Text Extraction and Correlation Analysis of Multi-Factor Mechanisms Influencing Traffic Accident Severity in Expressway Reconstruction and Expansion Projects

Jingshi Li[1], Zhongguang Wu*[2], Zechao Huang[3], Jiatian Hao[2], Yuanbo Zhang[1], Ying Li[1]

[1]Liaoning Provincial Expressway Operation Management Co., Ltd., Shenyang 110055, China;
[2]Research Center for Standards and Metrology, China Academy of Transportation Sciences, Beijing, 100029, China;
[3]Liaoning Provincial Transportation Construction Management Co., Ltd., Shenyang 110005, China;

Jingshi Li Email:15084113111@163.com
Zhongguang Wu Email:kinliwu@163.com
Zechao Huang Email:15040184766@163.com
Jiatian Hao Email:13020059227@163.com
Yuanbo Zhang Email:
Ying Li Email: 13804183203@163.com

Abstract. To accurately extract key factors from unstructured traffic accident texts and identify the interaction mechanisms among factors affecting the severity of accidents during such projects, a combined deep learning model based on BERT-BiLSTM-CRF-WApriori model is proposed. It blends Bi-directional Encoder Representation from Transformers (BERT), Bi-directional Long Short-Term Memory (BiLSTM), Conditional Random Field (CRF), and weighted Apriori association rule algorithm (WApriori),Using the BERT-BiLSTM-CRF-WApriori model, complex accident factors—including road factors, traffic factors, construction factors, environmental factors, driver factors and other complex factors in expressway reconstruction and expansion, are extracted from unstructured traffic accident texts related to expressway reconstruction and expansion projects, facilitating the analysis of interaction mechanisms influencing accident severity. The results show that the severity of sideswipe accidents in construction areas is relatively low, over speed vehicles in the lane closed construction area and involving truck will increase the severity of the accident; collisions between truck and fixture crash or intruding into the work area are likely to cause serious accidents. These research results can inform targeted measures for controlling traffic accidents and reducing injuries during expressway reconstruction and expansion project.

Keywords: traffic engineering; natural language processing; weighted correlation rules; expressway reconstruction and expansion; traffic accident severity; traffic accident causation

G. Chen et al. (eds.), *Proceedings of the 2024 International Conference on Rail Transit and Transportation (ICRTT 2024)*, Advances in Engineering Research 254,
https://doi.org/10.2991/978-94-6463-610-9_32

1 INTRODUCTION

In recent years, the number of expressway reconstruction and expansion projects has been increasing. The traffic organization form of construction areas often adopts the mode of simultaneous traffic and construction, where vehicles and construction personnel, machinery, etc., share the same horizontal plane, creating a dangerous and complex traffic scenario. As deemed by many researchers[1-3], such accidents involving vehicles breaking into construction areas occur frequently, and the severity of the accidents is much higher than that of normal sections. Analyzing the complex relationships between the factors causing traffic accidents and their severity during the reconstruction period is the basis and prerequisite for formulating measures to improve traffic safety and ensure the safety of construction and operation.

Research on factors affecting the severity of traffic accidents is a key focus in the field of traffic safety, with traffic accident reports serving as the primary data source for safety analysis. With the advancement of big data technologies, natural language processing (NLP) has matured and achieved widespread application across various domains. Arteaga et al. [4]combined machine learning-based text mining with GCV-LIME to analyze heavy vehicle crash data from 2007 to 2017 in Queensland, identifying possible causal factors influencing injury severity. Xiao et al. [5]developed a C-BiLSTM neural network by integrating Convolutional Neural Networks (CNN) and BiLSTM to predict the duration of traffic accidents. Sayed et al. [6]develop a classifier that applies text mining techniques to quickly find missed work zone crashes through the unstructured text saved in the crash narratives. Kwayu et al. [7]used the structural topic modeling (STM) and network topology analysis to generate and examine the prevalence and interaction of themes from the crash narratives that were mainly categorized into pre-crash events, crash locations and involved parties in the traffic crashes.

After extracting factors affecting accident severity using NLP techniques, statistical regression models and association analysis are commonly employed for accident factor identification and analysis. Hu et al.[8] analyzed five significant factors affecting the severity of highway traffic accidents by constructing a spatial generalized ordered Probit model. Ahmed et al. [9] used the binary logit model to establish the risk probability model of traffic safety accidents related to heavy vehicles in highway construction areas. Zhang et al. [10] constructed a risk probability assessment model based on multiple linear regression for the renovation and expansion work area, and analyzed the influence of the proportion of large vehicles and the length of the transition segment on the traffic safety risk in the work area. The above methods all consider the influence of some factors on the accident, but ignore the potential correlation among factors. Correlation analysis methods such as Apriori are effective methods to mine the internal correlation among data factors. Weng et al. [11] applied the Apriori association rule algorithm to mine and analyze the correlation between traffic accident factors in the construction area of expressway reconstruction and expansion. Yang et al. [12] improved the Apriori association rule mining algorithm based on analytic hierarchy Process (AHP) to improve the efficiency of association rule mining and highlight the main rules, and analyzed the factors related to the collision risk of expressway vehicles in different regional types and the internal correlation of each factor. Yuan et al. [13] applied the

Apriori association rule mining algorithm with orientation constraint to mine highway traffic accident data on working days and rest days respectively, and improved the algorithm to reduce the generation of redundant association rules.

Scholars both domestically and internationally have utilized natural language processing techniques to extract highway traffic accident data, conducting research on accident factor identification and analysis. However, the extraction of key information from texts still largely relies on rule-based feature extraction methods, which suffer from issues such as inaccurate semantic understanding and weak contextual dependency. These methods often require repeated revision and refinement of rules to achieve satisfactory extraction results. In terms of accident analysis, studies on the impact of combined accident factors on accident severity are relatively scarce. Moreover, due to the varying operational environments of expressways, the mechanisms affecting accidents during regular operation periods differ from those during reconstruction and expansion phases. This makes it difficult to apply traditional accident causation analysis results, derived from standard driving scenarios, to the context of expressway reconstruction and expansion projects. Therefore, this paper constructs a combined deep learning model based on BERT-BiLSTM-CRF to extract key factors from unstructured traffic accident texts, calculate the weight of accident factors by using Fuzzy Analytic Hierarchy Process (FAHP) algorithm, improve Apriori association rule algorithm, analyze the distribution characteristics of traffic accident factors and excavate the multi-factor interaction mechanism. To explore and quantify the impact of road reconstruction and expansion, traffic, construction, environment and driver factors on traffic accident injuries. It provides a basis for the engineering application of traffic accident control and injury prevention measures during expressway reconstruction and expansion.

2 METHODOLOGY

2.1 BERT-BiLSTM-CRF Text Extraction Algorithm

To effectively extract information from traffic accident text data, a deep learning model is developed that combines BERT with BiLSTM-CRF. This model utilizes BERT for text feature extraction and integrates BiLSTM and CRF for Named Entity Recognition and information extraction, making it suitable for extracting key factors from complex traffic accident texts.

BERT Embedding Layer. The BERT model is utilized to encode input traffic accident texts, extracting contextual semantic information. During its pre-training phase, BERT learns rich linguistic representations through a bidirectional Transformer structure, enabling it to effectively capture subtle semantic differences in Chinese texts. By inputting traffic accident texts into a pre-trained BERT model, contextual embeddings for each character or word are obtained.

BiLSTM-CRF Layer. The BiLSTM model, combined with a CRF layer, is employed to extract key information from unstructured narrative traffic accident texts. The BiLSTM model consists of two LSTM models that process information in both forward and backward directions. The hidden layers for forward and backward propagation are represented in equation (1~2).

$$\overrightarrow{h_t} = \sigma(W_{hx}x_t + W_{hh}\overrightarrow{h_{t-1}} + b_h) \tag{1}$$

$$\overleftarrow{h_t} = \sigma(W'_{hx}x_t + W'_{hh}\overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}) \tag{2}$$

Where $\vec{h}_{t-1}$ represents the hidden state from the previous time step, x_t is the current input value, W_{hx} is the weight matrix for the forward input layer, W_{hh} is the weight matrix for the forward hidden layer, $b_{\overrightarrow{h}}$ is the forward bias vector, σ denotes the activation function, W'_{hx} is the weight matrix for the backward input layer, W'_{hh} is the weight matrix for the backward hidden layer, $b_{\overleftarrow{h}}$ is the backward bias vector.

The parallel neurons merge the hidden state values into the output layer, yielding the output value h_t, as shown in Equation (3).

$$h_t = \vec{W}_{ho}\vec{h}_t + \vec{W}_{ho}\vec{h}_t + b_l \tag{3}$$

Where $\overrightarrow{W}_{ho}$ represents the weight matrix for the forward output layer, $\overleftarrow{W}_{ho}$ is the weight matrix for the backward output layer, and b_l is the bias vector for the output layer.

CRF is the final layer in the BERT-BiLSTM-CRF model, used to globally label a sequence. The eigenvector trained by BiLSTM layer is taken as the input of CRF layer. Given an input sequence $H = [h_1, h_2, ...h_n]$ from BiLSTM, CRF computes the conditional probability of the label sequence $y = [y_1, y_2, ..., y_n]$. CRF conditional probability formula as shown in equation (4).

$$P(y \mid H) = \frac{exp(score(H, y))}{\sum_{y' \in y} exp(score(H, y'))} \tag{4}$$

Where y' is a possible label sequence from the set of all label sequences, $score(H, y)$ is the scoring function for the hidden representation H and the label sequence y.

The scoring function consists of two parts: the association score between labels and hidden outputs and the transition score between adjacent labels, as shown in equation (5).

$$score(H, y) = \sum_{i=1}^{n} W_{y_i} \cdot h_i + \sum_{i=1}^{n} A_{y_{i-1}, y_i} \tag{5}$$

Where $W_{y_i} \cdot h_i$ is the association score between label y_i and the hidden representation h_i from BiLSTM, W_{y_i} is the weight vector, A_{y_{i-1}, y_i} is the transition score, capturing the cost (or reward) of transitioning from label y_{i-1} to label y_i.

2.2 Weighted Association Rule Algorithm

Weight Algorithm. The traditional association rule mining algorithm overlooks the importance of items in the dataset and is inefficient for large datasets. The Fuzzy Analytic Hierarchy Process (FAHP) method addresses the issue of fuzziness in factor importance division during weight allocation. The steps to calculate the weights of each factor using the FAHP method are as follows:

(1) Constructing the Judgment Matrix

A pairwise comparison judgment matrix is constructed using triangular fuzzy numbers, as shown in equation (6). Each element in the judgment matrix represents the fuzzy comparison of the relative importance between two indicators.

$$\tilde{\mathbf{A}} = \begin{pmatrix} \tilde{a}_{11} & \tilde{a}_{12} & \cdots & \tilde{a}_{1j} \\ \tilde{a}_{21} & \tilde{a}_{22} & \cdots & \tilde{a}_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{a}_{i1} & \tilde{a}_{i2} & \cdots & \tilde{a}_{ij} \end{pmatrix} \tag{6}$$

Where $\tilde{a}_{ij}$ represents the fuzzy comparison of the importance of indicator A_i relative to indicator A_j, expressed as a triplet $\tilde{a}_{ij} = (l_{ij}, m_{ij}, u_{ij})$, where l_{ij} is the fuzzy lower bound, m_{ij} is the fuzzy middle value, and u_{ij} is the fuzzy upper bound.

(2) Calculating Fuzzy Weights

The fuzzy judgment matrix is transformed into fuzzy weights. For each indicator A_i, the fuzzy weight $\tilde{W}_i$ is calculated using the geometric mean method as shown in equation (7).

$$\tilde{W}_i = \left(\frac{1}{P} \sum_{j=1}^{n} \tilde{a}_{ij} \right) \tag{7}$$

Where P is the number of indicators.

(3) Defuzzification and Normalization

The fuzzy weight vectors are defuzzified using the centroid method to obtain the determined weights W_i for individual indicators, as shown in equation (8).

$$W_i = \frac{l_i + m_i + u_i}{3} \tag{8}$$

where, l_i、m_i and u_i are the fuzzy lower bound, middle value, and upper bound weights of indicator A_i, respectively. The determined weights W_i are then normalized, as shown in equation (9), to obtain the normalized weights $W_i^{'}$.

$$W_i^{'} = \frac{W_i}{\sum_{i=1}^{n} W_i} \tag{9}$$

Weighted Association Rule Algorithm. The Apriori association rule mining algorithm is a classical unsupervised learning technique used to extract knowledge describing the interrelationships between data items from large datasets. Its basic principle is to define a transaction set as a set of itemsets. In the traffic accident dataset, itemsets contain various factors that might be related to accidents, and transactions represent individual accident records in the database. An association rule is represented as "A→B", where A is the antecedent (the condition leading to the conclusion) and B is the consequent (the result caused by the antecedent). However, the traditional Apriori algorithm does not adequately consider the varying importance of different causative factors in traffic accidents when mining association rules. To this end, the weighted association rule algorithm is established by introducing the factor weights calculated by fuzzy analytic hierarchy process (AHP) above. The weight factors are introduced in the generation stage of frequent itemsets, and the three indexes of weighted support, weighted confidence and lift of each item set are calculated when generating candidate itemsets. Weighted support ($WSupport$) represents the weighted probability of itemsets A and B occurring together in all transactions, as shown in Equation (10).

$$WSupport(A \to B) = \frac{\sum_{t \in T} \left(\prod_{i \in A \cup B} W_i^{'} \right)}{count(Total)} \tag{10}$$

Where T represents the transaction set containing all transactions t; $count(Total)$ denotes the total number of transactions in the transaction database.

Weighted confidence ($WConfidence$) represents the weighted conditional probability of the consequent occurring given the antecedent, as shown in Equation (11).

$$WConfidence(A \rightarrow B) = \frac{WSupport(A \cup B)}{WSupport(A)} \tag{11}$$

Lift represents the ratio of the weighted confidence of the rule to the weighted support of the consequent. If the lift value greater than 1 indicates that there is a positive correlation between the condition and the conclusion, as shown in Equation (12).

$$Lift(A \rightarrow B) = \frac{WSupport(A \cup B)}{WSupport(A) * WSupport(B)} \tag{12}$$

3 RESULTS AND DISCUSSIONS

3.1 Accident Factors Extraction

The BERT-BiLSTM-CRF text extraction algorithm established in this paper was used to extract the attribute factors of the accident from the collected accident reports, and the variables were classified into five categories: road, traffic, construction, environment and driver factors based on the meaning and numerical types of the accident factors, a total of 7865 traffic accident data of expressway reconstruction and expansion were obtained. The division of data variables is shown in Table 1. After data processing, the total number of data is 22,482.

Table 1. Classification of Traffic Accident Data for Expressway Reconstruction

Classes	Variable Name	Subclass
Road Factors	Number of Lanes	1
		2
		3
		4
		＞4
	Roadway surface conditions	Dry
		Wet
		Icy, Snowy, etc.
Traffic Factors	Crash Type	Rear-End
		Sideswipe
		Fixture Crash
		Head-On
		Other
	Speed Limit at Crash Site	＜40

		40-60
		＞60
	Number of Vehicles Involved	1
		2
		3
		＞3
	Truck Involved or Not	Not Involved
		Involved
Construction Factors	Construction Activity	Partial lane closure
		Lane Shift
		Work on Median
		Other
	Crash Location	Between First and Last Sign
		Before First Warning Sign
		No Warning Signs
Environmental Factors	Weather Conditions	Clear
		Cloudy
		Rain
		Snow
		Wind
		Fog
	Lighting Conditions	Daylight
		Dark
		Dusk
Driver Factors	Age	18~25
		26~60
		＞60
	Sex	Male
		Female
	Hazardous Actions	Careless Driving
		Improper Operation
		Speeding
		Other
		None

3.2 Multi-Factor Interaction Mechanism Analysis

In this paper, the established WApriori algorithm is used for accident factor coupling analysis, Set the minimum support at >0.05 and the minimum confidence at >0.55, meaning the probability of the rule's occurrence is greater than 5%, and the confidence level of its correlation with the accident severity is greater than 50%. Finally filtered out 2,928 strong association rules that meet these criteria. The mined strong association rules were categorized according to accident severity, and the top three rules with the highest lift were extracted after removing duplicate rules, as shown in Fig 1.

Figure 1 presents a bubble plot that visualizes the association rules generated by the WApriori algorithm. In this plot, the size of each bubble corresponds to the support value, with larger bubbles indicating higher support, which reflects the frequency with which a rule occurs in the dataset. The color intensity of the bubbles represents the lift value, where darker shades of orange denote higher lift values, signifying a stronger association between the antecedent and the consequent. The x-axis displays the consequents derived from the rules, each linked to its corresponding antecedents, while the y-axis lists the antecedents, with the rule numbers indicated before each antecedent for clarity.

Fig. 1. Bubble Plot of Association Rules

As shown in Figure 1, the overall lift values of the rules are relatively high, indicating a strong correlation and credibility between the antecedent and the consequent of the rules. Further, the similarities and differences among the causative factors of accidents of varying severity were analyzed, leading to the following interpretations:

(1) Minor Accidents

Rule 1 shows that when a hazardous action such as improper operation occurs in a sideswipe accident involving two vehicles, the likelihood of it resulting in a minor accident is notably high, with a lift value of 2.59, indicating a strong correlation. Rule 2 reveals that even on wet roads, under good lighting conditions, sideswipe accidents have a high likelihood of being minor, with a lift of 2.42. Rule 3 points to partial lane closures in construction zones where female drivers are involved, showing a lift of 2.12, suggesting a moderate correlation with minor accidents. Minor accidents tend to occur under conditions where road surface quality is compromised, or driver errors, such as improper operation, are prevalent. Despite such adverse factors, the overall accident severity remains low.

(2) General Accidents

Rule 4 highlights that in work zones with speed limits of 40-60 km/h, speeding during partial lane closures leads to general accidents with a lift of 2.56, signaling a strong association. Rule 5 indicates that accidents involving at least one truck on dry roads have a notable lift of 2.10, correlating strongly with general accidents. Rule 6 further supports the notion that work zones with speed limits of 40-60 km/h tend to produce general accidents when two vehicles are involved, with a lift of 1.89. For general accidents, the presence of trucks and speeding behavior are key contributing factors, even under favorable road conditions, which increases the likelihood of injury-causing accidents.

(3) Severe Accidents

Rule 7 underscores the impact of speeding in work zones with a speed limit of 40-60 km/h, leading to severe accidents with a lift of 1.98, indicating a significant relationship. Rule 8 demonstrates that crashes between a single truck and a fixed object in the same work zone yield a lift of 1.84, associating strongly with severe outcomes. Rule 9 suggests that rear-end collisions involving multiple vehicles, including trucks, in dark conditions have a moderate lift of 1.79, indicating a considerable likelihood of severe accidents. Severe accidents are predominantly influenced by speeding in work zones, particularly when trucks are involved, and the severity increases under low-light or nighttime conditions. The risk of severe accidents escalates when multiple vehicles are involved, or when trucks collide with stationary objects.

4 CONCLUSIONS

In this study, a combined deep learning model based on BERT-BiLSTM-CRF-WApriori is developed for extracting key factors from traffic accident texts. Additionally, a weighted association rule algorithm, improved using the FAHP, is applied to analyze the factors influencing the severity of traffic accidents during highway reconstruction and expansion. The conclusions are as follows:

(1)The analysis of multi-factor interaction mechanisms indicates that sideswipe accidents involving two vehicles in work zones have relatively low severity. However,

factors such as speeding, involvement of trucks, collisions with fixture or rear-end collisions tend to increase accident severity. The severity is further heightened when these factors are coupled with poor lighting conditions and involvement of multiple vehicles.

(2)In the strong association rules for accident severity on expressway reconstruction and expansion projects, rules with high confidence are predominant. These rules indicate a high conditional probability of the consequent occurring when the antecedent item sets occur. Breaking the associations of antecedent rules, especially those corresponding to severe accident types, can help mitigate the severity of traffic accidents.

(3) This study is limited by the richness and completeness of the traffic accident dataset, with the extracted accident factor indicators being constrained, as traffic flow data such as traffic volume and vehicle speed are not included. Consequently, the research findings have certain limitations. In the future, a more comprehensive and accurate accident data set is needed to support a more thorough coupling analysis of accident factors.

REFERENCES

1. Khattak, A.J., Khattak, A.J., Council, F.M. (2002) Effects of work zone presence on injury and non-injury crashes. Accident Analysis & Prevention, 34(1): 19-29. https://doi.org/10.1016/s0001-4575(00)00099-3.
2. Yang, H., Ozbay, K., Ozturk, O., et al. (2015) Work zone safety analysis and modeling: a state-of-the-art review. Traffic Injury Prevention, 16(4): 387-396. https://doi.org/10.1080/15389588.2014.948615.
3. Zou, H., Zhu, S., Jiang, R., et al. (2023) Traffic conflicts in the lane-switching sections at highway reconstruction zones. Journal of Safety Research, 84: 280-289. https://doi.org/10.1016/j.jsr.2022.11.004.
4. Arteaga, C., Paz, A., Park, J.W. (2020) Injury severity on traffic crashes: A text mining with an interpretable machine-learning approach. Safety Science, 132: 104988. https://doi.org/10.1016/j.ssci.2020.104988.
5. Xiao, S. (2021) Traffic accident duration prediction based on natural language processing and a hybrid neural network architecture. Proceedings of 2021 International Conference on Neural Networks, Information and Communication Engineering, SPIE, 11933: 194-202. https://doi.org/10.1117/12.2614987.
6. Sayed, M.A., Qin, X., Kate, R.J., et al. (2021) Identification and analysis of misclassified work-zone crashes using text mining techniques. Accident Analysis & Prevention, 159: 106211. https://doi.org/10.1016/j.aap.2021.106211.
7. Kwayu, K.M., Kwigizile, V., Lee, K., et al. (2021) Discovering latent themes in traffic fatal crash narratives using text mining analytics and network topology. Accident Analysis & Prevention, 150: 105899. https://doi.org/10.1016/j.aap.2020.105899.
8. Hu, Y.C., Wei, H., Zeng, Q. (2019) Analysis of freeway crash severity based on spatial generalized ordered probit model. Journal of South China University of Technology (Natural Science Edition), 51(1): 114-122. https://doi.org/10.12141/j.issn.1000-565X.210758.
9. Ahmed, F., Siddiqui, C., Huynh, N. (2023) Analysis of temporal stability of contributing factors to truck-involved crashes at work zones in South Carolina. Transportation Research Record, 2677(2): 1484-1499. https://doi.org/10.1177/03611981221112097.

10. Zhang, C., Wang, B., Yang, S., et al. (2020) The driving risk analysis and evaluation in rightward zone of expressway reconstruction and extension engineering. Journal of Advanced Transportation, 2020(1): 1-13. https://doi.org/10.1155/2020/8943463.
11. Weng, J., Zhu, J.Z., Yan, X., et al. (2016) Investigation of work zone crash casualty patterns using association rules. Accident Analysis & Prevention, 92: 43-52. https://doi.org/10.1016/j.aap.2016.03.017.
12. Yang, Y., Yuan, Z.Z., Sun, D.Y., et al. (2019) Analysis of the factors influencing highway crash risk in different regional types based on improved Apriori algorithm. Advances in Transportation Studies, 49(3): 165-178. https://doi.org/10.4399/978882552809113.
13. Yuan, Z.Z., Lou, C., Yang, Y. (2021) Analysis of highway traffic accidents causes under time differences. Journal of Beijing Jiaotong University, 45(3): 1-7. https://doi.org/10.11860/j.issn.1673-0291.20200120.