Convolutional Neural Network Algorithm–Based Novel Automatic Text Classification Framework for Construction Accident Reports

Xixi Luo[1]; Xinchun Li[2]; Xuefeng Song[3]; and Quanlong Liu[4]

Abstract: Construction sites remain one of the most hazardous workplaces globally. To improve workplace safety in the construction industry and reduce the personal injuries and socioeconomic impacts resulting from workplace accidents, tacit knowledge containing fundamental causes of accidents or specific contextual factors can be extracted from past accident narrative reports. However, manually analyzing unstructured or semistructured textual data stored in records is a daunting task, and requires the use of automated and intelligent technologies to achieve rapid and accurate knowledge acquisition. Therefore, this paper proposes a text self-classification model based on deep learning natural language processing (NLP) technology for automated classification of construction site accident cases by accident type. First, combined with two statistical measures, mutual information and information entropy, the preprocessed text data were subjected to phrase segmentation to identify more complete and accurate accident precursor information without human intervention. Then a complete multilayer and multisize convolutional neural network (CNN) model was constructed using pretrained Word2Vec word embeddings for text self-classification tasks. Finally, the test results of the CNN classification algorithm were compared with the practical application results of three shallow learning algorithms, and the performance of different types of classification algorithms was evaluated. The results showed that the CNN-based deep learning algorithm developed in this paper demonstrated excellent feature extraction and learning abilities in the task of automatic text classification in the field of NLP. This not only demonstrated that reliable accident prevention knowledge could be obtained from the textual descriptions of construction accidents, but also provided a novel model reference for document archiving and information retrieval. **DOI: 10.1061/JCEMD4.COENG-13523.** © *2023 American Society of Civil Engineers.*

Author keywords: Deep learning; Natural language processing (NLP); Construction safety; Text classification; Accident injury types.

Introduction

The development of the global economy is greatly and inextricably impacted by the construction industry, which not only makes a considerable contribution to a nation's gross domestic product (GDP), but also propels advances in urbanization and population employment. However, the construction sector is regarded as one of the most hazardous in the world due to its unique traits, including being labor-intensive, having complicated working environments and construction procedures, and having diverse work tasks and equipment (Guo et al. 2021; Xu and Patrick 2021). According to estimates by the International Labour Organization (2021), more than 2.78 million deaths every year are caused by occupational accidents, along with 374 million workers suffering from nonfatal occupational accidents. Of these, approximately one of six fatal accidents occur in the construction industry. It is crucial to carry out in-depth analyses of previous incidents and to pinpoint accident precursors in advance to prevent accidents and improve safety and health in the workplace. It is now possible to draw lessons from past accident experiences thanks to the adoption and implementation of the Safety Accident Reporting and Investigation System, which has resulted in a number of standardized and particular manufacturing accident reports (Love et al. 2018). However, because computers cannot automatically identify and process text accident reports in unstructured or semistructured form, it is necessary to combine a natural language processing (NLP) system and machine learning (ML) methods to analyze and share tacit knowledge contained in large textual data sets (Khurana et al. 2023).

In recent years, with a significant investment in information technology and digital technology, the explosive growth of diverse data and artificial intelligence methods has had a tremendous impact on the process of information handling and decision-making in organizations. The construction industry has demonstrated strong adaptability to this trend and has produced a series of research achievements that use NLP and artificial intelligence (AI) methods to retrieve and analyze text data. For example, Chi et al. (2014) collected different types of construction safety documents to find appropriate safety methods. By extracting unsafe scenarios corresponding to various activities and hazards in the documents and using text classification technology, automatic job hazard analysis (JHA) was realized in the form of ontology knowledge base. Zhou and El-Gohary (2017) proposed an ontology-based information extraction algorithm aimed at automatically extracting regulatory requirements from building energy conservation regulations. This algorithm utilizes text semantic and syntactic features to handle complex, lengthy sentences and has successfully achieved fully

[1] School of Economics and Management, Chang'an Univ., Xi'an, Shaanxi 710064, China. Email: 183509403@qq.com

[2] Professor, School of Management, China Univ. of Mining and Technology, Xuzhou 221116, PR China (corresponding author). Email: lxx17602934024@163.com

[3] Professor, School of Management, China Univ. of Mining and Technology, Xuzhou 221116, PR China. Email: TB19070004B2@cumt.edu.cn

[4] Associate Professor, School of Management, China Univ. of Mining and Technology, Xuzhou 221116, PR China. Email: xxl2019@cumt.edu.cn

Note. This manuscript was submitted on January 4, 2023; approved on July 25, 2023; published online on September 29, 2023. Discussion period open until February 29, 2024; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Construction Engineering and Management***, © ASCE, ISSN 0733-9364.

automated energy compliance inspections in the field of construction. Luo et al. (2020) proposed three prediction frameworks of artificial neural network (ANN), support vector machine (SVM), and long short-term memory (LSTM) neural network to predict building energy load with multiobjective energy consumption and load demand, and finally achieved effective building energy management. In addition, NLP and ML techniques have also been applied to research topics involving free-text data in construction engineering safety causal analysis (Zhong et al. 2020), development of building information modeling (BIM) database search engines (Jung and Lee 2019), and historical change order similarity case searches (Ko et al. 2021), greatly promoting the development of production management and decision support systems in the construction industry.

As one of the important problems in natural language processing, text classification is the task of assigning a large amount of text data to predefined categories or labels, which not only helps decision makers better understand and find relevant information, but also can be used to discover potential knowledge and patterns in the text. Therefore, text classification has important significance in information organization and retrieval, decision support, and knowledge discovery, and also has important application value in the field of occupational safety, which has a large amount of text data such as accident reports, safety inspection records, and hidden danger investigation lists (Pan and Zhang 2021). However, manual text classification tasks require a lot of time and personnel, and there is also a problem of strong subjectivity and classification errors, for which there is an urgently need for researchers to develop efficient, accurate, and reliable automated text classification techniques. At the same time, through the identification of accident precursor factors, control of the safety production process, and improvement of key accident causes, it is possible to effectively prevent and reduce the occurrence of construction safety accidents. Scientific and effective identification of accident precursor factors is the focus of safety management work, and how to extract feature attributes from the free-text narrative and accurately correspond to possible injury outcomes is an effective way to improve safety performance (Wang 2021). In summary, in the field of occupational safety, exploring suitable algorithms or building scientific models to extract key features and learn representations from complex text data to carry out classification and induction tasks in a timely and accurate manner can help to better conduct risk assessment, prediction, and prevention.

In the field of construction safety, there have been some research achievements and methods related to text classification. However, most of the research tends to focus on the specific application of shallow learning algorithms in text classification tasks, which usually require manual feature design and rely too much on human-extracted text features for learning. In addition, shallow learning algorithms have high computational complexity when processing large-scale, high-dimensional data, and may suffer from overfitting and underfitting, which limits the classification performance of the algorithms. However, deep learning methods have the ability to automatically extract features, and methods based on deep learning show better performance than machine learning methods that require manual distributed pretraining, especially in large-scale data processing and complex model training. Therefore, it is necessary to explore more efficient and accurate deep learning algorithms to better cope with practical application scenarios, and to adjust and optimize the model structure based on the text characteristics of different fields, so as to improve the effectiveness and application value of text classification tasks. In addition, although in recent years some deep learning methods based on recurrent neural network (RNN), LSTM, and bidirectional encoder representations from transformers (BERT) (Jing et al. 2022; Gupta et al. 2022) have emerged for information mining of occupational safety accident reports, given the complex network structure of deep learning there is an urgent need to supplement and promote research on the application of deep learning algorithms to capture key information in construction safety accident reports and achieve automated text classification. Therefore, this study seeks to bridge the gap by integrating the field of NLP and leveraging deep learning techniques to establish an automatic classification model for different types of construction safety accidents. The performance of this model will be evaluated through a comparative analysis with three shallow learning methods: SVM, logistic regression, and naive Bayes (NB). This paper aims to try to promote significant challenges that exist in text preprocessing and classification accuracy in previous research.

Literature Review

Construction site managers and engineers typically store safety information collected during safety inspections in the form of data, text, images, and videos. Identifying and analyzing the recorded content is considered an effective means of hazard identification and effective monitoring to improve safety performance. As an important form of information recording, it has been asserted that NLP techniques can be used for text mining of unstructured text to identify safety patterns and causal trends in engineering projects. Therefore, this paper starts with a brief review of text mining and natural language processing–related literature to provide a contextual backdrop for the developed complete classification model framework based on deep learning methods.

Previous Studies on Accident Narrative Text Mining

Occupational safety management is a series of measures taken by enterprises to ensure the safety and health of employees in the production process. For a long time, safety management during the production process mainly relied on expert opinions and management personnel experience. Researchers also mainly used traditional data collection methods such as brainstorming, questionnaire surveys, expert interviews, and literature research for domain knowledge acquisition (Lette et al. 2018; Soliman 2018). However, these methods not only consume a lot of time and costs, but also have shortcomings such as incomplete data and subjective interference factors. With the increasing understanding of the importance and reference value of accident statistics data to the production process, safety management personnel have gradually realized the significant research value of accident reports that contain accident text descriptions, including time, location, site configuration parameters, and accident narrative records. Descriptive accident text provides a basis for an effective retrospective management mechanism and can be analyzed to obtain valuable information implicit in the reports (Graves et al. 2015). Accident narrative text analysis has been applied to knowledge acquisition in research fields such as aviation (Tanguy et al. 2016), railways (Hughes et al. 2018), the chemicals industry (Song and Suh 2019), and construction (Luo et al. 2020), and has gained considerable attention.

Different forms of narrative text reports can be selected depending on the research topic. For example, Das et al. (2018) used the National Highway Traffic Safety Administration (NHTSA) vehicle complaint report to analyze potential trends in consumer complaints and identify major vehicle defects. Kakhki et al. (2019) used the industrial injury claim data of agricultural comprehensive enterprises in the Midwest United States. Combining three machine learning technologies and extracting factors such as the injury site, the nature of the injury, and the cause of injury in the report, a prediction model for the severity of occupational accidents was constructed, which verified the importance of quantitative analysis of historical injury data to safety science. Xing et al. (2019) extracted and formalized domain knowledge from relevant standards and technical manuals of subway design, construction, and management, and constructed a subway construction safety risk identification ontology (SRI-Onto) by means of a seven-step method of ontology development, which provided a theoretical framework for the standardization and formalization of safety risk knowledge. Zhang et al. (2021) used the National Transportation Safety Board (NTSB) aviation data excerpt text report as the original database to construct a deep learning algorithm for accident probability and severity prediction. The unstructured narrative text based on natural language contains complete information of the accident process, which is regarded as the original data of the accident process. It has potential research value in the in-depth analysis of accident precursor factors, causal relationships, and prevention information, so it has also received extensive attention in industries with different development types and models.

Natural Language Processing

NLP is a branch of data science that integrates linguistics, computer science, mathematics, and other related fields to systematically analyze, understand, and extract information from textual data. With the assistance of technological tools, it is possible to manage large amounts of textual data and perform automated tasks, profoundly affecting the way people recognize, acquire, and apply information. The commonly used methods for processing NLP problems are mainly based on rule, statistical, and deep learning approaches (Hobson et al. 2019). Rule-based language models require researchers to manually write knowledge expressions and call rules to develop algorithms. When there is an increase in the amount of information to be processed, the written rules may not meet the requirements, so the statistical-based methods have gradually replaced rule-based methods in NLP processes. Statistical language models build a large corpus to enable the machine to learn text features and make judgments based on probability models. For example, commonly used algorithm models include SVM, decision trees (DTs), *k*-nearest neighbor (KNN), and hidden Markov model (HMM). However, this method has a data sparsity defect, meaning it has poor handling capabilities for certain special cases or rare vocabulary. This is because these situations occur infrequently in the corpus, resulting in the model being unable to accurately learn their characteristics and probability distribution, thereby affecting the model's accuracy and generalization ability. With the concept of deep learning proposed by Hinton et al. (2006), its excellent feature extraction and learning ability have received widespread attention and become the mainstream technology in many tasks in the NLP field. The commonly used algorithm models include convolutional neural network (CNN), RNN and LSTM network.

NLP, which aims to use computer intelligence to process human language, has been widely applied in machine translation, intelligent voice assistants, text autoclassification, sentiment analysis, and subject segmentation (Jiang et al. 2021). Wu et al. (2019) proposed a lightweight dynamic convolution model, which optimized the performance of machine language translation by combining dynamic convolution with a self-attention mechanism, and showed more excellent results in English–German translation tasks. Goh and Ubeynarayana (2017) obtained public accident narration from the Occupational Safety and Health Administration (OSHA) website and evaluated the classification performance of six machine learning algorithms, including SVM, linear regression (LR), random forest (RF), KNN, DT, and NB. It was found that the linear SVM model had the best classification effect on accident narration. Jin et al. (2018) improved the bi-LSTM model and bagging algorithm and proposed Bi-LSTMM-B, a new sentiment analysis model. The proposed model combines the advantages of deep learning that can extract abstract features with the idea of ensemble learning that can achieve a common decision by multiple classifier, which improves the accuracy of social media users' evaluation of sentiment analysis. The application of NLP in many complex tasks is gradually advancing, but there are still many difficulties to be overcome in the process of research field expansion and research model construction.

Application of Machine Learning Algorithms in Text Analysis for Construction Safety Management

In summary, with the rapid development of artificial intelligence and big data technology, NLP has been widely applied in the field of occupational safety. In this context, the use of machine learning algorithms can help automate the processing of large amounts of construction safety management texts, such as safety reports, accident reports, work plans, and safety training materials, to extract useful information and knowledge. This can help managers better identify safety risks and formulate effective accident prevention strategies, thereby supporting construction management and safety decision-making. Therefore, the practical application of machine learning algorithms in the analysis of construction safety management texts has great potential in four areas: safety risk assessment, relation extraction, text classification, and sentiment analysis. For example, Tixier et al. (2016) developed a fully automated NLP system based on manually encoded rules and a keyword dictionary to assess the risk of construction worker injuries using various machine learning algorithms such as RF and stochastic gradient tree boosting (SGTB). They explored the influence of attribute factors on different accident outcomes (injury type, energy type, body part, and severity of injury) and the effectiveness and reliability of machine learning algorithms in construction safety risk assessment. Li et al. (2016) proposed an automated approach to check the compliance of utilities in engineering drawings using NLP and spatial reasoning. They used NLP techniques to convert the text descriptions in engineering drawings into spatial rules that can be processed by computers, and then executed the extracted spatial rules in a logical order according to the geographic information system (GIS) to identify noncompliance issues. Experimental results validated that spatial reasoning mechanisms can effectively detect compliance issues with utilities and have high accuracy and efficiency.

Zhang et al. (2019) used five baseline algorithms to construct an ensemble model with optimized weights and applied the sequential quadratic programming (SQP) algorithm to search for the best weights of the ensemble model to classify the causes of construction accidents, which can be used to take preventive measures to effectively eliminate or reduce potential risks identified. Yao et al. (2021) applied social network analysis and sentiment analysis to investigate the sharing of construction safety knowledge on Twitter. They collected and analyzed the social structure of Twitter networks and opinion leaders related to construction safety, and calculated graph metrics, performed cluster analysis, emotion analysis, and correlation analysis to analyze important nodes, community structures, and information flow patterns in the network. The research results provide insights for promoting the sharing and dissemination of construction safety knowledge. In addition to the research results in construction safety management, there are many potential applications of machine learning algorithms in the construction field, focusing on cost control, material and equipment optimization, frame structure design, prefabricated structure manufacturing, construction site progress and quality management, building visualization, and sustainable circular economy.

Traditional machine learning algorithms such as SVM and random forests have demonstrated good information learning capabilities in tasks such as text classification, factor recognition, and risk prediction by predicting and judging the underlying rules in large amounts of historical data through shallow data structures. However, the process of feature construction requires manual extraction and cleaning of data, which leads to low efficiency when facing complex and large-scale tasks. As an important branch of machine learning, deep learning has powerful functionality and flexibility. By using more complex multilayer neural network architectures, deep learning can simulate the operation mechanism of the human brain, automatically learn features from sample data, and form more abstract high-level semantic features, thereby analyzing and exploring complex and diverse relationships hidden in the data. Therefore, compared with machine learning methods that require manual distributed pretraining, deep learning shows better performance and can adapt more quickly to the application of the construction field. But to the authors' knowledge, there is still a lack of sufficient achievements in combining deep learning methods for natural language tasks in the field of construction safety. Therefore, it is necessary to supplement and optimize the research gap between traditional prediction models and the advantages of deep learning network architecture from different perspectives. Therefore, exploring the representative algorithm of deep learning (CNN), constructing a complete process framework to automatically extract precursor factors in construction accident reports, and improving the accuracy of text classification has important research significance and practical application value.

Methodology

A framework for automatic accident type text classification for construction accident reports using NLP techniques based on deep learning is shown in Fig. 1.

Text Cleaning

Text cleaning is a fundamental operation in NLP, which aims to organize text corpora for completeness and standardization, making subsequent text analysis tasks more accurate and reliable. First, missing data in the text database are checked and the collected text data are integrated into a fixed format. Specifically, the collected construction accident text reports are sorted according to accident description, cause, type, and severity, so that subsequent classification, clustering, sentiment analysis, and other tasks can be carried

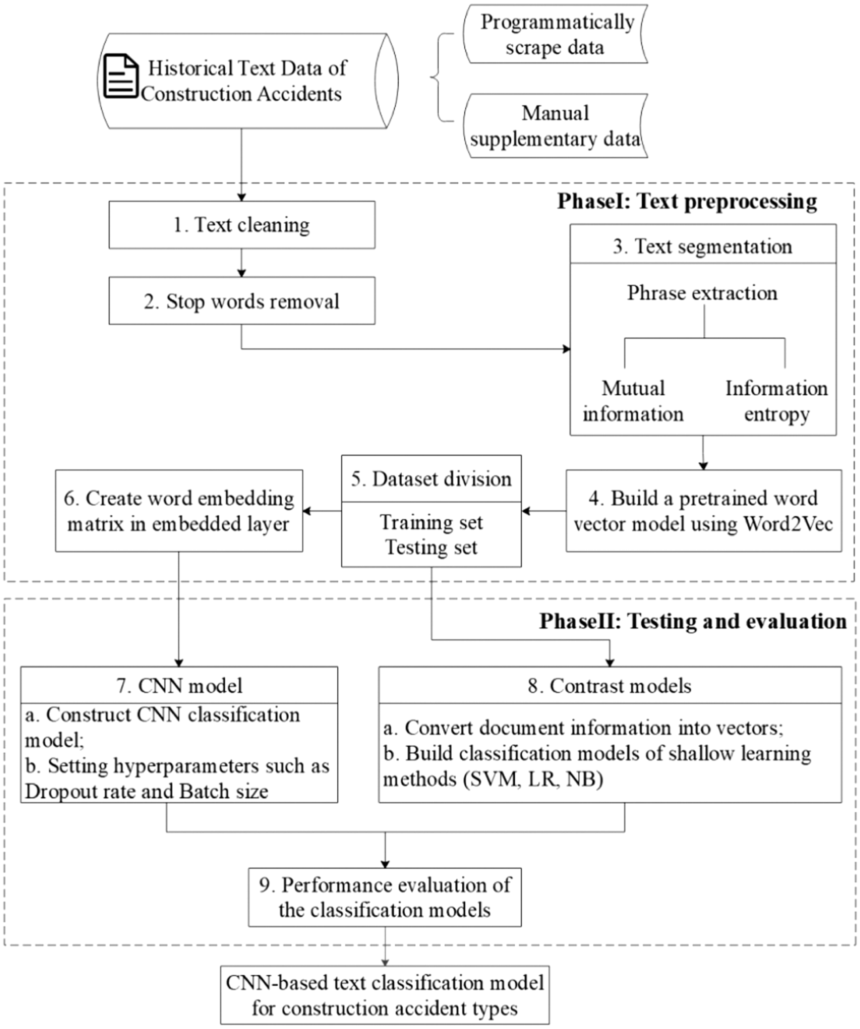
Historical Text Data of Construction Accidents
Programmatically scrape data
Manual supplementary data
PhaseI: Text preprocessing
1. Text cleaning
2. Stop words removal
3. Text segmentation
Phrase extraction
Mutual information
Information entropy
4. Build a pretrained word vector model using Word2Vec
5. Dataset division
Training set
Testing set
6. Create word embedding matrix in embedded layer
PhaseII: Testing and evaluation
7. CNN model
a. Construct CNN classification model;
b. Setting hyperparameters such as Dropout rate and Batch size
8. Contrast models
a. Convert document information into vectors;
b. Build classification models of shallow learning methods (SVM, LR, NB)
9. Performance evaluation of the classification models
CNN-based text classification model for construction accident types

Fig. 1. Research framework.

out more accurately. Additionally, text cleaning involves handling noise and useless information in the text. For example, data noise reduction, removal of faulty data, deletion of duplicate data, cleaning of punctuation and numbers, and text format conversion can be applied to text data.

Segmentation Using Phrase Extraction Technology

Chinese word segmentation is the first and foremost step in text processing and is also the foundation of human–machine natural language interaction. Therefore, the segmentation effect will directly affect the completeness and reliability of the basic information, thus further affecting the effectiveness of subsequent preaccident precursor extraction and text classification. Traditional Chinese text segmentation algorithms mainly rely on loading domain dictionaries, matching the target string with the words in the established dictionary through specific strategies, and the segmentation process is simple and efficient. However, the quality of text segmentation is directly influenced by the completeness and continuous updating of domain-specific dictionaries. To achieve more effective adaptive text segmentation and uncover unregistered vocabulary in construction accident reports, this paper investigates the practical application effects of phrase extraction techniques based on mutual information and adjacent information entropy (Chen et al. 2021).

Mutual Information

Mutual information reflects the degree of interdependence between the two variables. In text processing, mutual information between words refers to the degree of correlation between two words, which is represented by the mutual information value (MI) of the two words. The larger the mutual information value, the greater the correlation between the two objects, and the greater the probability they form a phrase (Peng et al. 2005). The calculation formula of mutual information can be expressed as

$$\mathrm{MI}(X,Y)=\log_2\frac{P(X,Y)}{P(X)P(Y)} \tag{1}$$

where X and Y = random variables of words; $P(X)$ and $P(Y)$ = probability the words X and Y appear in the corpus alone; $P(X,Y)$ = probability X and Y appear together in the corpus; and $\mathrm{MI}(X,Y)$ = correlation degree between X and Y. The larger the mutual information value MI, the higher the correlation between X and Y, and the greater the possibility of X and Y forming a phrase; on the contrary, the smaller the mutual information value, the lower the correlation between X and Y, and the more likely there is a phrase boundary between X and Y.

Information Entropy

The term *entropy* represents a measure of uncertainty of random variables. Information entropy can be used as a measure of system complexity. If the information entropy is relatively large, there are more types of different situations in the system, and the uncertainty is higher. Adjacency entropy is used to measure the uncertainty of left and right adjacent characters of the preselected words. The larger the left and right adjacency entropy, the more information the left and right adjacent characters contain, and the greater the possibility of the string becoming a word. Therefore, the left and right boundary of words can be determined by left and right adjacency entropy (Sun et al. 2006). The calculation formulas of left and right adjacency entropy are expressed as follows:

Left adjacency entropy

$$H_L(W)=-\sum_{W_l\in S_l}P(W_l|W)\log_2 P(W_l|W) \tag{2}$$

Right adjacency entropy

$$H_R(W)=-\sum_{W_r\in S_r}P(W_r|W)\log_2 P(W_r|W) \tag{3}$$

where S_l = set of left adjacent words of candidate word W; S_r = set of right adjacent words of candidate word W; $P(W_l|W)$ = conditional probability the left adjacent word is W_l when the candidate word W appears; and $P(W_r|W)$ = conditional probability the right adjacent word is W_r when the candidate word W appears.

The phrase extraction technique used in this study mainly relies on two statistical measures, mutual information and entropy, to identify the probability of forming a phrase based on the frequency of co-occurrence of adjacent combinations. It also evaluates the independence and semantic completeness of multiword expressions and uses both internal coherence and external boundaries as criteria to extract complete informative phrases. This approach can automatically extract meaningful phrase strings from text without loading domain-specific dictionaries, thereby improving the suitability of key phrase granularity obtained from text segmentation. The method is capable of extracting frequently co-occurring and repeatedly appearing collocations.

Deep Learning–Based Text Classification Model for Building Construction

CNN is one of the most representative neural networks in deep learning algorithms and has been widely applied in various fields, such as computer vision, NLP, and image processing (Fang et al. 2020). With the development of word embedding techniques and vector space models (VSMs) and the powerful feature learning and classification capabilities of deep learning methods based on CNN, it has achieved excellent experimental results in text classification tasks, proving that CNN cannot only handle tasks related to computer vision but also capture significant features in text using its multilayer perception structure, thus possessing more accurate text classification ability. A typical CNN consists of input layer, convolutional layer, pooling layer, fully connected layer, and output layer modules (Zhu and Chen 2020). The convolutional and pooling layers together form a convolution group, which learns local to global features by layer-wise feature extraction, and then the fully connected layer completes the vectorization representation of the extracted features. Finally, the feature vector is input to the softmax output layer to perform text classification according to accident types. The structure of the CNN text classification model designed in this paper is shown in Fig. 2.

Input Layer

This part of the content still belongs to the text preprocessing step, aimed at converting phrases in the corpus into machine-processable vector matrix form through word vector generation techniques, also known as the word embedding layer. Word2Vec and GloVe are two popular word embedding training models that can effectively identify various semantic and syntactic relationships, thus inputting the text corpus as continuous word vectors into the CNN model as the text matrix input. Word2Vec uses a shallow feedforward neural network for training to reflect the co-occurrence relationship between words and their neighbors, and is pretrained based on local corpus; GloVe uses logarithmic bilinear regression for training to reflect the overall co-occurrence relationship between words and is pretrained

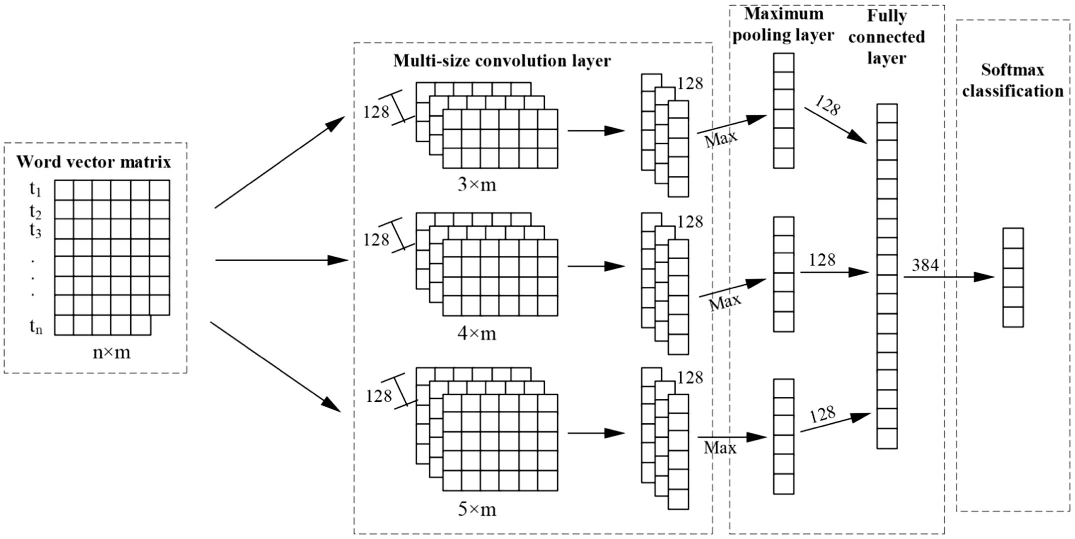

Fig. 2. Multiscale CNN text classification model.

based on the global corpus (Cerisara et al. 2018). Although the GloVe method can effectively utilize statistical information, it performs poorly in word analogy tasks, and some studies have shown that the word vectors obtained after pretraining the Word2Vec model perform better than those of the GloVe model (Khatua et al. 2019; Sunitha et al. 2022). Therefore, to ensure the robustness of the model, this study adopted the more mature Word2Vec model that has been verified by experiments to train word vectors. The specific steps are to compress the statistical information contained in the word context into a fixed-dimension, low-dimensional dense word vector, which is that if the maximum number of words in the text is n and the word vector dimension is m, then the input vector is $n \times m$, which realizes the numerical semantic representation of words. To enhance the model's generalization ability and improve its running speed, the idea of transfer learning can be introduced into the training process of the model. A large-scale unlabeled text corpus can be selected for pretraining of word vectors, and the pretrained word vectors can be used as the initial values of the word embedding layer.

Convolutional Layer

Convolutional layer, also known as the feature extraction layer, is the core component of CNN, with weight sharing and local connection features. This layer is responsible for most of the computations in the network, and by setting different sliding strides or different sizes of convolution kernels, it can extract features of different granularities from the input sequence. If the input of the convolution layer is a $d \times d$-dimension matrix, the convolution kernel size is $v \times v$, and $d \geq v$. The weight is ω, the output after convolution is c, and the calculation result at position (i, j) is as follows:

$$c_{i,j} = \sum_{k=1}^{v}\sum_{l=1}^{v} \omega_{k,l} x_{i+k-1,j+l-1}, \quad j \leq d - v + 1 \qquad (4)$$

where $x_{i,j}$ and $\omega_{i,j}$ = values respectively corresponding to the parameters of x, ω in position (i, j).

After performing convolutional operations on the features, an activation function is needed for nonlinear transformation, which makes the neural network structure better equipped to handle complex problems. Commonly used activation functions include sigmoid, tanh, and rectified linear unit (ReLU). However, after considering the characteristics and applicability of activation functions comprehensively, this study chose ReLU as the activation function because, compared with other nonlinear functions, ReLU has the ability to moderately sparsify and strongly generalize, making the extracted features more representative. When dealing with complex textual data, a multilayer, multisize convolutional neural network can capture features of different scales and semantic information at different levels by performing convolutional operations on convolutional kernels of different sizes. Therefore, this study designed and used multisize convolutional kernels with heights of 3, 4, and 5, with 128 convolutional kernels in each size to construct the convolutional layer, ultimately building a multilayer, multisize convolutional neural network (Wang et al. 2021).

Pooling Layer

The pooling layer is typically placed between two consecutive convolutional layers in a convolutional neural network and is used for dimensionality reduction and classification of text feature data extracted by the convolutional layer. When using the features extracted by the convolutional layer for training the classifier, an appropriate pooling method is adopted for the domain-specific feature points to downsample the feature map in the convolutional neural network, reducing the model parameters and computational complexity. Common pooling methods include max pooling and average pooling. Max pooling selects the maximum value in the feature map to preserve the main features of the image, while average pooling computes the average value in the feature map to normalize the entire image. However, average pooling can lead to information loss and text data usually exhibit strong position invariance; thus, max pooling has been widely used in text classification tasks. Finally, this study adopted the max pooling method for domain-specific feature points, selecting the maximum value of each feature information to extract and represent the convolutional features, which are used as inputs to the fully connected layer, further improving the performance of text classification.

Fully Connected Layer

Each dimension in the vector input to the fully connected layer represents the feature mapped by the text record. The fully connected layer connects all hidden units in the previous layer, which is equivalent to performing a linear weighted summation $Ax + b$ on all input units, where A represents the weight coefficient of the input unit, and b represents the bias value. The final generation vector of the fully connected layer is $f = (t_1, t_2, \ldots, t_n)$, where $t_1, t_2, \ldots, t_n$ are eigenvalues, and n is the dimension of f. The f is input into the softmax classifier for result classification, and the output result is an s-dimensional vector matrix, where the ith dimension represents the probability the sample is classified to the ith element. The category corresponding to the maximum value in the final matrix is the category output by the classifier. In this study, the expected output of the CNN model is five different accident types, so the softmax classifier will output a five-dimensional vector, where each element represents the probability of the text belonging to the corresponding accident type, and the accident type corresponding to the element with the highest probability in the final output vector will be considered as the prediction result of the CNN model.

Model Testing and Performance Evaluation

To evaluate the performance of the CNN model in text classification of construction texts, this study constructed three representative shallow learning comparison models, namely, SVM, NB, and LR. The parameters of the comparison models were set as the best state, and the same data set was used to compare the text classification model. There are two methods to evaluate the performance of classification models. One is using evaluation parameters, such as precision, recall, and $F1$-score, and the other is using the receiver operating characteristic (ROC) curve to visualize evaluation results.

Evaluation Parameters

The calculation announcements of the parameter precision, recall, and $F1$-score are as follows:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{6}$$

$$F1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

where TP = true positive, indicating the number of positive samples predicted to be true; FP = false positive, indicating the number of negative samples predicted to be true; and FN = false negative, indicating the number of positive samples predicted to be false.

In terms of indicator meaning, precision is a measure of how accurate the positive predictions are and recall is a measure of how many of the actual positives the model can identify. Additionally, precision and recall frequently exhibit a negative association, with an increase in one indicator sometimes being followed by a disproportionate decrease in the other (Cheng et al. 2020). The $F1$-score takes into account both precision and recall, and evaluates the overall performance of the classifier in the form of harmonic average. The larger the $F1$-score, the better the classification effect.

ROC Curve

ROC curve is an important metric tool in predictive analysis. After calculating the values of true positive rate (TPR) and false positive rate (FPR), a feature curve is formed with FPR as the x-axis and TPR as the y-axis. This calculation method takes into account the classifier's ability to classify positive and negative samples at the same time during the drawing process, eliminating the impact of sample class imbalance on the classifier, and making it easy to observe the classification accuracy from the graph. In addition, when the ROC curve does not clearly illustrate the classification performance, the area under the curve (AUC) metric is used to represent the area enclosed by the curve and the coordinate axis. The larger the area, the better the model performance.

Experiments and Results

Construction accident reports were gathered as thoroughly as possible for experiments to assess how well the suggested optimized text preprocessing session and the full text classification model based on deep learning perform. This part includes a detailed description of the experimental approach, including the data sources, text preprocessing, and experimental parameter settings, as well as a comparison and discussion of the four classifiers' final experimental findings.

Data Source

The data sets used in this study were mainly from the Ministry of Housing and Urban-Rural Development, the State Administration of Work Safety, the websites of various administrative departments, the safety management network, and the municipal government website (MOHURD 2021; SAOWS 2021; SMN 2021). First, Python language was used to crawl the accident data from 2013 to 2021 on the websites. Then, due to the scattered accident data, to ensure the accident data were as complete as possible, it was necessary to manually collect and supplement the safety production accident cases published on the website. Finally, 1,483 text reports of construction safety production accidents including five main accident categories were obtained, namely, 483 fall accidents, 266 collapse accidents, 145 electric shock accidents, 223 object strike accidents, and 366 mechanical injury accidents. These investigation reports were documents formed after the investigation of construction accidents by the housing and urban and rural construction departments and engineering construction units in the corresponding regions, which have considerable authority and authenticity. The reports mainly included the following five aspects: the overview of the accident occurrence units, the description of the accident process, the causes of the accident, the identification of the accident liability, and the preventive and corrective measures. This study focused on the precursory factors before the accident, and identified the basic and common injury precursors in the construction process from the accident process description module, excluding "collapse," "fall," "electric shock," and other words that indicate the outcome of the accident. For the 1,483 construction accident cases of five main accident categories in the database, each accident category was randomly divided into a training and testing database with a ratio of 8:2. In other words, 1,186 text reports were selected for training the CNN model, and 297 for testing its performance.

Text Preprocessing Process and Results Display

The process of text preprocessing was divided into two steps: one was to use the Jieba package in Python language for text corpus preprocessing, and the other was to use the gensim library to implement Word2Vec as a pretraining model to generate the word vector matrix of the embedding layer.

Text Segmentation Results Based on Phrase Extraction Technology

After cleaning and removing the stop words from the original text corpus, the TextRank algorithm and the phrase extraction technology based on mutual information and information entropy were used to segment words and extract keywords, respectively. The partial word segmentation results of the two comparison methods are given in Table 1.

Table 1. Comparison of word segmentation effects

Method	Partial display results for key words
TextRank	Left and right, work, proceed, ground, site, middle, personnel, discover, worker, install, suddenly, fall, responsible, platform, construction, operation, company, not, prepare
Phrase extraction	Operator, hanging cage, hit, truck crane, balance arm, construction hoist, concrete pouring, scaffold, unloading platform, bridge crane, hanging load, tower crane driver, operation platform, crane driver, directing cranes, arranging workers, moving scaffold, operation excavator, operation site

The preceding process compared two Chinese text segmentation methods: the TextRank algorithm and the phrase extraction technique. As shown in Table 1, the Chinese text segmentation results based on the TextRank algorithm have relatively less semantic integrity and information compared with those obtained by the phrase extraction technique, which mostly show up as bigram word strings. In contrast, the results obtained by the phrase extraction technique have more complete words, clearer semantics, and richer information, indicating the phrase extraction technique can better solve the problem of domain limitations and unknown words in the text to be processed, and can more accurately identify complex multicharacter words. Therefore, in practical applications, the Chinese text segmentation method based on the phrase extraction technique is more reliable and can provide better guarantees for subsequent feature extraction and text classification.

Using Pretraining Word Vector in the Embedding Layer

The pretrained word vectors for the segmented text were generated using the continuous bag-of-words (CBOW) approach of the Word2Vec model. When using this model, appropriate hyperparameters should be chosen based on factors such as data set size, task requirements, and text type to achieve better model performance and results. In this study, based on test results and references, the word vector dimension vector_size was set to 128, which provides relatively good expressive power while maintaining high efficiency. Additionally, the min_count filter size was set to 2 to filter out words that only appear once, retaining most common vocabulary, and improving model efficiency and effectiveness. In the embedding layer of the CNN, the pretrained Word2Vec model's word vectors and weights can be directly used to represent existing words in the corpus, and the resulting embedding layer can be used as the input layer of the CNN model to continue with subsequent text classification operations.

Hyperparameter Settings of CNN Model

This study ran in Python 3.9 software, and the experimental environment was mainly based on TensorFlow 2.6.0 version and Keras 2.7.0 version to construct a CNN text classification model. According to the test results and references, the hyperparameters in the model were selected as follows: the word vector dimension was set to 128. The convolution kernel was set to three different sizes, and the number of convolution kernels was 128, so the size of convolution kernels was 3 × 128, 4 × 128, and 5 × 128, respectively. The dropout rate was set to 0.5%, the gradient descent adopted the method of adaptive moment estimation (Adam), the learning rate was set to 0.5, and the number of iterations was set to 30. After setting the hyperparameters according to the preceding data, the performance of the obtained CNN model was relatively good, which supports subsequent operations of model testing and performance evaluation.

Classification Performance

Classification Results of the CNN Model

Based on the preceding text preprocessing workflow, it is possible to generate word vectors that represent the precursory factors of construction accidents. For example, for a specific accident description in a report, "The worker climbed over the railing onto the asbestos tiles to shovel material. In the process of shoveling, the asbestos tiles were stepped on and broke, resulting in a falling accident." After improved text preprocessing, the four precursory factors of climbed over the railing, asbestos tiles, shoveling, and broke can be extracted. After pretraining, the text-form precursory factors of accidents can be transformed into a vector mode that computers can directly recognize, and then used as input to the constructed model for result validation.

Although there have been many studies and applications of shallow learning methods in text classification tasks in the field of NLP, deep learning algorithms have been proven to be the most humanlike hierarchical intelligent learning method, and therefore become a key step in achieving machine learning intelligence. To more comprehensively and accurately evaluate the performance of deep learning algorithms in text classification tasks, a comparative analysis was conducted between a deep learning method based on CNN and three representative and important shallow learning methods: SVM, LR, and NB models. The parameters of the control models were adjusted to their optimal state to ensure fairness and accuracy of comparison, and the same data set was used for comparison of the performance of text classification models to reduce experimental errors and biases and evaluate the accuracy and efficiency of the CNN model in text classification tasks. Precision, recall, and $F1$-score (i.e., the harmonic mean of precision and recall) were used as the evaluation criteria for model performance.

Based on the data in Table 2, the test results of the four different classifiers are summarized as follows:

- According to the metrics used to measure the four classifiers' overall accuracy, the CNN model has the best accuracy (0.76),

Table 2. Results of text classification test

	CNN			SVM			LR			NB		
Accident type	$F1$-score	Precision	Recall	$F1$-score	Precision	Recall	$F1$-score	Precision	Recall	$F1$-score	Precision	Recall
Fall accidents (0)	0.85	0.8	0.91	0.79	0.83	0.75	0.72	0.94	0.58	0.62	0.45	0.97
Collapse (1)	0.85	0.85	0.84	0.82	0.80	0.85	0.77	0.69	0.87	0.58	1.00	0.41
Electric shock (2)	0.82	0.95	0.72	0.81	0.76	0.88	0.39	0.24	1.00	0.00	0.00	0.00
Object strike (3)	0.42	0.72	0.30	0.49	0.48	0.51	0.32	0.20	0.69	0.00	0.00	0.00
Mechanical injury (4)	0.72	0.64	0.82	0.67	0.67	0.67	0.69	0.71	0.68	0.63	0.67	0.60
Accuracy	0.76	—	—	0.73	—	—	0.66	—	—	0.54	—	—

followed by the SVM, LR, and NB models with 0.73, 0.66, and 0.54, respectively. In spite of the fact that the SVM model performs marginally better in the $F1$-score metric of the harmonic mean for the object strike accident type, the CNN model performed the best in $F1$-score for all other accident types when the classification of other accident types was performed. Overall, the CNN model generally has superior accuracy and stability for creating text classification models because it more accurately extracts detailed feature descriptions from complicated text than the other three shallow learning approaches. The test findings show once more that CNN models not only perform exceptionally well in the area of picture identification, but also have a lot of room for growth in the extraction and mining of text data accidents' antecedent information.

- The object strike accident type had the weakest performance out of the four classification models, and the classification results had the lowest $F1$-scores. However, two accident types, fall and collapse, performed better among the four classification models, and the index data obtained in the classification results were higher.

The main reasons for this result can be considered from the following three aspects: first, the quantity of text cases for different accident types. With text instances making up 32.8% of all cases, there are abundant training samples for fall accidents, which leads to successful training results in the four classification algorithms and high final performance assessment index values. The second is the degree of detail in the accident description. Although the number of text cases for collapse accidents accounts for only 16.8% of the total number of cases, the analysis of the descriptions and feature word extraction results in the collapse text reports shows that the accident descriptions in collapse accidents are more detailed and comprehensive regarding the construction environment, technical operations, and on-site management. Additionally, the extracted key feature words are more distinctive, such as "formwork support," "concrete leakage," and "no support," which can vividly represent the scene of the collapse accident. Therefore, the final model classification performance is also better. The third is the scene differentiation of the accident description. After examining the misclassification cases of object strike accident types that performed poorly, it was determined that the small number of cases (accounting for only 15.3% of all cases) was one reason for the low classification accuracy. Another more important reason was that there were many words related to high altitude and machinery in the event scene descriptions, and it was easy for the feature information extracted in the word embedding step to cause the classification model to misclassify them as falling from height or mechanical injury accident types. The test results have demonstrated that there are traceable indicators and trend information in different types of accidents based on the extracted feature factors or frequently occurring keyword combinations from the text accident reports. Examples of these key words and phrases include "temporary support during construction," "soil stability issues," and "material fatigue." These keywords and phrases may reveal specific tacit safety knowledge, such as the importance of temporary support during construction, the potential risks of soil stability issues, and the impact of material fatigue on the safety of building structures. This revealed safety knowledge can be practically applied in real-world construction scenarios, strengthening safety management practices. Construction managers can take appropriate safety management measures based on this tacit safety knowledge. By translating tacit safety knowledge into practical safety management practices, construction managers can have a more comprehensive understanding of and effectively address potential safety risks, thereby improving the level of safety management in construction scenarios.

The safety accident category of electric shock (10% of all cases) performs well in CNN and SVM classification models but poorly in LR and NB models. Model validation using the same accident safety reports eventually produced this phenomenon in agreement with the experimental findings of Tian et al. (2021), showing that the performance of the shallow learning algorithm was poorly trained in some of the classifiers, or more specifically, that the robustness of the classification models constructed by the shallow learning algorithm was insufficient. Furthermore, with 25.1% of all accident cases utilized in the study, the mechanical injuries safety accident category outperformed the other three classification methods generally, and could also be used to improve the classification accuracy by expanding the quantity of case reports. By automatically learning feature representations through multilayer and multiscale neural networks, the deep learning method no longer relies on manually defined features and rules, enabling a more comprehensive capture of information in complex and diverse construction safety texts. Comparing the results with traditional methods once again verifies that deep learning methods have significant advantages in discovering implicit knowledge compared with conventional approaches. Traditional methods rely on manually defined features and rules for text classification and analysis, but it is difficult for the manually defined features and rules to encompass all the information in complex and diverse construction safety texts. In contrast, deep learning methods have stronger data modeling and generalization capabilities, and they can extract and represent latent knowledge and patterns by learning complex patterns and correlations in large amounts of data. This allows for more accurate discovery of tacit knowledge in construction safety texts, including correlations between accident types, the importance of risk factors, and potential safety hazards in specific scenarios.

The combination of NLP and deep learning algorithms holds high practical and promotional value in construction safety management practices. It not only helps construction managers have a more comprehensive understanding of the nature and characteristics of accidents but also enables more accurate decision-making and the formulation of more effective preventive measures based on effectively extracted tacit knowledge. Therefore, the proposed deep learning method contributes to strengthening construction safety management practices and improving the effectiveness of accident

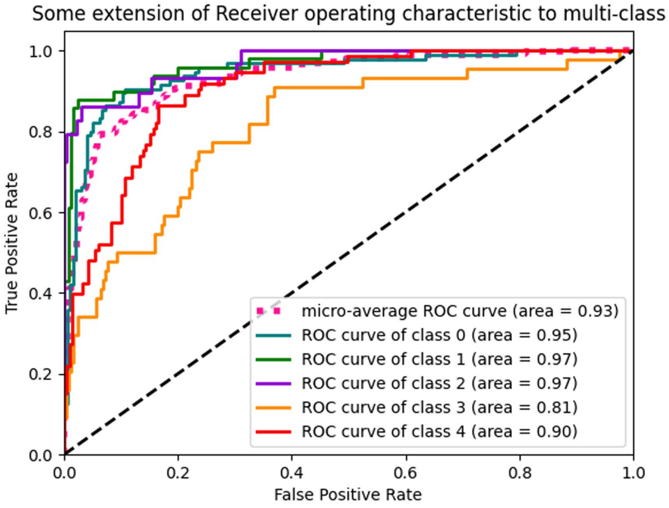

Fig. 3. ROC curve evaluation of text classification model for construction accident types.

prevention and management, and provides new directions and opportunities for future research and practices.

Evaluation of CNN Model Performance Using ROC Curve

The ROC curve and AUC index of the text classification model of construction accident type based on CNN algorithm are shown in Fig. 3. It can be seen from the figure that the area surrounded by the overall average ROC curve and the coordinate axis is 0.93. The recognition accuracy of Accident Types 0, 1, and 2 is relatively high, the recognition accuracy of Accident Type 4 is medium, and the recognition accuracy of Accident Type 3 is low. From the model effect, it can be seen that the text classification effect of construction accident types is generally good, and the model identification effect is closely related to the total number of cases of corresponding accident types.

Conclusions and Future Work

Due to excessive human intervention and a lack of information technology in the construction industry, the level of safety management has been progressing at an unusually slow pace, which is inconsistent with the goals of safety production and accident prevention. The field of safety production has a large number of accident reports stored in nonstructured or semistructured formats. These texts contain rich information that can provide in-depth support for safety management and risk prevention research. With the rapid development of computer technology and the effective improvement of information technology, text data containing more detailed and comprehensive knowledge are gradually being discovered. Although NLP in the field of construction site safety has not received sufficient research and emphasis, extracting tacit knowledge and information from accident text reports is of great significance for preventing accidents and reducing casualties in construction projects. However, manually organizing the increasing amount of data in databases is time-consuming and inefficient, making it difficult for managers to retrieve valuable information within the system. Therefore, it is necessary to utilize artificial intelligence to extract valuable text features and automatically classify text accident reports. Thus, in the field of NLP, a more complete and effective approach to text processing is proposed, optimizing the ability to automatically capture key feature information through multilayer learning methods, thereby achieving good results in text classification and providing more professional knowledge for subsequent industrial production processes in terms of work planning, process hazard analysis, and prompt post-incident response.

Previous studies have shown that shallow learning algorithms can achieve good classification results in text classification tasks in the construction industry. However, considering that deep learning algorithms possess stronger capabilities in data modeling, generalization, and automatic feature extraction, through training on large-scale text data and model building, they can capture nonexplicit associations and complex semantic relationships within the text, thereby acquiring deeper implicit knowledge. Therefore, introducing deep learning into text classification tasks in the construction domain holds the potential to comprehensively and accurately assist construction managers in identifying and assessing potential safety risks and in formulating safety decisions and preventive measures in a targeted manner, thus enhancing the level of safety management practices. Based on the preceding analysis, this paper proposed a deep learning algorithm framework based on CNN for the classification of accident types in construction site safety text reports. This framework thoroughly extracts and processes text feature information using multilayer and multiscale neural networks, enabling managers and operators to have a deeper understanding of the contextual elements of accidents. This not only allows for timely intervention to protect workers from harm but also enables accurate classification of different accident types, with the model achieving a classification accuracy of 76%. Therefore, the proposed text classification framework helps managers more effectively obtain valuable information from safety production accident text reports and promotes the digitization of safety production. This framework leverages the advantages of deep learning and achieves significant improvements in text feature extraction and modeling.

The main contributions of this paper are threefold:

1. It confirms the reliability of text analysis in predicting accident types. Through analyzing nonstructured accident texts containing valuable hidden information, premonitory signs and common trends can be obtained, which can achieve more accurate predictions of accident types, once again proving that the safety management of construction accidents can be quantitatively studied based on historical statistical data.
2. It improves the word segmentation process in text preprocessing. As the foundation of NLP tasks, the effect of text preprocessing can promote more accurate and reasonable processing of large amounts of collected data by automated means. To obtain more meaningful accident premonitory factors and weaken human intervention factors, a phrase extraction technique combining mutual information and information entropy was used to integrate advanced semantic and syntactic features in the text description without the need to construct and load professional domain dictionaries to obtain more complete and comprehensive text semantic segmentation results.
3. It validates the effectiveness of the constructed CNN model. Word2Vec technology was used to pretrain word vectors on the loaded corpus, and the pretraining results were placed in the embedding layer as the input layer of the CNN model. After reasonable settings of the hyperparameters in the constructed accident type text classification model, the final CNN model was superior to other shallow learning classification models in terms of average accuracy and $F1$-score. This indicates that the construction accident type text classification framework proposed in this paper, combining deep learning and natural language processing, can achieve accurate automatic text recognition and classification after improving text preprocessing and designing a complete CNN text classification process.

Despite the CNN model has shown promising results in the task of accident type text classification, there are still several areas for improvement in future research. First, this paper collected a total of 1,483 accident text reports, and for complex models like deep learning more representative and larger data samples can improve model performance. Therefore, future research should focus on creating high-quality massive databases to improve the accuracy of deep learning methods in text processing. Second, the NLP approach proposed in this paper is primarily focused on Chinese text data. Due to the differences in language structure, language rules, and lexical features between different languages, there are significant disparities between Chinese and English texts in aspects such as word segmentation, lemmatization, and stemming. Directly applying the NLP methods proposed to other languages may result in decreased recognition accuracy, which can subsequently affect the accuracy and generalization ability of the predictive models. Therefore, when applying these methods to other languages, it is necessary to make appropriate adjustments and optimizations based on the characteristics of the specific language. Additionally, it is crucial to thoroughly evaluate and validate their applicability in practical scenarios to ensure the reliability and accuracy of the research findings. Last, the focus of this paper is to experimentally demonstrate the significant advantages of the proposed framework, combining NLP and deep learning algorithms, in automatically acquiring implicit knowledge. However, there is still a lack of practical tool development to translate theoretical research into tangible applications. Future research tasks include further developing and implementing specific tools or systems that can receive construction accident reports or other related texts and automatically analyze and classify the information within them. Such tools would enable construction managers to quickly understand the types of accidents and potential safety risks, facilitating the adoption of appropriate prevention and management measures.

Data Availability Statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments

This work was supported by the Major Program of Philosophy and Social Science Research in Jiangsu University (Grant No. 2020SJZDA085), the China Postdoctoral Science Foundation (Grant No. 2019M662003), and the Jiangsu Planned Projects for Postdoctoral Research Funds (Grant No. 2019K171).

References

Cerisara, C., P. Kral, and L. Lenc. 2018. "On the effects of using word2vec representations in neural networks for dialogue act recognition." *Comput. Speech Lang.* 47 (Jan): 175–193. https://doi.org/10.1016/j.csl.2017.07.009.

Chen, C., J. B. Xi, J. P. Wang, and Y. Chen. 2021. "Mining of association rules for hidden safety hazards in hydropower project construction." *Chin. J. Saf. Sci.* 31 (8): 75–82. https://doi.org/10.16265/j.cnki.issn1003-3033.2021.08.011.

Cheng, M. Y., D. Kusoemo, and R. A. Gosno. 2020. "Text mining-based construction site accident classification using hybrid supervised machine learning." *Autom. Constr.* 118 (Oct): 103265. https://doi.org/10.1016/j.autcon.2020.103265.

Chi, N. W., K. Y. Lin, and S. H. Hsieh. 2014. "Using ontology-based text classification to assist job hazard analysis." *Adv. Eng. Inf.* 28 (4): 381–394. https://doi.org/10.1016/j.aei.2014.05.001.

Das, S., A. Mudgal, A. Dutta, and S. R. Geedipally. 2018. "Vehicle consumer complaint reports involving severe incidents: Mining large contingency tables." *Transp. Res. Rec.* 2672 (32): 72–82. https://doi.org/10.1177/0361198118788464.

Fang, W., L. Ding, P. E. Love, and C. Zhou. 2020. "Computer vision applications in construction safety assurance." *Autom. Constr.* 110 (Feb): 103013. https://doi.org/10.1016/j.autcon.2019.103013.

Goh, Y. M., and C. U. Ubeynarayana. 2017. "Construction accident narrative classification: An evaluation of text mining techniques." *Accid. Anal. Prev.* 108 (Nov): 122–130. https://doi.org/10.1016/j.aap.2017.08.026.

Graves, J. M., J. M. Whitehill, B. E. Hagel, and F. P. Rivara. 2015. "Making the most of injury surveillance data: Using narrative text to identify exposure information in case-control studies." *Injury* 46 (5): 891–897. https://doi.org/10.1016/j.injury.2014.11.012.

Guo, B. H., Y. Zou, Y. Fang, Y. M. Goh, and P. X. Zou. 2021. "Computer vision technologies for safety science and management in construction: A critical review and future research directions." *Saf. Sci.* 135 (Mar): 105130. https://doi.org/10.1016/j.ssci.2020.105130.

Gupta, A. K., C. G. Pardheev, S. Choudhuri, S. Das, and A. Garg. 2022. "A novel classification approach based on context connotative network (CCNet): A case of construction site accidents." *Expert Syst. Appl.* 202 (15): 117281. https://doi.org/10.1016/j.eswa.2022.117281.

Hinton, G. E., S. Osindero, and Y. W. Teh. 2006. "A fast learning algorithm for deep belief nets." *Neural Comput.* 18 (7): 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527.

Hobson, L., H. Cole, and H. Hannes. 2019. *Natural language processing in action*. Greenwich, UK: Manning Publications.

Hughes, P., D. Shipp, M. Figueres-Esteban, and C. Van Gulijk. 2018. "From free-text to structured safety management: Introduction of a semi-automated classification method of railway hazard reports to elements on a bow-tie diagram." *Saf. Sci.* 110 (Dec): 11–19. https://doi.org/10.1016/j.ssci.2018.03.011.

International Labour Organization. 2021. "Safety and health at work." Accessed December 23, 2021. http://www.ilo.org/global/topics/safety-and-health-at-work/lang--en/index.html.

Jiang, Y. Y., B. Jin, and B. C. Zhang. 2021. "Research progress of deep learning in the field of natural language processing." [In Chinese.] *Comput. Eng. Appl.* 57 (22): 1–14. https://doi.org/10.3778/j.issn.1002-8331.2106-0166.

Jin, Z., Y. Han, and Q. Zhu. 2018. "A sentiment analysis model with the combination of deep learning and ensemble learning." [In Chinese.] *J. Harbin Inst. Technol.* 50 (11): 32–39. https://doi.org/10.11918/j.issn.0367-6234.201709078.

Jing, S., X. Liu, X. Gong, Y. Tang, G. Xiong, and S. Liu. 2022. "Correlation analysis and text classification of chemical accident cases based on word embedding." *Process Saf. Environ. Prot.* 158 (Feb): 698–710. https://doi.org/10.1016/j.psep.2021.12.038.

Jung, N., and G. Lee. 2019. "Automated classification of building information modeling (BIM) case studies by BIM use based on natural language processing (NLP) and unsupervised learning." *Adv. Eng. Inf.* 41 (Aug): 100917. https://doi.org/10.1016/j.aei.2019.04.007.

Kakhki, F. D., S. A. Freeman, and G. A. Mosher. 2019. "Evaluating machine learning performance in predicting injury severity in agribusiness industries." *Saf. Sci.* 117 (Aug): 257–262. https://doi.org/10.1016/j.ssci.2019.04.026.

Khatua, A., A. Khatua, and E. Cambria. 2019. "A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks." *Inf. Process. Manage.* 56 (1): 247–257. https://doi.org/10.1016/j.ipm.2018.10.010.

Khurana, D., A. Koli, K. Khatter, and S. Singh. 2023. "Natural language processing: State of the art, current trends and challenges." *Multimedia Tools Appl.* 82 (3): 3713–3744. https://doi.org/10.1007/s11042-022-13428-4.

Ko, T., H. D. Jeong, and G. Lee. 2021. "Natural language processing–driven model to extract contract change reasons and altered work items for advanced retrieval of change orders." *J. Constr. Eng. Manage.* 147 (11): 04021147. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002172.

Lette, A., A. Ambelu, T. Getahun, and S. Mekonen. 2018. "A survey of work-related injuries among building construction workers in southwestern Ethiopia." *Int. J. Ind. Ergon.* 68 (Nov): 57–64. https://doi.org/10.1016/j.ergon.2018.06.010.

Li, S., H. Cai, and V. R. Kamat. 2016. "Integrating natural language processing and spatial reasoning for utility compliance checking." *J. Constr. Eng. Manage.* 142 (12): 4016074.1. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001199.

Love, P. E., S. Jim, and T. Pauline. 2018. "Putting into practice error management theory: Unlearning and learning to manage action errors in construction." *Appl. Ergon.* 69 (May): 104–111. https://doi.org/10.1016/j.apergo.2018.01.007.

Luo, X. J., L. O. Oyedele, A. O. Ajayi, and O. O. Akinade. 2020. "Comparative study of machine learning-based multi-objective prediction framework for multiple building energy loads." *Sustainable Cities Soc.* 61 (Oct): 102283. https://doi.org/10.1016/j.scs.2020.102283.

MOHURD (Ministry of Housing and Urban-Rural Development). 2021. "A guide for planning and construction of rural and urban areas in China." Accessed September 20, 2021. https://www.mohurd.gov.cn/ess/.

Pan, Y., and L. Zhang. 2021. "Roles of artificial intelligence in construction engineering and management: A critical review and future trends."

Autom. Constr. 122 (Feb): 103517. https://doi.org/10.1016/j.autcon.2020.103517.

Peng, H., F. Long, and C. Ding. 2005. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (8): 1226–1238. https://doi.org/10.1109/TPAMI.2005.159.

SAOWS (State Administration of Work Safety). 2021. "A non-ministerial agency for the regulation of risks to occupational safety and health in China." Accessed September 20, 2021. https://www.mem.gov.cn/was5/web/.

SMN (Safety Management Network). 2021. "A non-ministerial agency for the regulation of risks to occupational safety and health in China." Accessed September 22, 2021. https://www.safehoo.com/Manage/.

Soliman, E. 2018. "Risk identification for building maintenance projects." *Int. J. Constr. Project Manage.* 10 (1): 37–54.

Song, B., and Y. Suh. 2019. "Narrative texts-based anomaly detection using accident report documents: The case of chemical process safety." *J. Loss Prev. Process Ind.* 57 (Jan): 47–54. https://doi.org/10.1016/j.jlp.2018.08.010.

Sun, Y. L., W. Yu, Z. Han, and K. R. Liu. 2006. "Information theoretic framework of trust modeling and evaluation for ad hoc networks." *IEEE J. Sel. Areas Commun.* 24 (2): 305–317. https://doi.org/10.1109/JSAC.2005.861389.

Sunitha, D., R. K. Patra, N. V. Babu, A. Suresh, and S. C. Gupta. 2022. "Twitter sentiment analysis using ensemble based deep learning model towards COVID-19 in India and European countries." *Pattern Recognit. Lett.* 158 (Jun): 164–170. https://doi.org/10.1016/j.patrec.2022.04.027.

Tanguy, L., N. Tulechki, A. Urieli, E. Hermann, and C. Raynal. 2016. "Natural language processing for aviation safety reports: From classification to interactive analysis." *Comput. Ind.* 78 (May): 80–95. https://doi.org/10.1016/j.compind.2015.09.005.

Tian, D., M. Li, J. Shi, Y. Shen, and S. Han. 2021. "On-site text classification and knowledge mining for large-scale projects construction by integrated intelligent approach." *Adv. Eng. Inf.* 49 (Aug): 101355. https://doi.org/10.1016/j.aei.2021.101355.

Tixier, A. J. P., M. R. Hallowell, B. Rajagopalan, and D. Bowman. 2016. "Application of machine learning to construction injury prediction." *Autom. Constr.* 69 (Sep): 102–114. https://doi.org/10.1016/j.autcon.2016.05.016.

Wang, B., Y. Lei, N. Li, and W. Wang. 2021. "Multi-scale convolutional attention network for predicting remaining useful life of machinery." *IEEE Trans. Ind. Electron.* 68 (8): 7496–7504. https://doi.org/10.1109/TIE.2020.3003649.

Wang, S. H. 2021. "Basic concept analysis of dual prevention mechanism and discussion on creation method." [In Chinese.] *Ind. Saf. Environ. Prot.* 47 (3): 63–67. https://doi.org/10.3969/j.issn.1001-425X.2021.03.015.

Wu, F., A. Fan, A. Baevski, Y. N. Dauphin, and M. Auli. 2019. "Pay less attention with lightweight and dynamic convolutions." Preprint, submitted July 25, 2019. http://arxiv.org/abs/1901.10430.

Xing, X., B. Zhong, H. Luo, H. Li, and H. Wu. 2019. "Ontology for safety risk identification in metro construction." *Comput. Ind.* 109 (Aug): 14–30. https://doi.org/10.1016/j.compind.2019.04.001.

Xu, X. X., and X. Z. Patrick. 2021. "Discovery of new safety knowledge from mining large injury dataset in construction." *Saf. Sci.* 144 (Dec): 105481. https://doi.org/10.1016/j.ssci.2021.105481.

Yao, Q., R. Li, L. Song, and M. Crabbe. 2021. "Construction safety knowledge sharing on Twitter: A social network analysis." *Saf. Sci.* 143 (Nov): 105411. https://doi.org/10.1016/j.ssci.2021.105411.

Zhang, F., H. Fleyeh, X. Wang, and M. Lu. 2019. "Construction site accident analysis using text mining and natural language processing techniques." *Autom. Constr.* 99 (Mar): 238–248. https://doi.org/10.1016/j.autcon.2018.12.016.

Zhang, X., P. Srinivasan, and S. Mahadevan. 2021. "Sequential deep learning from NTSB reports for aviation safety prognosis." *Saf. Sci.* 142 (Oct): 105390. https://doi.org/10.1016/j.ssci.2021.105390.

Zhong, B., X. Pan, P. E. Love, L. Ding, and W. Fang. 2020. "Deep learning and network analysis: Classifying and visualizing accident narratives in construction." *Autom. Constr.* 113 (May): 103089. https://doi.org/10.1016/j.autcon.2020.103089.

Zhou, P., and N. El-Gohary. 2017. "Ontology-based automated information extraction from building energy conservation codes." *Autom. Constr.* 74 (Feb): 103–117. https://doi.org/10.1016/j.autcon.2016.09.004.

Zhu, Y., and S. P. Chen. 2020. "Text classification models with nearest neighbor attention and convolutional neural networks." [In Chinese.] *Small Microcomputer Syst.* 41 (2): 375–380. https://doi.org/10.3969/j.issn.1000-1220.2020.02.025.