases involving different models, as depicted in Table 6. For example (a), both BERT and ERNIE gave an incorrect sentiment prediction. For example (b), all three models predicted the correct sentiment. For example (c), BERT predicts correct sentiment, whereas ERNIE predicts NEG sentiment. For example (d), all models predicted correctly. In conclusion, the fine-tuned BERT accurately identified most sentiment labels. However, in cases where the sentiment was ambiguous, BERT was unable to determine whether the comment was POS or NEG. ERNIE, although not fine-tuned with information from the food safety domain, consistently predicted NEG sentiment in all four test cases, possibly due to the limitation of generalized corpora to express specific words. The fusion model precisely identified sentiment in all four test cases, and accurately recognized positive sentiment tendencies based on BERT, demonstrating its effective domain adaptation capabilities.

Table 6. Predictions of different methods on four test samples. POS and NEG denote No FSA, and FSA sentiments, respectively

Text\Model	Manual Label	BERT	ERNIE	Fusion Model
(a) 一点也不好吃 还这么贵，全是肥肉。(It's not good at all and it's so expensive. It's all fat.)	POS	(NEG, ✘)	(NEG, ✘)	(POS, ✔)
(b) 吃出来一只蚊子，胃口全无#9寸奥良烤鸡披萨# (Eat out a mosquito and lose your appetite #9' Orleans Grilled Chicken Pizza.)	NEG	(NEG, ✔)	(NEG, ✔)	(NEG, ✔)
(c) 味道一般，菜品也只能说一般，毕竟在那个地段，而且服务员态度也不热情。(The flavor was average, the food was only average, after all, it was in that location, and the waiters were not welcoming.)	POS	(POS, ✔)	(NEG, ✘)	(POS, ✔)
(d) 是不是放了几天了？太难吃了！有点馊的味道！(Has it been sitting there for a few days? It's awful! It tastes a bit rancid!)	NEG	(NEG, ✔)	(NEG, ✔)	(NEG, ✔)

5. Supplementary Experiments

5.1. Ablation Experiments

The fusion model includes three variants, all utilizing the same hyperparameters as our proposed methodology. The results can be found in Table 7 and Figure 4.

- **w/o LE** (Label Embedding):

Token embedding is achieved using the baseline fine-tuning BERT pre-training model that we have adopted. The embedded vector then undergoes two layers of Bi-LSTM to extract the last layer of the hidden state from both directions. The obtained vectors are spliced to generate fixed vectors, which are then input into the classifier.

- **w/o ATT** (Attention):

This method excludes the attention mechanism applied to the keyword vectors generated by the Label Learning layer. The resulting vectors, along with the directional quantities of sentence features, are spliced in the last dimension and fed into the matcher to obtain the match score.

- **w/o MC** (MCNN):

This variant is based on our optimal model but excludes the MCNN. It uses only CNN with observation horizons 1, 2, or 3, selecting the result with the best observation horizon.

Table 7. The outcomes of the ablation experiments

	Macro_Precision	**Macro_Recall**	**Macro_F1**
Fusion Model	0.8753	0.8993	0.8866
w/o LE	0.8707	0.8893	0.8796
w/o ATT	0.8673	0.8942	0.8799
w/o MC	0.8611	0.8777	0.8691

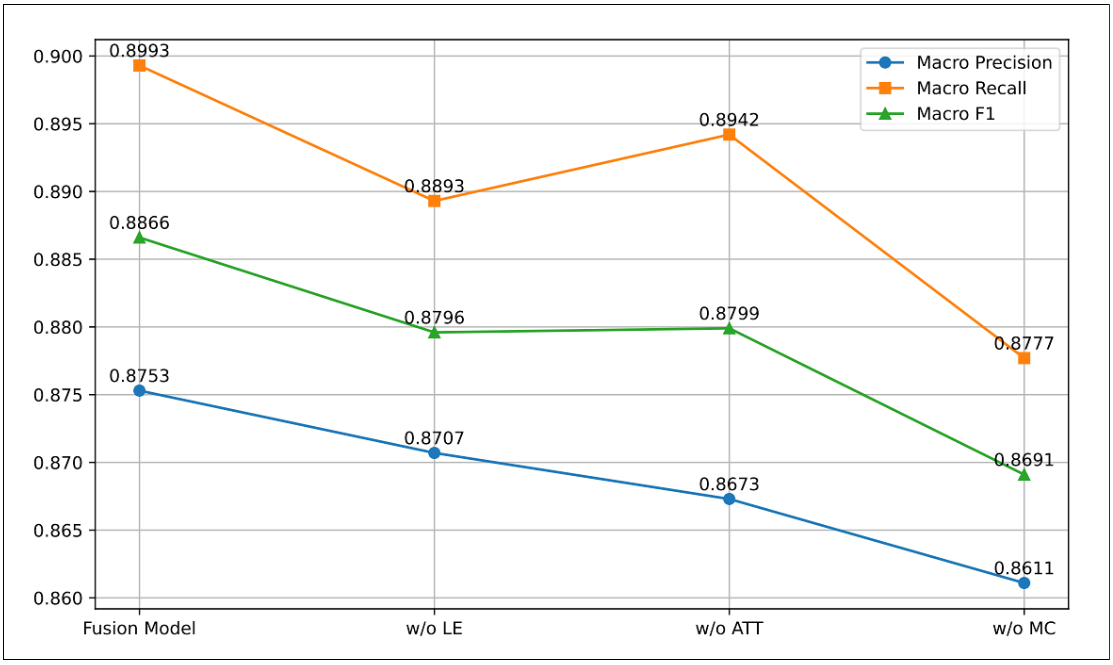


Figure 4. Macro_Precision, Macro_recall, Macro_F1 for the ablation experiments

The analysis of Table 7 reveals several key insights into the performance of our proposed methods and their interactions with different model components:

- **w/o LE Variant:** The w/o LE variant is a pairing of BERT and Bi-LSTM, which has shown consistent performance improvement in SA tasks. This indicates a synergistic effect between our model and BERT variants, emphasizing the improved performance attained through their combination.
- **w/o ATT Variant:** The w/o ATT variant, which eliminates the Attention layer with the same parameters, shows a noteworthy impact on performance. Macro_Precision, Macro_Recall, and Macro_F1 demonstrate a decrease, implying that the attention mechanism plays a crucial role in capturing features that the model overlooks or fails to learn.
- **w/o MC Variant (Without MCNN):** The model's performance is only moderate across all metrics when the MCNN is removed based on different step sizes. This indicates that the MCNN significantly contribute to the overall effectiveness of the model, and their exclusion leads to a noticeable impact on performance.

In conclusion, the analysis emphasizes the collaborative nature of the model components, with the Attention layer and MCNN playing pivotal roles in achieving superior performance. Furthermore, the positive interaction with BERT variants confirms the effectiveness and versatility of our proposed label embedding model.

5.2. Comparison Experiments

This section aims to investigate whether the length of the word list influences the performance of our model. The hypothesis is that a longer word list, and consequently more embedded information, could potentially enhance the model's learning capabilities. To investigate this, a comparative experiment is conducted using a word list length of 4 as the benchmark under the same parameters. This exploration aims to illuminate the correlation between word list length and model performance, offering insights into how the quantity of embedded information may influence the learning process and outcomes. The findings are presented in Table 8 and Figure 5.

Table 8. The outcomes of the comparative experiments

	Macor_Precision	Macro_Recall	Macro_F1
1/4 word	0.8622	0.8820	0.8716
2/4 word	0.8753	0.8993	0.8866
3/4 word	0.8706	0.8796	0.8750
All word	0.8693	0.8929	0.8804

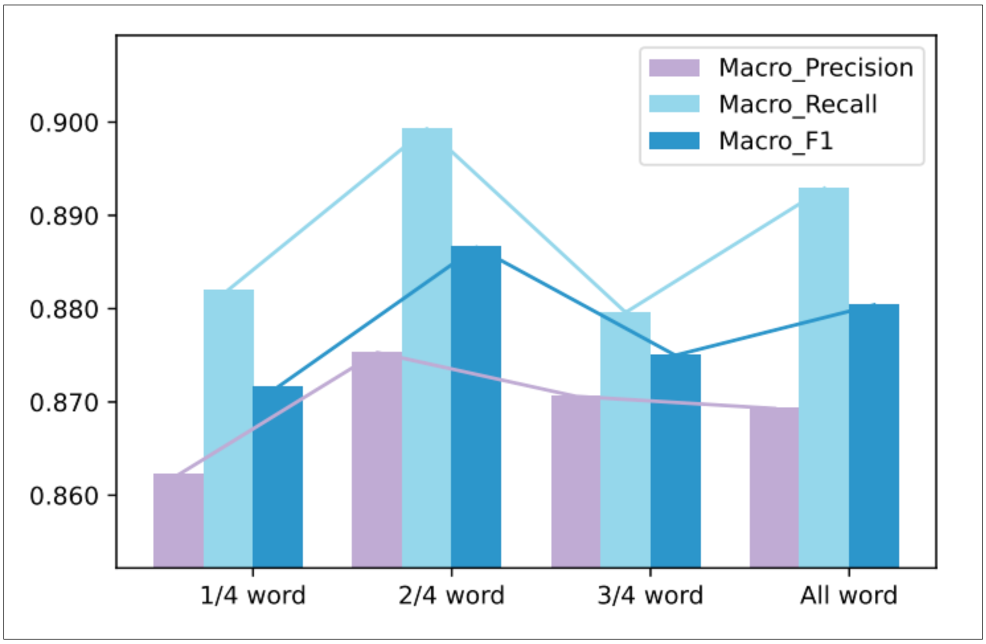


Figure 5. Macro_Precision, Macro_recall, Macro_F1 for different word list lengths

The analysis of Table 8 reveals some intriguing patterns in contrast to our initial expectations. As the word list length increases, the model performance initially improves, followed by a decline, and then a subsequent rise. After conducting numerous experiments and careful consideration, we focused on the pivot word selection weight indicator (WLLR). Upon closer examination, it was observed that the negative generic sentiment words in the early part of the word list have higher relevance to known domains. However, as the word list length expands, the relevance decreases, and there is even a risk of misclassification. We hypothesize that the expansion of the word list length introduces a noise problem. The inclusion of the Attention layer in our model, when the word list length reaches its full extent, synthesizes this information, reducing the weight of unfavorable words and consequently benefiting the model.

This analysis underscores the importance of continuous optimization in our pivot word algorithm. Developing a more refined algorithm for the selection of general pivot words in sentiment analysis remains an area for improvement, ensuring the model's robustness and effectiveness.

6. Conclusion

In this paper, we present an innovative method to classify food safety comments using fused self-semantic and self-knowledge feature models. Introducing an auxiliary task filters positive and negative pivot words from known domains as label description information, embedding them into expressions through co-embedding. This not only tackles the challenge of domain adaptation but also introduces a novel labeling representation that incorporates labeling information into a neural network model. The result is a significant improvement in the effectiveness of domain-heavy text classification tasks, especially for short texts related to food safety.

Looking ahead, our future research will focus on the field of multi-label fine-grained cross-field classification tasks, which is a cutting-edge and challenging domain in research. Considering the wide-ranging applications of multi-label, fine-grained classification tasks in real-world scenarios, we aim to extend the methods proposed in this paper to these domains. This exploration will provide valuable insights and inspiration, contributing to the enhancement of the universality and applicability of text classification methods within the food safety domain. We expect that ongoing research will continue to advance the application and development of NLP in the field of food safety and beyond.

7. Declarations

7.1. Author Contributions

Conceptualization, Y.Z.; methodology, Y.Z.; formal analysis, Y.Z.; investigation, X.Z.; data curation, J.F.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z. and H.L.; supervision, X.Z.; project administration, H.L. All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7.4. Acknowledgements

Special thanks to Shufeng Xiong, Distinguished Professor of Henan Agricultural University, for his guidance and help from the Natural Language Processing Laboratory.

7.5. Institutional Review Board Statement

Not applicable.

7.6. Informed Consent Statement

Not applicable.

7.7. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. References

[1] Seo, S., Almanza, B., Miao, L., & Behnke, C. (2015). The Effect of Social Media Comments on Consumers' Responses to Food Safety Information. Journal of Foodservice Business Research, 18(2), 111–131. doi:10.1080/15378020.2015.1029384.

[2] Lobb, A. (2005). Consumer trust, risk and food safety: A review. Food Economics-Acta Agriculturae Scandinavica, Section C, 2(1), 3-12. doi:10.1080/16507540510033424.

[3] Jin, C., Bouzembrak, Y., Zhou, J., Liang, Q., Van Den Bulk, L. M., Gavai, A., ... & Marvin, H. J. (2020). Big Data in food safety-A review. Current Opinion in Food Science, 36, 24-32. doi:10.1016/j.cofs.2020.11.006.

[4] Wang, J., & Yue, H. (2017). Food safety pre-warning system based on data mining for a sustainable food supply chain. Food Control, 73, 223–229. doi:10.1016/j.foodcont.2016.09.048.

[5] Geng, Z., Shang, D., Han, Y., & Zhong, Y. (2019). Early warning modeling and analysis based on a deep radial basis function neural network integrating an analytic hierarchy process: A case study for food safety. Food Control, 96, 329–342. doi:10.1016/j.foodcont.2018.09.027.

[6] Van de Brug, F. J., Lucas Luijckx, N. B., Cnossen, H. J., & Houben, G. F. (2014). Early signals for emerging food safety risks: From past cases to future identification. Food Control, 39(1), 75–86. doi:10.1016/j.foodcont.2013.10.038.

[7] Huang, Y., Wang, X., Wang, R., & Min, J. (2022). Analysis and Recognition of Food Safety Problems in Online Ordering Based on Reviews Text Mining. Wireless Communications and Mobile Computing, 2022, 1–15. doi:10.1155/2022/4209732.

[8] Li, Y., Gao, X., Du, M., He, R., Yang, S., & Xiong, J. (2020). What causes different sentiment classification on social network services? Evidence from weibo with genetically modified food in China. Sustainability (Switzerland), 12(4), 1345. doi:10.3390/su12041345.

[9] Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference, 2, 90–94.

[10] Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference, 1, 655–665. doi:10.3115/v1/p14-1062.

[11] Zhang, H., Xiao, L., Chen, W., Wang, Y., & Jin, Y. (2018). Multi-task label embedding for text classification. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP, 4545–4553. doi:10.18653/v1/d18-1484.

[12] Tan, C., Ren, Y., & Wang, C. (2023). An adaptive convolution with label embedding for text classification. Applied Intelligence, 53(1), 804–812. doi:10.1007/s10489-021-02702-x.

[13] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[14] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1, 4171–4186. doi:10.48550/arXiv.1810.04805.

[15] Miyazaki, T., Makino, K., Takei, Y., Okamoto, H., & Goto, J. (2019). Label embedding using hierarchical structure of labels for twitter classification. EMNLP-IJCNLP - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 6317–6322. doi:10.18653/v1/d19-1660.

[16] Zhang, K., Wu, L., Lv, G., Chen, E., Ruan, S., Liu, J., Zhang, Z., Zhou, J., & Wang, M. (2023). Description-Enhanced Label Embedding Contrastive Learning for Text Classification. IEEE Transactions on Neural Networks and Learning Systems, 1-14. doi:10.1109/TNNLS.2023.3282020.

[17] Hambrick, D. Z., & Meinz, E. J. (2011). Limits on the predictive power of domain-specific experience and knowledge in skilled performance. Current Directions in Psychological Science, 20(5), 275–279. doi:10.1177/0963721411422061.

[18] Akata, Z., Perronnin, F., Harchaoui, Z., & Schmid, C. (2016). Label-Embedding for Image Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(7), 1425–1438. doi:10.1109/TPAMI.2015.2487986.

[19] Qu, X., Che, H., Huang, J., Xu, L., & Zheng, X. (2023). Multi-layered semantic representation network for multi-label image classification. International Journal of Machine Learning and Cybernetics, 14(10), 3427–3435. doi:10.1007/s13042-023-01841-6.

[20] Rodriguez-Serrano, J. A., & Perronnin, F. (2013). Label embedding for text recognition. BMVC 2013 - Electronic Proceedings of the British Machine Vision Conference, 1-12. doi:10.5244/C.27.5.

[21] Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. Advances in Neural Information Processing Systems, 26, 2121-2129.

[22] He, S., Guo, T., Dai, T., Qiao, R., Shu, X., Ren, B., & Xia, S. T. (2023). Open-Vocabulary Multi-Label Classification via Multi-Modal Knowledge Transfer. Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023, 37(1), 808–816. doi:10.1609/aaai.v37i1.25159.

[23] Palatucci, M., Pomerleau, D., Hinton, G., & Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference, 22, 1410–1418.

[24] Zhang, H., Meng, X., Cao, W., Liu, Y., Ming, Z., & Yang, J. (2023). Graph embedding based multi-label Zero-shot Learning. Neural Networks, 167, 129–140. doi:10.1016/j.neunet.2023.08.023.

[25] Tang, J., Qu, M., & Mei, Q. (2015). PTE: Predictive text embedding through large-scale heterogeneous text networks. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 1165-1174. doi:10.1145/2783258.2783307.

[26] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., ... & He, L. (2020). A survey on text classification: From shallow to deep learning. arXiv preprint, arXiv:2008.00364. doi:10.48550/arXiv.2008.00364

[27] Chen, X., Qiu, X., Zhu, C., Wu, S., & Huang, X. (2015). Sentence modeling with gated recursive neural network. Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, 793–798. doi:10.18653/v1/d15-1092.

[28] Zagoruyko, S., & Komodakis, N. (2015). Learning to compare image patches via convolutional neural networks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June-2015, 4353–4361. doi:10.1109/CVPR.2015.7299064.

[29] Yu, J., & Jiang, J. (2016). Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings, 236–246. doi:10.18653/v1/d16-1023.

[30] Church, K. W. (2017). Word2Vec. Natural Language Engineering, 23(1), 155-162. doi:10.1017/S1351324916000334.

[31] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. EMNLP - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 1532–1543. doi:10.3115/v1/d14-1162.

[32] Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-Training with Whole Word Masking for Chinese BERT. IEEE/ACM Transactions on Audio Speech and Language Processing, 29, 3504–3514. doi:10.1109/TASLP.2021.3124365.

[33] Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., ... & Wu, H. (2019). Ernie: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223. doi:10.48550/arXiv.1904.09223.

[34] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. arXiv preprint, arXiv:1412.6980. doi:10.48550/arXiv.1412.6980.