FREEWAY ACCIDENT DURATION PREDICTION BASED ON SOCIAL NETWORK INFORMATION

K.K. Ji,[*] *Z.Z. Li,*[*] *J. Chen,*[†] *G.Y. Wang,*[‡] *K.L. Liu,*[†] *Y. Luo*[†]

Abstract: Accident duration prediction is the basis of freeway emergency management, and timely and accurate accident duration prediction can provide a reliable basis for road traffic diversion and rescue agencies. This study proposes a method for predicting the duration of freeway accidents based on social network information by collecting Weibo data of freeway accidents in Sichuan province and using the advantage that human language can convey multi-dimensional information. Firstly, text features are extracted through a TF-IDF model to represent the accident text data quantitatively; secondly, the variability between text data is exploited to construct an ordered text clustering model to obtain clustering intervals containing temporal attributes, thus converting the ordered regression problem into an ordered classification problem; finally, two nonparametric machine learning methods, namely support vector machine (SVM) and k-nearest neighbour method (KNN), to construct an accident duration prediction model. The results show that when the ordered text clustering model divides the text dataset into four classes, both the SVM model and the KNN model show better prediction results, and their average absolute error values are less than 22 %, which is much better than the prediction results of the regression prediction model under the same method.

Key words: *duration prediction, freeways, intelligent transportation, natural language processing, social networks*

Received: December 27, 2021
Revised and accepted: April 30, 2022

DOI: 10.14311/NNW.2022.32.006

1. Introduction

Traffic accidents are the leading cause of non-recurring highway congestion [1]. The national traffic incident management association estimates that 25 % of congestion on U.S. roadways is caused by traffic accidents that result in lane closures, resulting in reduced roadway capacity and consequent traffic congestion or delays [2]. Traffic accident impact is usually measured by traffic accident duration, and reducing the

[*] KeKe Ji; ZhengZhong Li – Corresponding author; Tianjin Transportation Research Institute, Tianjin, 300074, China, E-mail: jikk520@126.com lizhengzhonglzz@163.com

[†] Jian Chen; KeLiang Liu; Yi Luo; Shool of Traffic and Transportation, Chongqing Jiaotong University, Chongqing, 400074 China

[‡] GuanYan Wang; Tianjin Dongfang Tairui Technology Co. LTD., Tianjin 300192, China

duration of the accident is considered the most critical task to mitigate the impact of the accident [3]. Therefore, timely and accurate accident duration prediction will help road management to develop reasonable response strategies and reduce the subsequent impact caused by traffic accidents.

Studies on the duration of road traffic accidents have attracted much attention from transportation experts and scholars. In the existing literature, models to predict the duration of freeway accidents can be generally grouped into two main categories: parametric models and nonparametric models [1]. In statistics, parametric models usually assume that the data population follows some distribution, mainly regression, and hazard-based approaches. For example, Garib et al. built a linear regression model to predict the collected 205 road traffic accidents [4]. Valenti et al. verified that the multiple linear regression model had higher accuracy in predicting the duration of short-term accidents [5]. Unlike regression analysis, the hazard-based models have the advantage of capturing the effects of duration and are widely studied by many researchers [6]. Chung established the log-logistic accelerated failure time (AFT) metric model to predict the duration of freeways accidents in Korea and verified the model's validity [7]. Li prediction of the duration of traffic accidents on urban expressways at different stages is achieved by building an accelerated failure time hazard model [8]. Other types of hazard-based models, such as proportional hazards and Cox proportional hazards models, have also been widely used in previous studies [9, 10].

Nonparametric models typically make no assumptions about the overall distribution of data, mainly including support vector machine (SVM), k-nearest neighbour (KNN), artificial neural networks (ANN), and so on. This type of model has more flexibility than the parametric model and can handle the complex and nonlinear relationship among variables more easily [11]. Hamad et al. summarizes the application of five state-of-the-art machine learning (ML) models for predicting traffic incident duration, including regression decision tree, support vector machine (SVM), ensemble tree (bagged and boosted), Gaussian process regression (GPR), and artificial neural networks (ANN). The results showed that the SVM and GPR models outperformed other models. [12]. Li et al. established a two-level model consisting of a cost-sensitive Bayesian network and a weighted k-nearest neighbor model. The experiment results show that compared with the traditional classical model, this model has higher accuracy in predicting accident duration [13]. Lee et al. established a prediction model of accident duration based on ANN and verified that significant variables have an important influence on the prediction of accident duration [14]. Although the above studies have achieved good prediction results, the established duration prediction models often rely on specific accident characteristics. They are influenced by the source, quantity, and quality of accident data to a greater extent, which has more significant limitations. Therefore, a prediction method based on natural language processing techniques provides a new way of thinking for conducting accident duration prediction research. Pereira et al. combined a time series prediction model with a topic model based on natural language processing to predict the duration of road accidents. The results demonstrate that natural language processing techniques can effectively improve the performance of prediction models [15]. Li et al. developed a hybrid model based on polynomial logic and parametric risk based on the study of Pereira et al., which effectively improved the model prediction accuracy [9]. However, the studies mentioned above also only use text data to complement specific accident characteristics and cannot predict the accident duration entirely from the perspective of text data.

Due to the heterogeneous case-by-case nature of traffic accidents, plenty of relevant information is recorded in free-flow text fields instead of constrained value fields [9]. This also results in many accident characteristics that cannot be quantified. As an essential way of communication, especially in the Internet era, social network platforms have become an essential part of people's lives and the leading carrier of human language. Hundreds of millions of unstructured text data are published daily on major social network platforms. These social network platforms contain a large amount of road traffic accident information, which contains both objective description and subjective analysis, and is richer and more comprehensive than the accident status information collected in the traditional physical space [16]. More and more scholars are increasingly using social networking information to conduct relevant research. For example, Gu et al. used traffic events extracted from a Twitter text message and compared with actual traffic survey results, finding that Twitter can be used as a new traffic event detector. [16] Salas et al. performed real-time detection of traffic event information on Twitter through algorithmic models such as text classification, geolocation, and sentiment analysis [17].

However, before using social network information for related research, textual information needs to be quantified and processed. The commonly used text vectorization models include the topic model: like latent Dirichlet allocation (LDA) model; vector space model: like Term frequency-inverse document frequency (TF-IDF) model; and word vector model: like word to vector (Word2Vec) model [18]. The TF-IDF model is widely used owing to its simplicity and efficiency. In order to better characterize all the information content contained in the text, the text vectorization representation will show certain sparsity and high dimensionality, which is very unfavorable to the traditional regression model prediction methods. Moreover, studies have shown that under similar context, including time and location [13]. Therefore, this proposed highway accident duration prediction study is based on social network information. First, the heterogeneity among text data is eliminated by establishing an ordered text clustering model to reduce vectors' high dimensionality and sparsity after vectorization. Thus, the ongoing accident duration regression prediction problem is transformed into a classification problem. Then, two nonparametric machine learning methods, namely, support vector machine (SVM) and k-nearest neighbor method (KNN), are selected to implement the temporal prediction of accident text data. Finally, the performance of the created models is evaluated.

The remainder of this paper is divided into the following sections. Section 2 presents data sources and data analysis. Section 3 provides a brief introduction to the methodologies and a model evaluation indicator. Section 4 illustrates the evaluation results of four prediction models. Finally, Section 5 presents concluding remarks and suggestions for future research.

2. Data description

2.1 Data sources

This study selected the accident information published in the Weibo of "Sichuan freeway" as the experimental data. Weibo is a social media platform based on social relationships, which can share and interact with information instantly in the form of text, pictures, videos and other multimedia, similar to Twitter. These data contain information content such as the description of the accident condition, the time of the accident, and the loca tion of the accident, which meet the requirements of the experimental data for this study.

By using Python-based web crawler technology to mine the information in the Weibo of "Sichuan freeway", a total of more than 60 000 pieces of data were obtained from June 2019 to October 2020. The dataset contains three main blocks of information categories, release time and text information, which are partially shown in Tab. I.

Since this study focuses on the information of accident category, only the data of category traffic accident category and traffic recovery category are retained. The above data set is manually screened to find the traffic recovery information corresponding to each traffic accident, and the duration of a complete accident is calculated by Eq. (2), as show in Tab. II. Finally, 1 073 valid accident data with complete time tags were obtained.

$$Y_i = (recover\_time_i - rescue\_time_i) \times 60, \tag{1}$$

where:

– Y_i denotes the duration of the accident (unit: min);

– $rescue\_time_i$ denotes the time stamp for traffic management to start rescue (unit: min, converted to minutes);

– $recover\_time_i$ denotes the time stamp for traffic recovery (unit: min, converted to minutes).

2.2 Data description

The accident duration in this study represents the period between the time an accident is reported and when all handlers leave the accident site. The minimum, maximum, and average durations for 1 073 accidents were 3, 804, and 87 minutes. Tab. III shows the number, frequency, and cumulative frequency of durations for 1 073 accidents in different time intervals. For about 90 % of the accidents, the duration was less than 180 minutes. The percentage of accidents with more than 180 minutes of durations was 9.88 %, and 3.36 % for durations greater than 270 minutes.

By fitting the probability distribution to the experimental data in this paper, it is found that the accident duration approximately shows a Poisson distribution, as shown in Fig. 1. However, the logarithm of the accident duration approximation shows a standard normal distribution, in agreement with the findings of Giuliano, Skabardonis, and others [17, 18], as shown in Fig. 2.

No.	Class	Posting Time	Texts
1	Entrance close	Oct. 16 12:15	Notification from fressway traffic officer: In the Chengdu to Yaan section of the G5 Beijing-Kunming fressway, the entrance of the Xinjinnan Toll Station from Chengdu to Yaan is closed because of the traffic accident in the direction from Chengdu to Yaan at 1846 kilometers.
2	Entrance open	Oct. 16 13:18	Notification from freeway traffic officer: In the Chengdu to Yaan section of the G5 Beijing-Kunming fressway, the entrance of the Shuangliubei Toll Station and Shuangliunan Toll Station, which were closed because of the traffic accident in the direction from Chengdu to Yaan at 1846 kilometers, are in service.
3	Travel tips	Oct. 16 13:36	In the Guangyuan section and Guangyuan to Shaanxi section, the entrance of Jianmenguan Toll Station, Guangyuan Toll Station, Chaotian Toll Station, Zhongzi Toll Station, and Zhaohua Toll Station is open to trucks with three axes and above, which were prohibited except for fresh produce vehicles because of the road construction in the Mianyang section.
4	Traffic accident	Oct. 16 14:48	In the Chengdu to Yaan section of the G5 Beijing-Kunming freeway, there is a rear-end crash with two vehicles in the direction from Yaan to Chengdu at 1814 kilometers. The first lane, the second lane, and the third lane have been occupied. The emergency lane can be used.
5	Traffic recovery	Oct. 16 15:37	In the Chengdu to Yaan section of the G5 Beijing-Kunming freeway, the handling of the accident in the direction from Yaan to Chengdu at 1814 kilometers has finished. The traffic recovers to normal.

Tab. I *"Sichuan freeway" Weibo data.*

No.	Texts	Time/min
1	In the Chengdu to Yaan section of the G5 Beijing-Kunming freeway, there is a rear-end crash with two vehicles heading from Yaan to Chengdu at 1814 kilometers. The first lane, the second lane, and the third lane have been occupied. The emergency lane can be used.	49

Tab. II *Duration of accident information schedule.*

Accident Duration/min	Samples	Frequency [%]	Cumulative Frequency [%]
0-30	170	15.84	15.84
30-60	332	30.94	46.78
60-90	234	21.81	68.59
90-120	114	10.62	79.22
120-150	74	6.90	86.11
150-180	43	4.01	90.12
180-210	37	3.45	93.57
210-240	19	1.77	95.34
240-270	14	1.30	96.64
Over 270	36	3.36	100.00

Tab. III *Accident duration divided into interval statistics.*

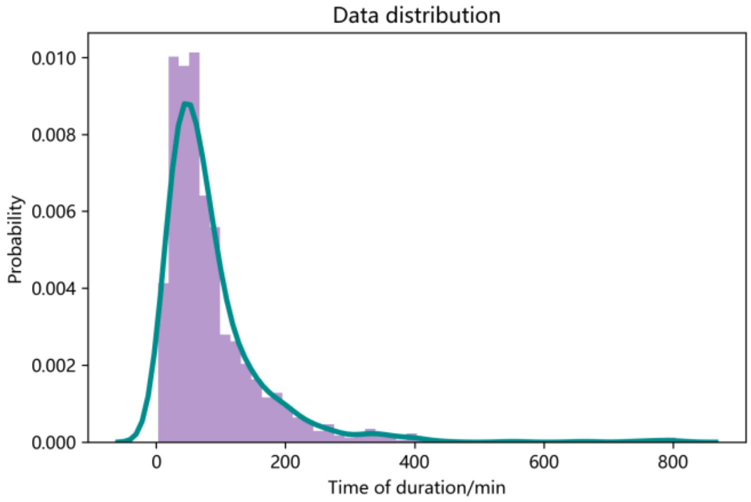


Fig. 1 *Accident duration distribution.*

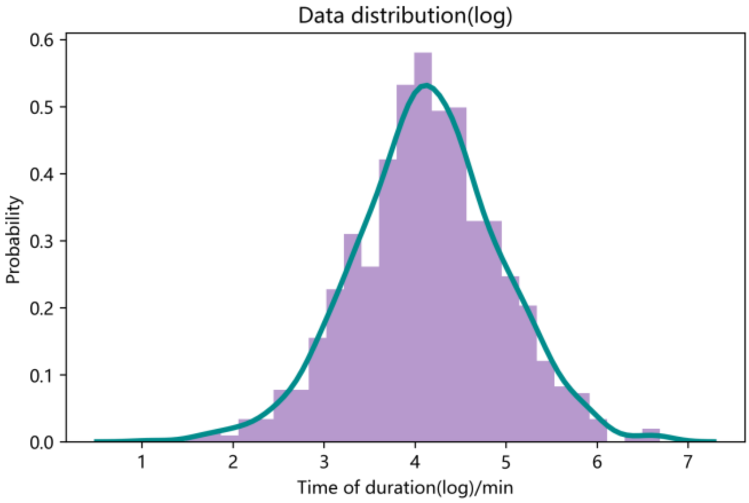


Fig. 2 *Accident duration log distribution.*

3. Model construction

The accident duration prediction is carried out in the following steps, as shown in Fig. 3.

Step 1: Data set division. First, consider dividing the accident text dataset obtained above into two parts: the training and test sets. The training set is used for the later ordered text clustering model and the parameter calibration and training of the classification model. Furthermore, the test set is used to verify the prediction effectiveness of the trained model on the accident duration.

Step 2: Natural language processing. This part mainly includes pre-processing and vectorized representation of incident text data. The pre-processing is mainly to complete the text data word separation and deactivation operations. Moreover, the text vectorized representation is designed to convert text data into structured data that computers can recognize.

Step 3 Ordered text clustering model. Using the sum of squares of deviations to characterize the degree of differences among text features, with the slightest difference between similar text features and the most significant difference between different categories of text features, to achieve ordered text clustering, and then transform the regression problem into a classification problem.

Step 4: Construction, training, and performance evaluation of text classifiers. Based on the above steps, a text data classifier is constructed, word vectors from the training set are input to the classifier for training, and the classifier parameters

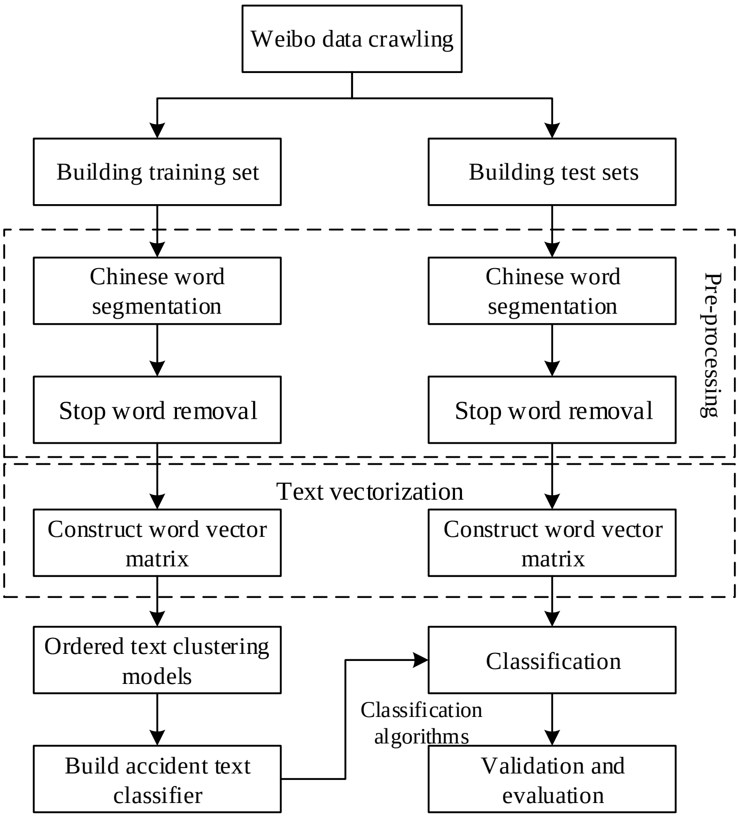

Weibo data crawling
Building training set
Building test sets
Chinese word segmentation
Chinese word segmentation
Pre-processing
Stop word removal
Stop word removal
Text vectorization
Construct word vector matrix
Construct word vector matrix
Ordered text clustering models
Classification
Classification algorithms
Build accident text classifier
Validation and evaluation

Fig. 3 *Technical framework diagram.*

are adjusted to achieve optimal performance. Finally, the performance is evaluated on the test set.

3.1 Natural Language Processing

Natural language processing (NLP) is the process of transforming the language used by people in their daily communication into a language that can be understood by computers through a series of processing, in order to solve the problem of "making natural language understandable by machines", which mainly includes word separation, deactivation and vectorized representation of text [19].

(1) Word segmentation and stop word removal

Word segmentation and stop word removal operations are performed on the training and test sets. The Chinese word segmentation is Chinese information processing foundation. In the word segmentation process, select the precise word segmentation mode in Jieba word segmentation tool in Python environment. By adding new words, customizing and expanding the participle word lists to achieve more accurate word segmentation.

It is also important to remove all words that have no semantic relevance for the problem, such as articles, prepositions, and pronouns, thus reducing the number of redundant text features. In stopping words, the common stopping words lists in natural language processing are chosen as the basic stopping words lexicon. Moreover, custom addition of nonsense words present in the experimental data, thus constituting a new stopping words lists. Examples are shown in Tab. IV.

It can be seen from the original texts in Tab. IV. This contains a lot of words that have no real meaning, but only serve as a link. For example, where the in, the, and to, etc. In order to reduce the influence of nonsense words during the experiment. Therefore, it is necessary to delete the words in this category. The result after word segmentation and stop word removal is shown in preprocessed texts.

	Texts
Original texts	In the Chengdu to Yaan section of the G5 Beijing-Kunming fressway, there is a rear-end crash with two vehicles in the direction from Yaan to Chengdu at 1814 kilometers. At present, the first lane, the second lane and the third lane have been occupied. The emergency lane can be used.
Preprocessed texts	Beijing-Kunming fressway/ Yaan/Chengdu/ two vehicles/ rear-end/the first lane/the second lane/the third lane/ emergency lane

Tab. IV *Example of pre-processed text.*

(2) Text vectorization

Text data is an unstructured data type that cannot be read directly by the computer itself and must be converted into a digital language that the computer can recognize through text data processing. The commonly used methods for vectorized text representation include the Boolean model, statistical topic model, and vector space model. One of the vector space models is by extracting feature terms in text data and then transforming these feature terms into the form of vector combinations in vector space. The more common methods include the term frequency-inverse document frequency (TF-IDF) model, which is efficient and straightforward. Each document is represented as a vector in n-dimensional vector space in the TF-IDF model.

The TF-IDF model is a prevalent research method in natural language processing (NLP) used in the implementation of the algorithm described in this article [20].

TF-IDF model is the product of term frequency (TF) and inverse document frequency (IDF). Wherein TF – term frequency of term i in document j

and IDF – inverse document frequency of term i [21]. TF and IDF can be calculated as Eq. (2), (3).

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{2}$$

where

– $tf_{i,j}$ denotes the number of times word t_i occurs in document d_j,

– $n_{i,j}$ denotes the frequency of t_i words in the d_j document,

– k denotes the total number of words in the d_j document,

– $n_{k,j}$ denotes the total number of occurrences of all words in the d_j document.

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}| + 1}, \tag{3}$$

where

– idf_i denotes the inverse document frequency exponential function of t_i words,

– $|D|$ denotes the total number of documents,

– $|\{j : t_i \in d_j\}|$ denotes the number of documents containing t_i words.

TF-IDF can be calculated as (4):

$$TF - IDF = tf_{i,j} \times idf_i \tag{4}$$

Based on the TfidfVectorizer in Python language environment, the TF-IDF model parameters are set as shown in the following Tab. V.

Parameters	Parameters Value	Parameters	Parameters Value
norm	L2	ngram_range	(6,10)
sublinear_tf	True	max_df	0.6

Tab. V *TF-IDF parameter table.*

In this study, a total of 969 accident sample sets with a duration of less than 180 min were used as the experimental set. Using the phrase [Beijing-Kunming freeway, Yaan, Chengdu, two vehicles, rear-end, the first lane, the second lane, the third lane, emergency lane] as a practical example, the TF-IDF values of each word were calculated as shown in Tab. VI.

From Tab. VI, “Two vehicles” and “Rear-end” have the highest weight value in the sentence. It also reflects the leading cause of the accident and the adaptability of TF-IDF to the experimental data in our research.

Words	TF-IDF values	Words	TF-IDF values
Beijing-Kunming fressway	0.119	the first lane	0.285
Yaan	0.151	the second lane	0.254
Chendu	0.359	the third lane	0.123
Two vehicles	0.549	Emergency lane	0.469
Rear-end	0.498		

Tab. VI *TF-IDF values.*

3.2 Ordered text clustering models

Due to the diversity of language expression and the heterogeneity of traffic accidents, there may be multiple expression schemes for the same accident. Simply using the quantified text data for regression prediction analysis will inevitably lead to deviation from the actual results. In addition, a slight bias in time prediction does not affect freeway travelers. Therefore, considering a clustering means to divide the experimental data, which is no intuitive rule, to reduce the influence of multivariate data expression on the prediction accuracy.

The traditional clustering method considers that the data are not different and equal to each other, so the clustering results are disorganized. Nevertheless, the accident text data exists in a time-ordered form. Therefore, this study is based on the idea of optimal segmentation method to design a clustering method that minimizes the variability between similar samples and maximizes the variability between samples of different classes while ensuring the same order of experimental samples data. The modeling process is as follows:

(1) Define class diameter

Suppose a class of ordered experimental data $G = \{X_i, X_{i+1}, \ldots, X_j\}$ $(j > i)$, where, X_i, X_j both denote a vector in the data set G after text vectorization. Then the mean vector $\bar{X}_G$ of this class is denoted as follows:

$$\bar{X}_G = \frac{1}{j-i+1} \sum_{t=i}^{j} X_t. \tag{5}$$

$D(i,j)$ denotes the diameter of this class of experimental data, which is expressed by the formula:

$$D(i,j) = \sum_{t=i}^{j} (X_t - \bar{X}_G)(X_t - \bar{X}_G)^{\mathrm{T}}. \tag{6}$$

(2) Define the loss function

$b(n,k)$ denotes a division of the n ordered experimental sample data into k classes.

$$G_1 = \{i_1, i_1 + 1, \ldots, i_2 - 1\},$$

$$G_2 = \{i_2, i_2 + 1, \ldots, i_3 - 1\},$$

$$G_k = \{i_k, i_k + 1, \dots, n\}\,, \tag{7}$$

where the classification points are $1 = i_1 < i_2 < \cdots < i_k < n = i_{k+1} - 1(i_{k+1} = n + 1)$.

The loss function of the above classification method is defined as

$$L\left[b(n,k)\right] = \sum_{t=1}^{k} D(i_t, i_{t+1} - 1), \tag{8}$$

when n, k are fixed, a smaller value of the loss function $L[b(n,k)]$ indicates a smaller sum of squares of the deviations between the various types of sample data. The ordered text clustering model is to find a classification method that minimizes the loss function $L[b(n,k)]$, which is then denoted as $p[n,k]$.

(3) Recurrence formula

The recurrence formula to obtain the classification as $p[n,k]$ with the smallest loss function value $L[b(n,k)]$ is calculated as shown in (9):

$$\begin{cases} L[b(n,2)] = \min_{2\le j\le n}\left\{D(1, j-1) + D(j,n)\right\}, \\ L[b(n,k)] = \min_{k\le j\le n}\left\{L[P(j-1,k-1)] + D(j,n)\right\}. \end{cases} \tag{9}$$

The above formula indicates that if the optimal segmentation of n ordered samples into k classes is found, it should be based on the optimal segmentation of $j - 1(j = 2, 3, \dots, n)$ ordered samples into $k - 1$ classes.

(4) Optimal segmentation

Suppose the number of classifications is $k(1 < k < n)$, and find the optimal classification $p[n,k]$ under the minimum loss function.

First, find the segmentation point j_k that makes the recurrence formula reach the minimum value, the value of the loss function $L[P(n,k)] = L[P(j_k - 1, k - 1)] + D(j_k, n)$, and the kth class's segmentation interval is $G_k = \{i_k, i_k + 1, \dots, n\}$.

Second, find the j_{k-1} segmentation point that makes the recurrence formula reach the minimum value, at which point the value of the loss function is $L[P(j_{k-1}, k-1)] = L[P(j_k - 1, k-2)] + D(j_{k-1}, j_{k-1})$, and the segmentation interval for the k-th class is $G_k = \{i_{k-1}, i_{k-1} + 1, \dots, j_k - 1\}$.

Based on the above method, all classification results $G_1, G_2, \dots, G_k$ can be obtained and expressed as $P\,(n,k) = \{G_1, G_2, \dots, G_k\}$.

(5) Determine the optimal number of categories

The curve method and the β test were chosen to determine the optimal classification status by reading the relevant literature.

- Curve method

  By plotting the variation curve of the minimum loss function L with the number of classifications k and taking the classification number k corresponding to the point where this curve becomes slower or abruptly changed as the optimal classification number.

- β test

  We are using Eq. (10) to derive the optimal classification result.

  $$\beta = \frac{L\left[b(n,k)\right]}{L\left[b(n,k+1)\right]}, \tag{10}$$

  when the value of β is more considerable, it means that the segmentation into $k+1$ classes is better than k classes; the closer the value of β is to 1, then the number of classes k in the segmentation can be considered the optimal number of categories.

4. Classifier construction

Based on a promising performance in the literature, this study employed two non-parametric machine learning methods, namely the support vector machines and k-nearest neighbour, to construct the accident duration prediction model.

(1) Support vector machines

Support vector machine (SVM) is a supervised learning algorithm based on the statistical learning theory. It can be used for classification and regression tasks and has a low generalization error rate and high learning ability. This method has been widely used in previous accident duration prediction studies [22,23].

Support vector machines are used to find an optimal separation hyperplane by constructing a function that maximizes these separation margins or the distance between it and the nearest data point of each class [24]. When SVM models are applied to text classification problems, the first step is to map the input text vector into a feature space, find an optimized linear divider in this feature space, and then construct a hyperplane separating the categories, enabling the classification of text data.

(2) k-nearest neighbour (KNN)

The k-nearest neighbour algorithm is also an algorithmic model that can be used for both classification and regression tasks, and it has received much attention in accident duration prediction studies [14,25,26]. The "neighbors" in the KNN method represent the training instances in the metric space, and the metric space is represented as the feature value of the distance between all set members.

The procedure of KNN for calculation is as follows:

- For a set of training samples $X_{tr} = \{x_1, x_2, \ldots x_i\}$, $X_{te} = \{x_1, x_2, \ldots x_j\}$, $Y_{tr} = \{y_1, y_2, \ldots y_i\}$ and $Y_{te} = \{y_1, y_2, \ldots y_j\}$, where X_{tr} represents the training set, X_{te} represents the test set, Y_{tr} represents the training set label, Y_{te} represents the test set label, and each sample has the same number of features n.

- Calculate the distance between training data x_i and testing data x_j using the Euclidean distance shown in Eq. (11).

$$d\left(x_i, x_j\right) = \sqrt{\sum_{n=1}^{N} \left(x_i^n - x_j^n\right)^2}, \tag{11}$$

where x_i denotes the ith training data, $x_i \in X_{tr}$. x_j denotes the jth training data, $x_j \in X_{te}$. N denotes the number of data sets.

- Sort $d\left(x_i, x_j\right)$ for each testing data x_j in ascending order and elect the first K closest training data set $\{y_1, y_2, \ldots y_k\}$.
- Eq. (12) is used to average the first K closest training data set $\{y_1, y_2, \ldots, y_k\}$ as the estimated y_j.

$$\hat{y}_j = \frac{\sum\limits_{k=1}^{K} d\left(x_k, x_j\right) \times y_k}{\sum\limits_{k=1}^{K} d\left(x_k, x_j\right)}. \tag{12}$$

4.1 Performance evaluation

In order to evaluate the performance of the prediction model, the accuracy of the prediction results needs to be evaluated. Taking reference from the metrics mean absolute error (MAE) and mean square error (MSE) used to measure the performance of regression tasks, mean absolute percentage error (MAPE) is introduced as a model prediction performance evaluation criterion, which is a summary measure widely used for evaluating the accuracy of prediction results [25, 27].

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{v_i^{\text{actual}} - v_i^{\text{predicted}}}{v_i^{\text{actual}}} \right| \times 100\%, \tag{13}$$

where

- MAPE denotes mean absolute percentage error;
- n denotes the number of experimental sample sets;
- v_i^{actual} denotes the actual value of observation;
- $v_i^{\text{predicted}}$ denotes the predicted value of observation.

The lower the MAPE value is, the more accurate the prediction model will be. When the MAPE value is less than 10 %, the prediction accuracy is the highest; the MAPE value is between 11 % and 20 %, it means the prediction accuracy is good; the MAPE value is between 21 % and 50 %, it means the prediction is reasonable [14].

Since the results of the predicted observations in this study exhibit different categories, the mean value of each category is considered as its corresponding predicted value. The refined Equation is shown as (14), and the evaluation criteria of MMAPE are the same as those of MAPE.

$$\text{MMAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{v_i^{\text{actual}} - v_i^{\text{Mtypes}}}{v_i^{\text{actual}}} \right| \times 100\%, \tag{14}$$

where

- MMAPE denotes refined mean absolute percentage error;
- v_i^{Mtypes} denoted as the mean value corresponding to each category.

5. Evaluation

Since the fundamental theory of SVM and KNN training algorithms is stochastic-oriented, different settings of initial training parameters may lead to different convergence states. However, the existing studies have not provided an optimal parameter setting scheme. Therefore, during the training process, it is necessary to combine the characteristics of the experimental data itself and train at an appropriate scale to select a model with optimal performance.

In this study, a total of 969 accident sample sets with a duration of less than 180 min were used as the experimental set, ten experiments were conducted for examining the proposed methodology. In each experiment, 80 % of the data was randomly selected as the training set from the sample set and the remaining 20 % of data served as the testing set.

This study developed four types of accident duration prediction models based on a third-party "Sklearn" machine learning library. Among them, both Model1 and Model2 are based on the results of the ordered text clustering model and use SVM and KNN methods for classification prediction studies. Model3 and Model4 are based on nonlinear data fitting methods and use SVM and KNN methods for regression prediction studies. The same training and test sets were used for all four models in each experiment.

5.1 Results of the ordered text clustering model

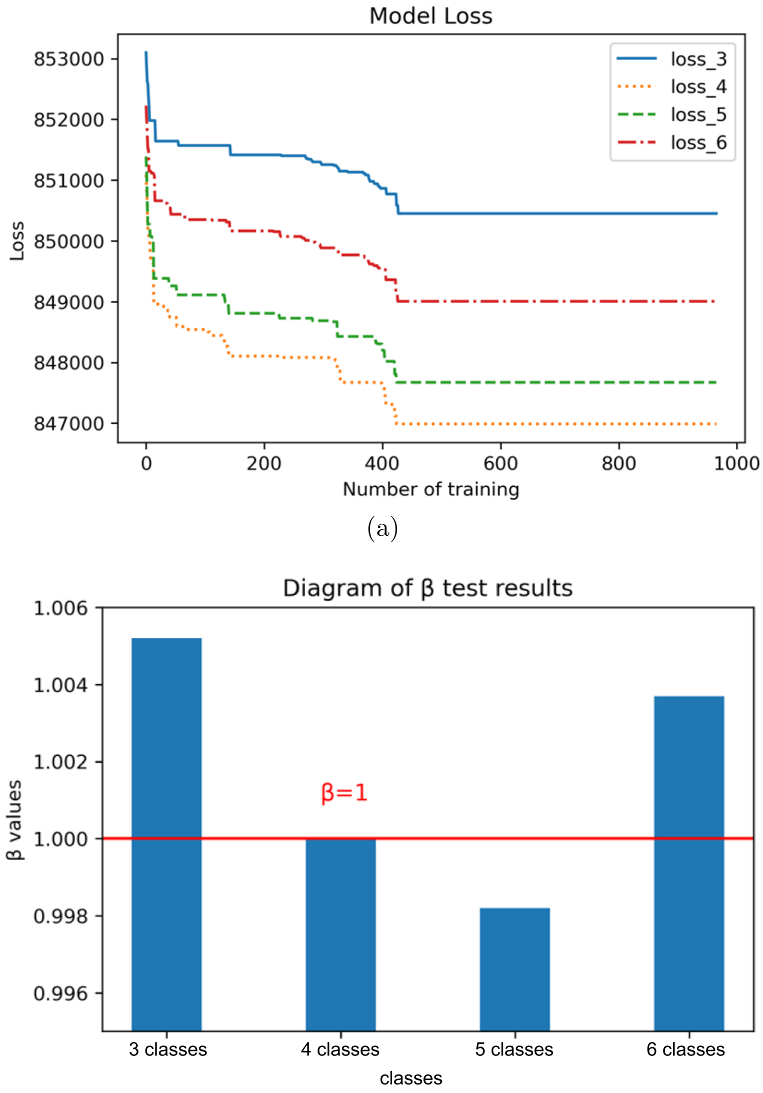
Tab. VII presents the results of the ordered text clustering model, including the values of the loss functions corresponding to different clustering intervals and the β values. When the cluster number is 4, the smallest loss function value and the β value are the closest to 1. According to the β-test, the clustering results in

Number of clusters	Clustering results /min						Loss function	β values
3 classes	<19	20-53	>53				851028.467	1.0052
Time Averages	13.479	36.466	92.824					
4 classes	<19	20-53	54-98	>98			846612.148	1.000
Time Averages	13.479	36.466	72.892	131.807				
5 classes	<19	20-53	54-97	98	>98		847511.784	0.9982
Time Averages	13.479	36.466	72.163	98	131.807			
6 classes	<19	20-53	53	54-97	98	>98	849004.946	1.0037
Time Averages	13.479	36.466	53	72.163	98	131.807		

Tab. VII *Clustering results and loss function values, β values.*

this state are considered the optimal clustering results, which are less than 19 min, 20-53 min, 54-98 min, and greater than 98 min, respectively.

Fig. 4(a) shows the change process of the loss function values of different clustering results. After the training times exceed 400, the loss function values of each clustering result are close to smooth, indicating that the optimal clustering result has been reached. Fig. 4(b) demonstrates that the β value is closest to 1 when the clustering results in 4 classes, satisfying the criteria of the β test method.



(a)

(b)

Fig. 4 *Loss function diagram and β test.*

5.2 Results of the accident duration prediction model

Based on the accident duration prediction model established above, the prediction results of the four models were compared and analyzed separately, as shown in Tab. VIII. With the ordered text clustering models established in this study, the average MAPE value of each model is below 22 %, which is at a reasonable prediction level value. In model 1, the SVM model performed the best with the smallest MAPE value of 19.8 %. In model 2, the KNN model performed the best with the smallest MAPE value of 19.7 %. Through the comparative analysis of the two models, the KNN model prediction performance is slightly better than the SVM model prediction performance in this experimental dataset.

	Model1	Model2	Model3	Model4
Method for model construction Experiments	S-SVM [%]	S-KNN [%]	SVM [%]	KNN [%]
1	22.4	21.4	8812	6048
2	23.8	21.0	8470	6104
3	20.9	19.7	8751	6897
4	19.8	20.0	9271	7105
5	21.9	20.2	8352	6578
6	22.3	21.4	8687	6264
7	23.2	20.3	8769	6029
8	20.2	22.6	9106	6482
9	21.7	21.8	8962	7206
10	20.5	19.9	8546	6563
Average	21.7	20.8	8773	6528

Tab. VIII *Comparison of prediction performance of different models.*

However, for the traditional regression prediction model, see Model3 and Model4 prediction results shown. The average MAPE values all exceeded 6500 %, and even the average MAPE value of the SVM model reached 8773 %, which was 404.3 times higher than that of the S-SVM model. Comparing the four sets of models shows that Model1 and Model2 can be considered effective modeling methods for predicting accident duration when a set of accident text description data is provided.

Fig. 5 and Fig. 6 show the performance assessment concerning the predicted accident duration vs. actual accident duration. Wherein Fig. 6, Time1 means less than 19 min; Time2 means 20-53 min; Time3 means 54-98 min; Time4 means more than 98 min.

As can be seen, the prediction results of the SVM model in Model1 and the KNN model in Model2 are consistent with the actual outcome categories, and the KNN model can better match the actual outcome categories. This means that the KNN model can better catch the relationship between the text vectorized feature values and the temporal labels.

For the prediction results of the regression model, the SVM model predicts less fluctuating results, basically concentrating on about 60 min. Nevertheless, the KNN

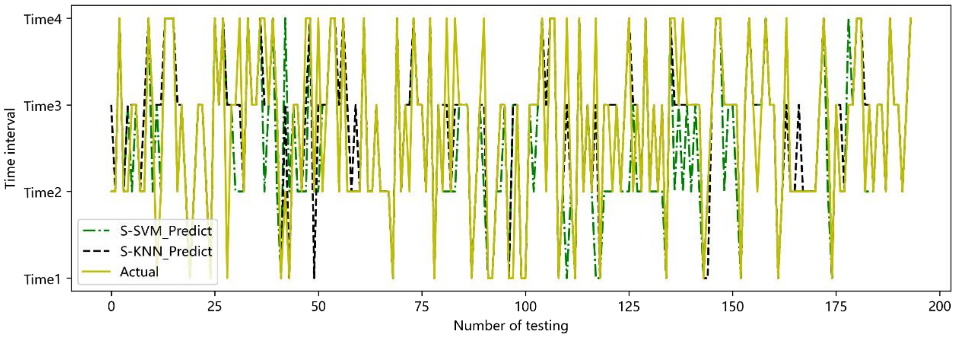


Fig. 5 *Model1 and Model2 duration prediction results.*

model is consistent with the actual value trend, with some fluctuation changes, but still a significant gap with the actual value. Therefore, the above comparative analysis shows that the prediction effect of using an ordered text clustering model is better than the traditional regression prediction model.

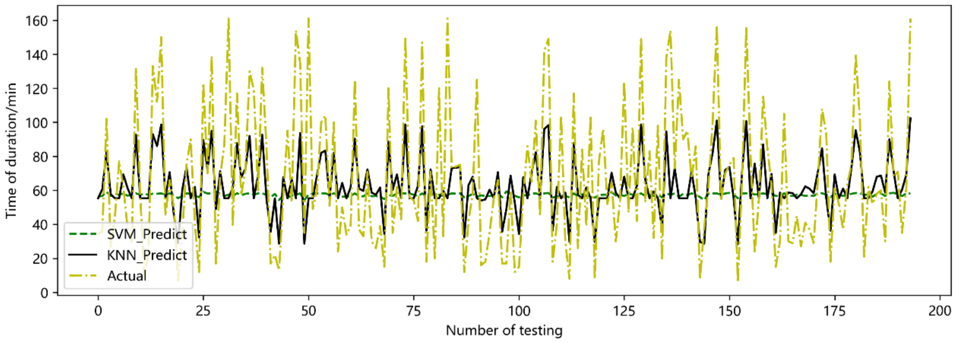


Fig. 6 *Model3 and Model4 duration prediction results.*

6. Conclusion

In this study, 969 traffic accident text data of the Sichuan freeway were collected by web crawler technology, and an accident duration prediction model based on social network information was established. Before model building, the TF-IDF method is used to extract and quantify the features of the accident text data. Considering the complex structure and high dimensionality of the accident data in social network information after quantification, an ordered text clustering model is established to make the slightest difference between similar samples and the most significant difference between different classes, from reducing the dimensionality of accident duration. After obtaining different clustering time intervals, the support vector machine (SVM) and k-nearest neighbour (KNN) methods were employed to develop the prediction model. The evaluation results of prediction models indicate that the proposed ordered text clustering model predicts much better than the traditional prediction model in various model compositions based on mean absolute percent error values, and the KNN model performs the best performance.

For future studies, more accident data should be collected and optimize model training parameters to facilitate the learning ability of the model. Above all, exploring the integration with structured fixed-value data is also an issue for future research.

Acknowledgement

This work was supported by the Scientific and Technological Innovation Projects of Chongqing Municipal Education Commission, Tianjin Transport Science and Technology Development Plan Project under Grant KJQN202001611, 2020-01.

References

[1] LI L., SHENG X., DU B., WANG Y., RAN B. A deep fusion model based on restricted Boltzmann machines for traffic accident duration prediction, Engineering Applications of Artificial Intelligence, 93, 2020.

[2] SHANG Q., TAN D., GAO S., FENG L. A hybrid method for traffic incident duration prediction using BOA-optimized random forest combined with neighborhood components analysis, Journal of Advanced Transportation, 2019, 11 pages.

[3] TANG J., ZHENG L., HAN C., YIN W., ZHANG Y., ZHOU Y., HUANG H. Statistical and machine-learning methods for clearance time prediction of road incidents: a methodology review, Analytic Methods in Accident Research, 27, 2020.

[4] GARIB A., RADWAN E., AL-DEEK H. Estimating magnitude and duration of incident delays, J. Transp. Eng., 1997, 123(6), pp. 458–466.

[5] VALENTI G., LELLI M., CUCINA D. A comparative study of models for the incident duration prediction, European Transport Research Review, 2010, 2(2), pp. 103–111.

[6] MA X., DING C., LUAN S., WANG Y., WANG Y. Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method, IEEE Trans. Intell. Transp. Syst., 2017, 18(9), pp. 2303–2310.

[7] CHUNG Y. Development of an accident duration prediction model on the Korean Freeway Systems, Accid. Anal. Prev., 42(1), pp. 282–289, 2010.

[8] LI R. Traffic incident duration analysis and prediction models based on the survival analysis approach, Iet Intelligent Transport Systems, 2015, 9(4), pp. 351–358.

[9] LI R., PEREIRA F.C., BEN-AKIVA M.E. Competing risks mixture model for traffic incident duration prediction, Accident Anal. Prevention, 2015, 75, pp. 192–201.

[10] LIN L., WANG Q., SADEK A.W. A combined M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations, Accident Anal. Prevention, 2016, 91, pp. 114–126.

[11] MA X., TAO Z., WANG Y., YU H., WANG Y. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data, Transp. Res. C Emerg. Technol., 2015, 54, pp. 187–197.

[12] HAMAD K., KHALIL M.A., ALOZI A.R. Predicting freeway incident duration using machine learning. International Journal of Intelligent Transportation Systems Research, 2019, 18, pp. 367–380.

[13] KUANG L., YAN H., ZHU Y., TU S., FAN X. Predicting duration of traffic accidents based on cost-sensitive bayesian network and weighted k-nearest neighbor, J. Intell. Transp. Syst., 2019, 23(2), pp. 161–174.

[14] LEE Y., WEI C.-H., CHAOV K.-C. Non-parametric machine learning methods for evaluating the effects of traffic accident duration on freeways, Archives of Transport, 2017, 43(3), pp. 91–104.

[15] PEREIRA F.C., RODRIGUES F., BEN-AKIVA M. Text analysis in incident duration prediction, Transp. Res. C Emerg. Technol., 2013, 37, pp. 177–192.

[16] GU Y., QIAN Z.S., CHEN F. From Twitter to detector: Real-time traffic incident detection using social media data, Transp. Res. C Emerg. Technol., 2016, 67, pp. 321–342.

[17] SALAS A., GEORGAKIS P., NWAGBOSO C., AMMARI A., PETALAS I. Traffic Event Detection Framework Using Social Media[C] 2017 International Conference on Smart Grid and Smart Cities. IEEE, 2017.

[18] ROUL R.K. An effective approach for semantic-based clustering and topic-based ranking of web documents, International Journal of Data Science and Analytics, 2018, 5(3), pp. 1–16.

[19] LI Y., WANG X., XU P. Chinese text classification model based on deep learning, Future Internet, 2018, 10(11), p. 113.

[20] LI F., YIN Y., MAO X., SHI R., SHI J. Method of feature reduction in short text classification based on feature clustering, Appl. Sci., 2019, 9(8), pp. 1578.

[21] LILLEBERG J., ZHU Y., ZHANG Y. Support vector machines and Word2vec for text classification with semantic features, Proc. IEEE 14th Int. Conf. Cogn. Inform. Cogn. Comput. (ICCICC), 2015, pp. 136–140.

[22] GHOSH S., DESARKAR M.S. Class specific TF-IDF boosting for short-text classification: Application to short-texts generated during disasters, Companion of the The Web Conference 2018 (WWW '18), 2018, pp. 1629–1637.

[23] TRSTENJAK B., MIKAC S., DONKO D. KNN with TF-IDF based framework for text categorization, Procedia Eng., 2014, 69, pp. 1356–1364.

[24] YU B., WANG Y.T., YAO J.B., WANG J.Y. A comparison of the performance of ANN and SVM for the prediction of traffic accident duration, Neural Network. World, 2016, 26(3), pp. 271–287.

[25] LIN Y., LI R. Real-time traffic accidents post-impact prediction: Based on crowdsourcing data, Accident Anal. Prevention, 2020, 145.

[26] SHAH K., PATEL H., SANGHVI D., SHAH M. A comparative analysis of logistic regression, random Forest and KNN models for the text classification, Augmented Human Research, 2020, 5(1), pp. 1–16.

[27] CHEN H., RAKHA H.A. Real-time travel time prediction using particle filtering with a non-explicit state-transition model, Transp. Res. C Emerg. Technol., 2014, 43, pp. 112–126.