Traffic & Transportation

# Analysis of Beijing Traffic Violations Based on the BERT-CRF Model

Jie LI[1], Yuntao SHI[2], Shuqin LI[3]

Original Scientific Paper
Submitted: 26 June 2023
Accepted: 28 Dec. 2023

[1] lijie1986@ncut.edu.cn, North China University of Technology
[2] Corresponding author, shiyuntao@ncut.edu.cn, North China University of Technology
[3] lsq@ncut.edu.cn, North China University of Technology



Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

**ABSTRACT**
Traffic violations are a major cause of traffic accidents, yet current research falls short in comprehensively analysing these violations and the named entity method fails to extract the name of traffic violation events from records, thereby lacking in providing guidance for managing urban traffic violations. By expanding the People's Daily dataset from 71,456 words to 95,291 words, the BERT-CRF (Bidirectional Encoder Representations from Transformers-Conditional Random Field) model achieves an accuracy rate of 88.53%, a recall rate of 92.90% and an F1 score of 90.66%, successfully identifying event, time and location named entities within traffic violations. The data of traffic violations is then enhanced through forward geocoding and the Bayesian formula, and traffic violations are analysed from time, space, administrative region, gender and weather, to provide support for the dynamic allocation of law enforcement forces on traffic scenes and the precise management of traffic violations.

**KEYWORDS**
traffic violation; traffic accident; name entity; People's Daily; BERT-CRF; Bayesian formula.

## 1. INTRODUCTION

According to the 2021 China Statistical Annual Report, there were 3,872 traffic accidents in Beijing in 2020, with 964 deaths and 3,369 injuries, resulting in 49.242 million yuan [1] in direct property losses. The occurrence of traffic accidents is associated with factors such as distracted driving [2] and traffic violations [3]. There is considerable research on distracted driving, but less attention has been given to traffic violations. With the significant improvement in the digitalisation of urban governance and the widespread application of artificial intelligence [4, 5], there has been an accumulation of data resources related to urban traffic law enforcement. However, most of the data resources are described through natural language, which cannot be directly used for data analysis, cannot directly guide the construction of urban traffic violation monitoring system and cannot support the accurate management of traffic violations. Therefore, there is an urgent need to augment existing data resources through data enhancement methods, analyse the augmented data, and provide support for the governance of urban traffic violations.

Wang et al. [6] utilised the BERT-BiLSTM-CRF model to extract event information related to traffic violations, constructing a traffic violation event graph. However, they did not augment or enhance the traffic violation data and the data scale is small. Zhao et al. [7] created a heat map for traffic violations in Fuzhou city based on the temporal and spatial dimensions, omitting considerations for factors such as weather. Li et al. [8] analysed common traffic violations like running red lights and illegal parking using data from Automated Enforcement Systems (AES) but did not involve the analysis of serious traffic violations. Existing research faces challenges such as difficulty in acquiring data and incomplete analyses. However, with the increasing openness of information, records of urban road traffic violations can be promptly obtained

through internet platforms. By employing natural language processing techniques for named entity recognition (NER) on these records, and enhancing extracted information such as time and location, comprehensive support can be provided for the dynamic allocation of law enforcement personnel on urban traffic scenes and the precise governance of traffic violations.

The NER is the central link in the field of information extraction and knowledge graph construction, aiming to extract specific types of entities such as person names, place names, organisation names and so on from complex structured, unstructured and semi-structured data, and classify these entities with specific meaning [9], which has piqued the academics' interest. Xiao et al. [10] used the BiLSTM-CRF model to recognise the named entities in traditional Chinese medical case texts, advancing text mining in traditional Chinese medicine; Chen et al. [11] applied the BERT-BiLSTM-CRF model to named entity identification in Chinese rock description texts, with the goal of automating geological knowledge extraction; Zhao et al. [12] employed the BERT model to recognise named entities in agricultural text information, hence addressing the issue of polysemy in agricultural texts; Zeng et al. [13] used the BERT model to recognise named entities in publicly available verdict documents from the Shanghai High People's Court, introducing a new way for extracting important information from verdict records; Liu et al. [14] used the BERT-BiLSTM-CRF model to perform an intelligent analysis of power industry accident reports, introducing a new technique to assessing incident reports in the power industry; Wang et al. [15] used the BERT-BiLSTM-CRF model to recognise earthquake emergency information in online media, swiftly and accurately retrieving earthquake emergency information; The BERT-CRF model was utilised by Li et al. [16] to extract maize breeding entity relationships, establishing the groundwork for the construction of a maize breeding knowledge graph and other downstream tasks. The NER method is applied in various fields such as civil engineering, agriculture and judiciary, but it is less commonly used in the field of traffic safety. Using the NER method to extract key information related to traffic violations for analysis provides a certain level of technical support for traffic safety analysis.

This paper employs a comprehensive traffic safety service management platform to address the concerns of incomplete analysis and inability to identify effective named entities in the field of traffic safety connected to serious traffic violations. The study focuses on serious traffic violations in Beijing and analyses them using a BERT-CRF model. Firstly, the BIO annotation method is used to add event labels to traffic violation events, expanding the People's Daily dataset from 71,456 to 95,291 words. This is done to enhance the generalisation and learning capabilities of the BERT-CRF model through an expanded data set and compared to other models, the average accuracy of the BERT-CRF model grew by 2.04%, and the average F1 score increased by 1.64%, allowing it to accomplish the task of extracting named entities of traffic violations more effectively. Secondly, methods such as forward geocoding and Bayesian formulas are applied to enhance the data on severe traffic violations, acquiring features such as longitude, latitude, weather, and day of the week, thus providing technical support for the governance of urban traffic violations. Lastly, a precise analysis is conducted from multiple dimensions including month, week, hour, gender and space on six categories of severe traffic violations: forged or altered motor vehicle driving license, drunk driving, hit-and-run, driving after drinking alcohol, no valid motor vehicle driving license and speeding 50% over. This can provide support for the dynamic allocation of law enforcement personnel on urban traffic scenes and enabling precise governance of traffic violations.

## 2. METHODS

By observing and pre-processing the collected dataset on severe traffic violations, we employed a series of methods, including the BERT-CRF model, forward geocoding, Beautiful Soup method and Bayesian formula, to enhance data quality. The overall process of data acquisition is illustrated in *Figure 1*, with a key emphasis on the application of the BERT-CRF model. The main steps are described and illustrated in the following subsections.

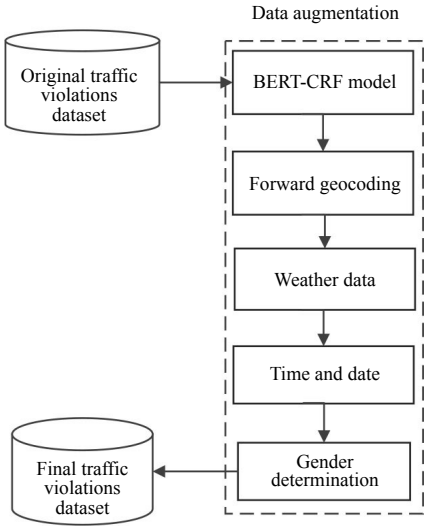


*Figure 1 – The process of enhancing data on traffic violations*

## 2.1 The original traffic violations dataset

Beijing is in northern China, north of the North China Plain, with Tianjin City to the east and Hebei Province to the west. Beijing is a world-famous ancient capital and a modern international city which has a semi-humid and semi-arid monsoon climate, with hot and rainy summers, cold and dry winters, and short spring and autumn seasons. It includes the Dongcheng District, Xicheng District, Chaoyang District and other 16 districts. This paper primarily focuses on the records of severe traffic violations in Beijing (https://bj.122.gov.cn/) and analyses a dataset comprising 62,886 instances of traffic violations recorded from 1 January 2016 to 31 December 2021 [17].

The 16 different fields that make up the original traffic violation data include the vehicle plate type, vehicle plate number, fine fact, penalty type, penalty result, etc. Since the original traffic violation data included sensitive information including the driver's name and legal name, the data was desensitised by replacing it with numbers. Because there are so many serious traffic violations every day, in order to address and record them quickly, only important information such as the driver's name, license plate type, case name, penalty category, penalty result and penalty fact are usually recorded, but this information does not include the weather, longitude and latitude information where the violations occurred and the case name is not as clear as the description of the penalty facts.

The case name of serious traffic violations are split into six primary categories based on data observation and statistics, to define the case name types in *Table 1* as standards.

Using these six categories of serious traffic violations as the study target, this study gets weather, latitude and longitude information from traffic violations through data augmentation and examines traffic violations in Beijing from time, place and weather perspectives.

*Table 1 – The type conversion of serious traffic violation cases*

| Types of traffic violation | Traffic violation case identification |
|---|---|
| Forged or altered motor vehicle driving license | 0 |
| No valid motor vehicle driving license | 1 |
| Driving after drinking alcohol | 2 |
| Hit-and-run | 3 |
| Drunk driving | 4 |
| Speeding 50% over | 5 |

## 2.2 Data augmentation

*BERT-CRF model*

For traffic violation data, the key entities are the name, time and place of the traffic violation. Since the People's Daily dataset used only contains four types of entity labels: personal name, time, place name and organisation name, it is impossible to extract the names of traffic violations. In order to enable the BERT-CRF model to extract the name of the traffic violation, the People's Daily dataset was expanded by using manual labelling data.

Location keywords and names of traffic violation were randomly selected from the original dataset of traffic violations, and 300 pieces of data were generated and labelled in batch according to the sentence pattern of "where and what illegal behaviours will be fined". The generated data was labelled with BIO labelling method, and the data set of People's Daily was expanded from 71,456 words to 95291 words.

There are three commonly used tagging strategies: BIO, BMES, and BIOES. This paper uses the BIO tagging method, where "B" indicates the beginning of an entity, "I" indicates the inside of an entity or the end of an entity, and "O" indicates an entity that is not of interest, *Table 2* shows the entity label corresponding to the BIO label.

*Table 2 – BIO label*

| Entity type | Start-tag | End-tag |
|---|---|---|
| EVENT | B-EVENT | I-EVENT |
| ORGANISATION | B-ORGANISATION | I-ORGANISATION |
| PERSON | B-PERSON | I-PERSON |
| TIME | B-TIME | I-TIME |
| LOCATION | B-LOCATION | I-LOCATION |
| NON-ENTITY | O | O |

BERT [18] can produce deeper semantic features and deep bidirectional language representations that combine left and right context information. The model structure can be changed to suit various downstream needs. The context is linked in this method and a more precise semantic representation of the text sequence characters is extracted. The CRF [19] can mark the labels of the characters in the text sequence and calculate the output to complete the text sequence prediction which is to solve the conditional probability distribution of the state sequence based on the observation sequence and form the random field of the state sequence.

Take "On 6 October 2020 at 18:34, a violation occurred in the section from Changqiao Hutong West Mouth of Deshenmen Inner Street to Xinghua Hutong West Mouth, where a vehicle was driven that did not match the types of vehicles specified on the driver's license." as an example; the identification process is shown in *Figure 2* to better illustrate how the BERT-CRF model works.

The BERT model, as shown in *Figure 2*, comprises of the embedding layer and the transformer layer, with the [CLS] identifier added at the beginning of the input sequence and the adjacent sequences separated by the identifier. Token embeddings, segment embeddings and position embeddings are all part of the embedding layer. Token embeddings are used to represent the vector of each character in the input sequence and derive the input sequence's word vector. Segment embeddings are used to convey the sentence's global semantic information and to obtain the sentence vector of the input sequence. Position embeddings can be used to encode the position information of characters in an input sequence in order to retrieve the input sequence's position vector. Combine these three vectors to create a 768-dimensional composite vector, then feed it into each transformer for learning via the multi-head attention mechanism, and then feed the learnt data into the CRF model. Each character's label is predicted and the conditional probability label is output using the BIO labelling method.

The key part of BERT is the multi-head attention mechanism in the transformer structure [20], which allows the model to learn different knowledge in different representation subspaces, forming information

with multiple dimensions. The calculation process of the mechanism is shown in *Formula 1 and 2.*

$$head_i = attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{1}$$

$$MultiHead(Q,K,V) = head_1 \oplus head_2 \oplus \ldots \oplus head_i \tag{2}$$

where $Q$ represents the query vector, $K$ represents the key vector, $V$ represents the value vector, $W_i^Q$ represents the parameter for the $i$-th round of linear transformation of the query vector, $W_i^K$ represents the parameter for the $i$-th round of linear transformation of the key vector, $W_i^V$ represents the parameter for the $i$-th round of linear transformation of the value vector, $\oplus$ represents concatenation.

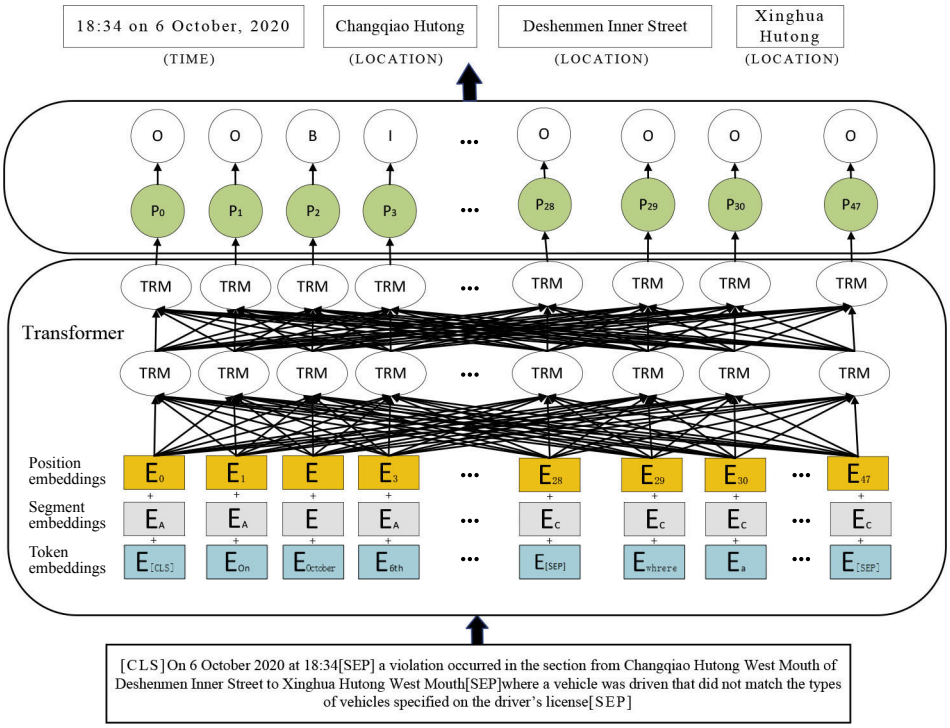


*Figure 2 – The named entity recognition process for traffic violations*

The word feature vectors created by the BERT layer are independent of one another, and when they are fed into the fully connected layer to forecast the labels of each word in the input sequence, the dependency link between word labels is ignored. The CRF model is used to find the globally optimal label sequence in order to tackle this challenge. As demonstrated in *Formulas 3 and 4,* CRF is a conditional probability distribution that can achieve the ideal anticipated sequence through the relationship between neighbouring labels.

$$S(x) = \sum_y \exp\left[\sum_k w_k\left(f_k(x,y)\right)\right] \tag{3}$$

$$P(y|x) = \frac{1}{S(x)} \exp\left[\sum_k w_k\left(f_k(x,y)\right)\right] \tag{4}$$

where $f$ represents the feature function, $W$ represents the weight corresponding to the feature function, $x$ represents the labelled observation sequence, $y$ represents the output sequence, $x$ represents the corresponding label sequence, $S(x)$ and represents the sum of the score values of all possible output sequences. During prediction, the Viterbi algorithm can be used to obtain the label sequence with the highest probability for a given observation sequence.

The primary metrics used to evaluate the model are precision in *Formula 5*, recall in *Formula 6* and F1 score in *Formula 7*. The F1 score is a model performance indicator that combines precision and recall rates.

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

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

$$\text{F1 - } score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{7}$$

Where *TP* indicates the number of correctly named entities identified, *FP* indicates the number of incorrectly named entities identified as correct, *FN* indicates the number of correctly named entities identified as incorrect.

The model training was done in the CUDA version 11.1, GPU RTX2060 and TensorFlow1.15 software environments, with the training set, validation set and test set divided at 8:1:1. *Table 3* displays the parameters chosen for model training.

*Table 3 – Model parameter values*

| Parameter | Values |
|---|---|
| max_seq_length | 32 |
| batch_size | 32 |
| learning_rate | 0.00005 |
| clip | 0.1 |
| warmup_proportion | 0.1 |
| dropout_rate | 0.2 |

In this study, the BERT model with 12 levels and 768 dimension hidden layers is employed, along with the 12-head mode's attention mechanism and a total of 110M parameters. Multiple tuning experiments were utilised to get superior training outcomes according to the values in *Table 3* in order to choose the optimum mix of parameters for training. *Table 4* displays the extracted entities using the trained BERT-CRF model.

*Table 4 – Annotation results of different entities on the BERT-CRF model*

| Entity | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|
| EVENT | 98.52 | 98.88 | 98.69 |
| LOCATION | 95.45 | 97.46 | 96.44 |
| PERSON | 94.90 | 98.68 | 96.75 |
| ORGANISATION | 71.58 | 90.67 | 80 |
| TIME | 94.62 | 91.79 | 93.18 |

Ablation experiments were also done on the proposed model to investigate the function of various components of the BERT-CRF model. The ablation studies were validated on the same dataset by keeping the CRF or BERT layer but adding the BiLSTM (Bidirectional Long Short-Term Memory) layer, and the findings are displayed in *Table 5*. According to the model's training findings, when compared to other models, the average accuracy of the BERT-CRF model grew by 2.04%, and the average F1 score increased by 1.64%, allowing it to accomplish the task of extracting named entities of traffic offenses more effectively.

*Table 5 – The results of different models*

| Model | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|
| CRF | 87.03 | 91.32 | 89.12 |
| BiLSTM-CRF | 86.06 | 91.32 | 88.61 |
| BERT | 86.37 | 92.50 | 89.33 |
| BERT-BiLSTM-CRF | 84.77 | 91.30 | 87.92 |
| BERT-CRF | 88.53 | 92.90 | 90.66 |

*Forward geocoding, weather data, time and date*

To analyse the frequency of traffic violations in various locations, obtain the latitude and longitude in-

formation of traffic violations by using the geocoding method provided by the Baidu Map Open Platform to identify the locations identified by the named entity; then draw a heat map and mark off the traffic network using Folium components to intuitively analyse the spatial law of traffic violations.

To analyse the frequency of traffic violations at different time points, the time identified by the named entity is standardised into the canonical format of date "YYYY-MM-DD" and "time (HH:MM:SS)"; then, according to the segmentation rules, extract the year, month, day and time of traffic violations, and find the implied rules of traffic violations at a different time granularity.

The standardised date is used to obtain the weather, wind direction and wind data of traffic violations from the Beijing Historical Weather Platform, and the average number of weather and wind values was used to fill in the inaccessible weather information.

*Gender determination*

According to ROLISON et al. [21, 22], the gender of the driver has an impact on traffic accidents, and the Bayesian formula can be used to predict the gender based on the name to improve the data characteristics of traffic violations. *Formula 8* depicts the Bayesian formula.

$$P(G \mid N) = \frac{P(N \mid G) \cdot P(G)}{P(N)} \tag{8}$$

where $P(G|N)$ represents the probability of identifying the gender based on the name, $P(N|G)$ represents the probability of the name appearing based on the known gender, $P(G)$ represents the probability of the gender appearing and $P(N)$ represents the probability of occurrence of a given name.

When the condition is independent, $P(N|G)$ can be transformed as shown in *Formula 9*.

$$P(N \mid G) = \prod_{i=1}^{n} P(N_i \mid G) \tag{9}$$

where $P(N_i|G)$ represents the probability of occurrence associated with the specified name.

## 3. THE ANALYSIS OF TRAFFIC VIOLATIONS

### 3.1 The analysis of time characteristics

The frequency of traffic violations varies depending on the time of day, including morning and evening rush hours, holiday adjustments and road conditions. Statistical analysis of traffic violations at the granularity of months, days and hours can help traffic managers formulate different treatment strategies at different time points by intuitively analysing the peak value of traffic violations.

According to *Figure 3*, the months with the greatest traffic infractions are October and November. In China, the statutory National Day holiday in October and the rest of the year will generally extend the holiday time to 7 days, and people usually travel to Beijing during the holiday, resulting in an increase in traffic flow in Beijing. On the other hand, in November, there is no statutory holiday and the month is longer, resulting

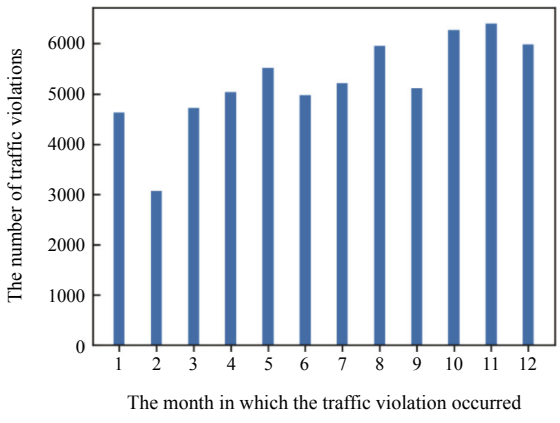


*Figure 3 – The number of traffic violations by month*

in more frequent commuting and an increase in traffic violations. Values 0–5 in *Figure 4* correspond to the types of traffic violations in *Table 1*. According to *Figure 4*, traffic violations involving forged or altered motor vehicle driving licenses are more common in October, November and December; traffic violations involving no valid motor vehicle driving license are more common in August, October and November; and traffic violations involving speeding 50% over are more common in February and April.

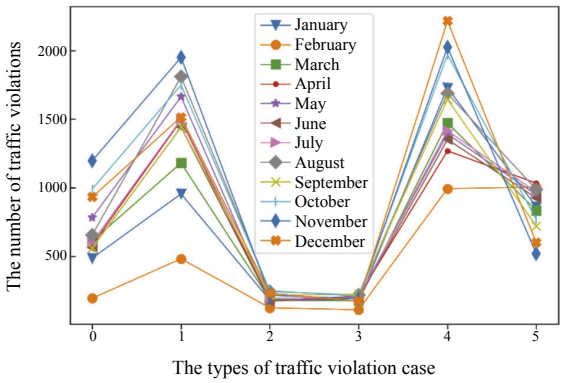


*Figure 4 – The number of traffic violation cases in different months*

*Figure 5* shows that more traffic violations occurred during the week on Thursdays and on Saturdays during the weekend; *Figure 6* shows that no valid motor vehicle driving license violations occurred on Tuesdays, Wednesdays and Thursdays; and drunk driving occurred on Tuesdays, Wednesdays and Thursdays. In China, there are seven days in a week, with five days for labour and two days for rest. On-site law enforcement is more prevalent on Monday, resulting in fewer traffic violations. People tend to go out on Saturday to unwind, which boosts traffic congestion and the number of traffic infractions compared to Sunday.

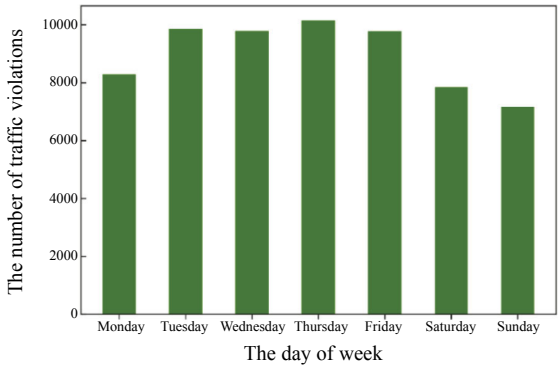


*Figure 5 – The number of traffic violations by day*

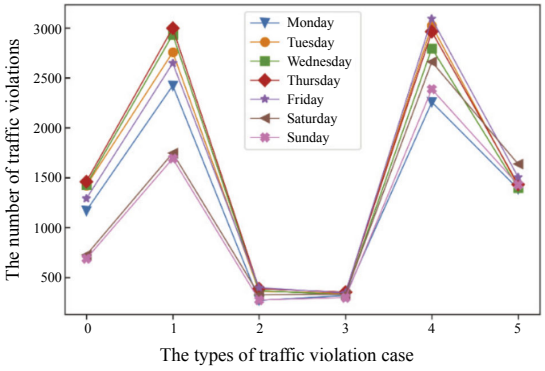


*Figure 6 – The number of traffic violation cases on different days*

*Figures 7 and 8* show that the distribution of hours with the most traffic violations on weekdays has two peaks (9:00 a.m. and 9:00 p.m.); the distribution of hours with the most traffic violations on weekends also

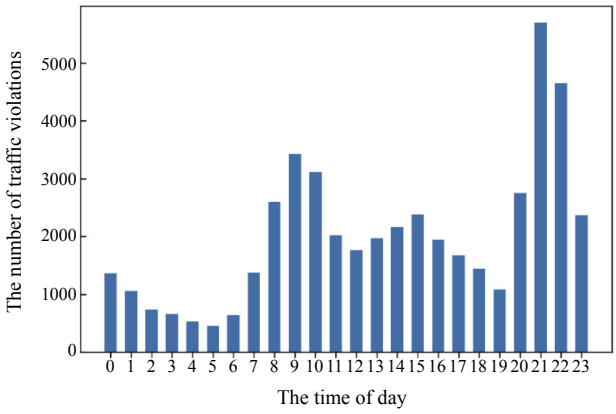


*Figure 7 – The number of traffic violations in 24 hours on working days*

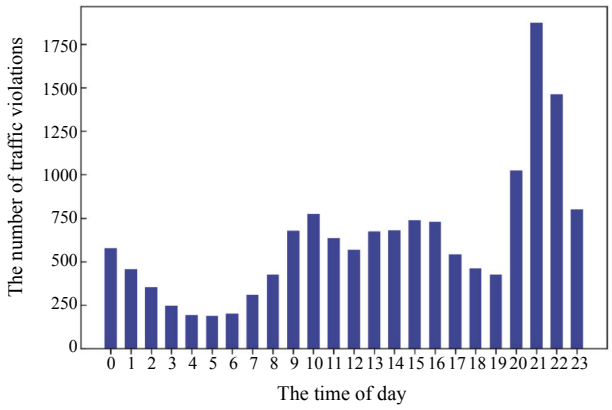


*Figure 8 – The number of traffic violations in 24 hours on weekends*

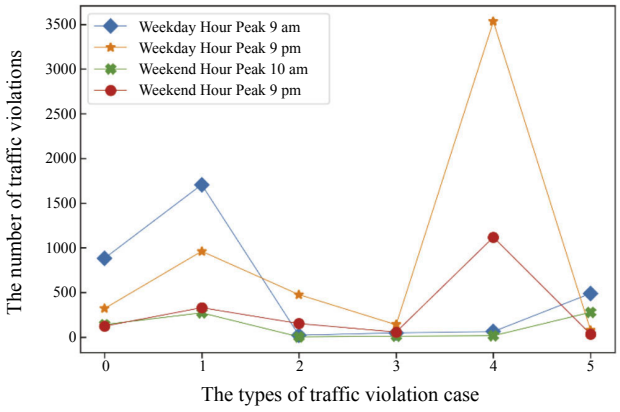


*Figure 9 – The number of traffic violation cases at the same time*

has two peaks (10:00 am and 9:00 pm). Further examination of the hours with the highest peak value in *Figure 9* reveals that traffic violations of no valid motor vehicle driving license occurred more frequently on weekdays at 9:00 a.m.; traffic violations of drunk driving occurred more frequently at 9:00 p.m.; and traffic violations of speeding 50% over occurred more frequently on weekends at 10:00 a.m.

### 3.2 The analysis of administrative district characteristics

The frequency of traffic violations varies by administrative region due to traffic facilities, residential population, schools, and so on. The frequency of traffic violations for each administrative region can be analysed, and traffic officers in different administrative regions can be assisted in developing different traffic strategies.

The Chaoyang and Haidian districts of Beijing have greater rates of traffic violations, owing to their dense populations where people live, work and go to school. Chaoyang District, known for business and commerce, has increasing traffic flow, but the Haidian District, which is home to a big number of elementary and secondary schools, has a huge number of younger drivers who engage in unsafe behaviour. As indicated in *Figures 10 and 11*, this combination of circumstances leads to a higher occurrence of violations such as driving without a valid license and drunk driving in these areas.

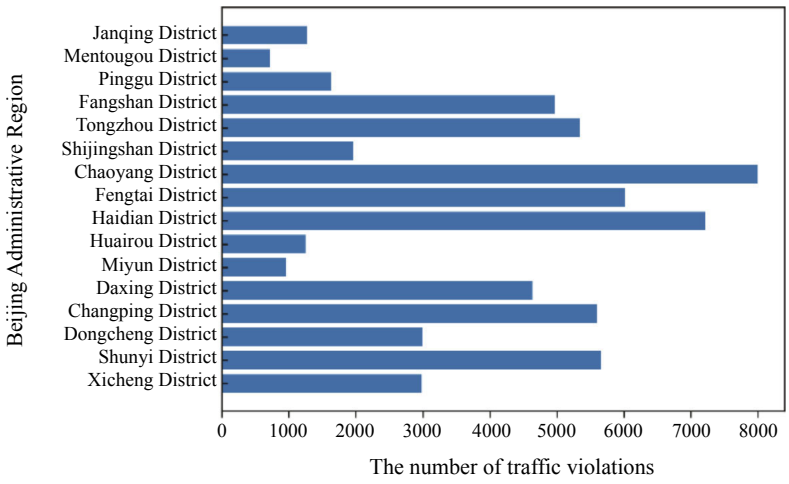


*Figure 10 – The number of traffic violations by administrative region*

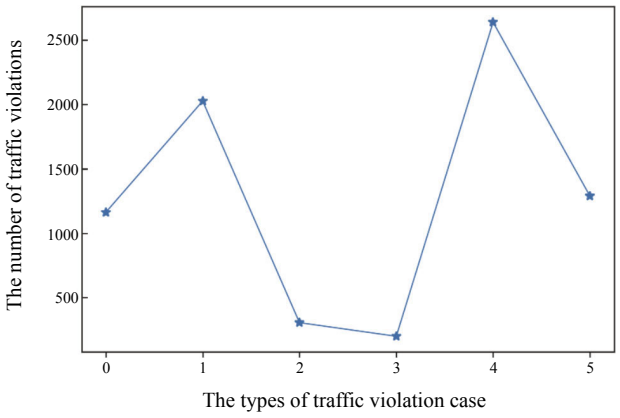


*Figure 11 – The number of traffic violation cases in Chaoyang District*

### 3.3 The analysis of weather characteristics

Weather conditions may contribute to traffic offenses. Weather reduces visibility, which can lead to major car accidents [23], and weather influences driving behaviour to some extent [24, 25]. Climate change will steadily worsen the living environment, necessitating an examination of weather conditions. A statistical analysis of traffic violations based on meteorological parameters is undertaken to validate the frequency of traffic violations under different weather conditions.

*Figure 12* shows that traffic violations occur more frequently in sunny and cloudy weather, which is related to the number of weather occurrences and driver psychology, indicating that a good driving vision effect will lead drivers to relax their vigilance but make them more prone to traffic violations. Further analysis of the sunny and thunderstorm weather types in *Figure 13* shows that while there are more occurrences of no valid motor vehicle driving license and drunk driving in sunny weather, the frequency of traffic cases with speeding 50% over is relatively high in thunderstorm weather. This demonstrates that drivers are more dangerous in sunny weather than in thunderstorm conditions.

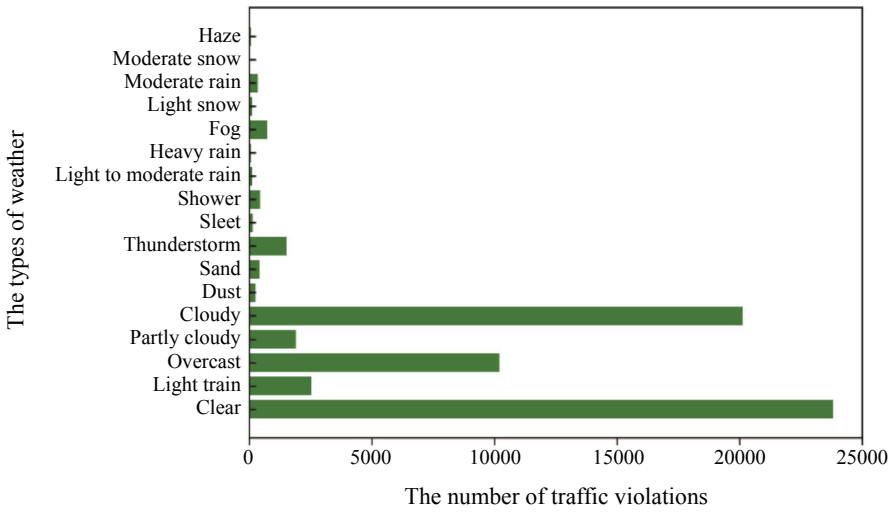


*Figure 12 – The number of traffic violations by weather*

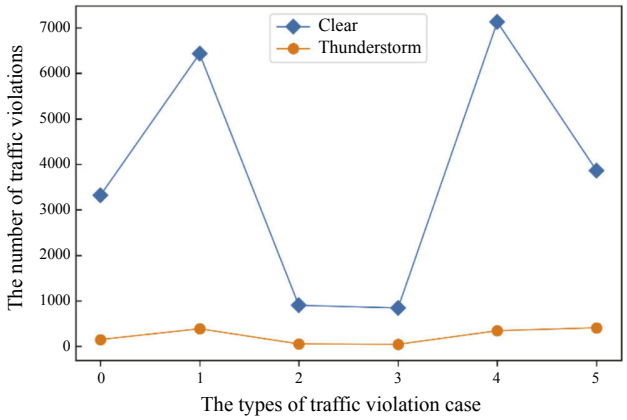


*Figure 13 – The number of traffic case types in clear and thunderstorms*

### 3.4 The analysis of gender characteristics

*Table 6* shows the number of traffic violations that occurred in Beijing based on a statistical analysis of traffic violations by gender. There are more traffic violations for driving without a valid license and drunk driving violations for men; for women, there are more traffic violations for speeding 50% over. It is further confirmed that male drivers drive drunk more frequently than female drivers [26], male drivers are more prone to risky driving, while female drivers need more attention [27, 28] to avoid traffic violations.

*Table 6 – The number of traffic cases by gender*

| Types of traffic violation | Gender | |
|---|---|---|
| | Male | Female |
| Forged or altered motor vehicle driving license | 7379 | 775 |
| No valid motor vehicle driving license | 15700 | 1483 |
| Driving after drinking alcohol | 2198 | 176 |
| Hit-and-run | 2060 | 247 |
| Drunk driving | 17556 | 1607 |
| Speeding 50% over | 7817 | 2366 |

### 3.5 The analysis of spatial features

Spatial feature analysis may quickly identify the geographical distribution of traffic offenses, and combining the geographical map and heat map can identify areas with a high frequency of traffic violations.

*Figure 14* depicts a statistical analysis of traffic violations in different administrative districts of Beijing in 2021 using a colour-patch map.

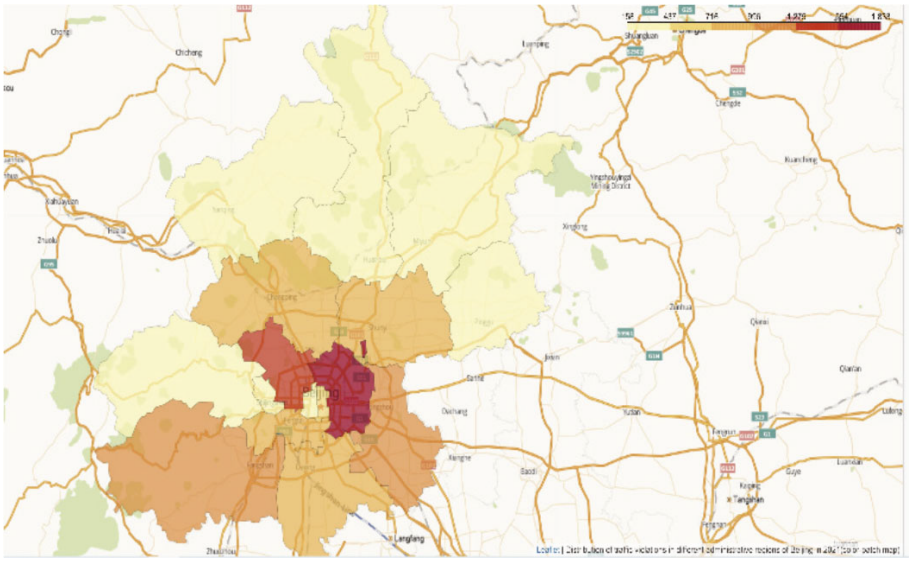

*Figure 14 – The number of traffic violations occurring in different districts of Beijing in 2021*

*Figure 14* demonstrates that the Chaoyang District and Haidian District had the most traffic violations in 2021, which is consistent with the pattern of administrative district traffic violations. In China, major medical, education, transportation facilities and other resources are more concentrated in the core area than in the peripheral area. Generally, people are more willing to live in the core area, resulting in more prominent traffic demand problems and a higher tendency of traffic violations. Heat maps and scatter plots were used to assess the street network of traffic violations in Chaoyang District in 2021. *Figure 15* depicts the heat map, whereas *Figure 16* depicts the scatter plot. Through heat maps and scatter diagrams, it can be seen that traffic violations are easy to occur in commercial centres, hospitals and other places, and the distribution of traffic violations around traffic hubs is wider, and it is easy to appear in the branch road connected with the main road.

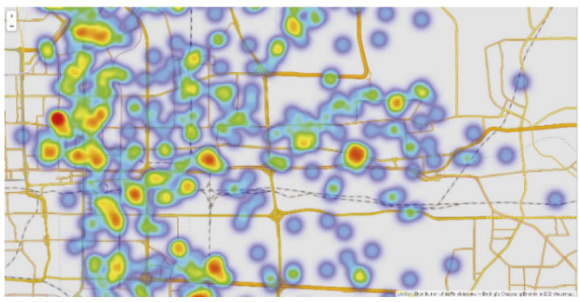

*Figure 15 – Heatmap of traffic violations in Chaoyang district of Beijing in 2021*

*Figure 16 – The scatter plots of traffic violations in the Chaoyang District of Beijing in 2021*

## 4. CONCLUSION

Based on the demand for governing urban traffic violations and focusing on the records of such violations, a city-specific traffic violation feature analysis method based on the BERT-CRF model is proposed. From the perspectives of a month, week, time, gender, space and other dimensions, the analysis of six types of traffic violations, such as forged or altered motor vehicle driving license, drunk driving, hit-and-run, driving after drinking alcohol, no valid motor vehicle driving license and speeding 50% over, provides decision-making reference for the allocation of traffic law enforcement. Weather factors and urban road network factors can also be combined to reveal the causes of traffic violations.

In this paper, the proposed method is put into practice based on the data of traffic violations recorded in Beijing from 2016 to 2021. The analysis results show that drunk driving has the highest frequency of traffic violations, followed by no valid motor vehicle driving license. The early peak of traffic violations on working days appeared at 9 a.m., while the early peak of traffic violations on weekends appeared at 10 a.m. The male drivers with no valid driver's license and driving a motor vehicle under the influence of alcohol comprised 59.46% and female drivers driving a motor vehicle under the influence of alcohol and speeding 50% over reached 57.11%. Extreme weather conditions are associated with the highest frequency of exceeding the speed limit by 50%, whereas non-extreme weather conditions correlate with a higher frequency of drunk driving violations. The traffic violations mostly occur on secondary roads that connect with main streets.

The above analysis can be used to guide the construction of a monitoring system for urban traffic violations and provide support for the governance of urban traffic violations. Due to the complexity in describing the locations of traffic violations, the next steps will delve into further research on the BERT-CRF model's recognition of compound or overlapping locations. This will involve an analysis of factors such as traffic facilities and road conditions to identify potential influences on traffic violation behaviour. Additionally, leveraging data mining methods will allow for a more in-depth analysis of traffic violations, leading to more targeted decision-making recommendations.

## ACKNOWLEDGMENTS

This work was supported by the Corresponding Topics of Beijing Science and Technology Plan (Grant No. Z211100004121010), and the by National Key R&D Program of China (Grant No. 2023YFC3306405). We are very grateful to Professor Shi Yun Tao's team for their strong support, which promoted the research of this study with the support of the information extraction technology foundation and fund projects.

## REFERENCES

[1] National Bureau of Statistics. China statistical yearbook. Beijing: China Statistics Press;2021. https://www.stats.gov.cn/sj/ndsj/2021/indexch.htm [Accessed 10th Sep. 2021].

[2] Yigitcanlar T, Li R, Inkinen T, Paz A. Public perceptions on application areas and adoption challenges of AI in urban services. *Emerging Science Journal.* 2022. DOI: 10.28991/ESJ-2022-06-06-01.

[3] Jamshid A, Kudratulla A, Ilkhomjon S. Method of analysis of the reasons and consequences of traffic accidents in Uzbekistan cities. *International Journal of Safety and Security Engineering.* 2020;10(4):483-490. DOI: 10.18280/ijsse.100407.

[4] Sorum NG, Pal D. Effect of distracting factors on driving performance: A review. *Civil Engineering Journal.* 2022. DOI: 10.28991/cej-2022-08-02-014.

[5] Abbas HA, Obaid HA, Alwash A. Enhanced road network to reduce the effect of (external – external) freight trips on traffic flow. *Civil Engineering Journal*. 2022. DOI: 10.28991/cej-2022-08-11-015.

[6] Wang C, Hu HT, Deng SH. Development of a knowledge base for reasoning penalty for traffic violations based on event evolutionary graph. *Journal of Transport Information and Safety.* 2022;40(1):36-44. DOI: 10.3963/j.jssn.1674-4861.2022.01.005.

[7] Zhao ZY, et al. Study on the method of identifying the characteristics of the traffic violation behavior based on the spatial and temporal hotspot analysis approach. *Journal of Geo-Information Science.* 2022;24(7):1312-1325. DOI: 10.12082/dqxxkx.2022.210599.

[8] Li YX, et al. Analysis and prediction of intersection traffic violations using automated enforcement system data. *Accident Analysis & Prevention*. 2021;162:106422. DOI: 10.1016/j.aap.2021.106422.

[9] Ji ZY, et al. Named entity recognition based on deep learning. *Computer Integrated Manufacturing Systems.* 2022;28(6):1603-1615. DOI: 10.13196/j.cims.2022.06.001.

[10] Xiao R, Hu FJ, Pei W. Chinese medicine text named entity recognition based on BiLSTM-CRF. *World Science and Technology-Modernization of Traditional Chinese Medicine.* 2020;22(07):2504-2510. DOI: 10.1155/2021/6696205.

[11] Chen ZL, Yuan F, Li XH, Zhang MM. Based on BERT-BiLSTM-CRF model the named entity and relation joint extraction of Chinese lithological description corpus. *Geological Review.* 2022;68(02):742-750. DOI: 10.16509/j.georeview.2022.01.115.

[12] Zhao PF, Zhao CJ, Wu HR, Wang W. Recognition of the agricultural named entities with multi-feature fusion based on BERT. *Transactions of the Chinese Society of Agricultural Engineering.* 2022;38(3):112-118. DOI: 10.11975/j.issn.1002-6819.2022.03.013.

[13] Zeng LL, Wang YS, Chen PF. Named entity recognition based on BERT and joint learning for judgment documents. *Journal of Computer Applications*. 2022;1-7. DOI: 10.11772/j.issn.1001-9081.2021091565.

[14] Liu F, Wen Z, Wu Y. Intelligent analysis on text of power industry accident based on BERT-BILSTM-CRF model. *Journal of Safety Science and Technology*. 2023;19(01):209-215. DOI: 10.11731/j.issn.1673-193x.2023.01.031.

[15] Wang ZH, et al. Intelligent recognition of key earthquake emergency Chinese information based on the optimized BERT-BiLSTM-CRF algorithm. *Applied Sciences*. 2023;13(5):3024. DOI: 10.3390/app13053024.

[16] Li SQ, Pang WT. Joint extraction method of entity and relation in maize breeding based on BERT-CRF and word embedding. *Transactions of the Chinese Society for Agricultural Machinery.* 2023;1-16. DOI: 10.6041/j.issn.1000-1298.2023.11.028.

[17] Li J, Shi YT, Li SQ, Wang Q. Construction of knowledge graph based on traffic violations in Beijing. *2022 4th International Conference on Intelligent Information Processing (IIP), 14-16 Oct. 2022, Guangzhou, China.* 2022. p. 113-116. DOI: 10.1109/IIP57348.2022.00030.

[18] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 2019, Minneapolis, Minnesota*. 2019. p. 4171-4186. DOI: 10.18653/v1/N19-1423.

[19] Lafferty JD, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*. 2001. p. 282-289. DOI: 10.1109/ICIP.2012.6466940.

[20] Vaswani A, et al. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Red Hook, NY, USA*. 2017. p. 6000-6010. DOI: 10.48550/arXiv.1706.03762.

[21] Rolison J, Regev S, Moutari S, Feeney A. What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records. *Accident Analysis & Prevention*. 2018;115:11-24. DOI: 10.1016/j.aap.2018.02.025.

[22] Regev S, Rolison J, Moutari S. Crash risk by driver age, gender, and time of day using a new exposure methodology. *Journal of Safety Research*. 2018;66:131-140. DOI: 10.1016/j.jsr.2018.07.002.

[23] Wu YN, Abdel-Aty M, Lee J. Crash risk analysis during fog conditions using real-time traffic data. *Accident Analysis & Prevention*. 2018;114:4-11. DOI: 10.1016/j.aap.2017.05.004.

[24] Theofilatos A, Yannis G. A review of the effect of traffic and weather characteristics on road safety. *Accident Analysis & Prevention*. 2014;72:244-256. DOI: 10.1016/j.aap.2014.06.017.

[25] Xing F, et al. Hourly associations between weather factors and traffic crashes: Non-linear and lag effects. *Analytic Methods in Accident Research*. 2019;24:100-109. DOI: 10.1016/j.amar.2019.100109.

[26] Oppenheim I, Oron-Gilad T, Parmet Y, Shinar D. Can traffic violations be traced to gender-role, sensation seeking, demographics and driving exposure?. *Transportation Research Part F: Traffic Psychology and Behaviour.* 2016;43:387-395. DOI: 10.1016/J.TRF.2016.06.027.

[27] Ozkan T, et al. Cross-cultural differences in driving skills: A comparison of six countries. *Accident Analysis & Prevention.* 2006;38(5):1011-1018. DOI: 10.1016/j.aap.2006.04.006.

[28] Li XM, Oviedo-Trespalacios O, Rakotonirainy A, Yan XD. Collision risk management of cognitively distracted drivers in a car-following situation. *Transportation Research Part F: Traffic Psychology and Behaviour.* 2019;60:288-298. DOI: 10.1016/J.TRF.2018.10.011.

李杰，史运涛（通讯作者），李书钦

基于BERT-CRF模型的北京市交通违法行为分析

摘要：

交通违法行为是引发交通事故的重要原因，现有的研究对交通违法行为的分析不够全面且命名实体方法无法从交通违法行为记录中抽取到交通违法行为名称，无法为

城市交通违法行为的治理提供参考。通过将人民日报数据集从71456词扩展到95291词，使基于转换器的双向编码表征-条件随机场（BERT-CRF）模型取得了88.53%的准确率、92.90%的召回率和90.66%的F1值，识别了交通违法行为中的事件、时间和地点命名实体，并通过前向地理编码和贝叶斯公式对交通违法行为数据进行增强，从时间、空间、行政区域、性别和天气等多维度对交通违法行为进行分析，为交通执法现场的动态调配和交通违法行为的精准管理提供支持。

关键词：
交通违法行为；交通事故；命名实体；人民日报；BERT-CRF;贝叶斯公式