International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 11 Nov 2023, pp: 468-476 www.ijaem.net ISSN: 2395-5252

The Development Trend of 5G Technology Application in Land Transportation by Text Mining

Huang, Kuan-Wei, Wu, Meng-Yue, Wang, Chien-Hua
Business School, Lingnan Normal University, Zhanjiang, China
Business School, Lingnan Normal University, Zhanjiang, China
, Business School, Lingnan Normal University, Zhanjiang,China

Date of Submission: 01-12-2023 Date of Acceptance: 10-12-2023

ABSTRACT： With the continuous improvement of mobile communication technology, the research of the fifth generation communication technology (5G) in the field of land transportation has also made rapid development. Thus, to explore the development trend of 5G technology application in land transportation is an important issue for this study. This study is divided into two technical routes: (1) construct Word2Vector and BPN named WVBP model to judge whether key words belong to the field of 5G land transportation; (2) construct TF-IDF and decision tree named TIDT mode; to explore the “core words” of 5G applications in land transportation. The accuracy of the two models is 92%; According to the decision tree model, “on-board”, “dispatching”, “rail transit” and “automatic driving” are the core words of 5G in the field of land transportation, and the key development trends are “5G on-board terminal technology”, “intelligent dispatching in the field of 5G urban public transport”, “intelligent and automatic rail transit” and “automatic driving 5G cloud computing”. The contribution of this study is to mine, analyze and summarize the important information of the development trend of 5G land transportation from a large number of literature and it would be exanimated that the further research according our results in 5G land transportation.

Keywords： Text mining; WVBP; TIDT; 5G Land transportation

I. INTRODUCTION

The Chinese government has announced lots of policies to promote the 5G in land transportation. The “Regulations of Shenzhen Special Economic Zone on the Administration of Intelligent Connected Vehicle” issued by the Shenzhen Municipal People's Congress directly contributed to the transition from road testing to commercialization of intelligent connected vehicles. In the field of public land transportation, in 2019, with the support of the Henan Provincial Government and China Mobile's 5G technology, the Wisdom Island 5G smart bus started its first departure. This is the world's first driverless bus line operating on open roads [1]. In 2020, the Outline of the Railway First Plan for a Transportation Powerful Country in the New Era clearly stated that a modern railway network will be built by 2035. The railway train will have a “Super Brain” led by technologies such as 5G communication and BeiDou navigation [2].

To achieve greater success in the urban transport equipment market and in the field of public land transport, it is necessary to seek developments in the field of land transport. Simultaneously, the arrival of 5G is a brand new intelligent world. A smart city is a major product in the 5G era, and it will shape a city that operates intelligently and rationally in all aspects. In addition, the high speed and low latency of 5G technology can lay a solid foundation for the development of the transportation industry [3]. In the field of urban transportation, urban transportation equipment is increasingly connected to the Internet. According to the "Position paper 5G Applications" [4], it is pointed out that the Internet of Vehicles technology will be one of the ten application scenarios in the 5G era.“Position paper 5G Applications” pointed out that the internet of vehicles technology will be one of the ten application scenarios in the 5G era [5].

Research on 5G applications in the land transportation has emerged one after another, and the development of various sub-fields in the land transportation is also in full swing. The relevant academic literature is an important data support; generally speaking, the content of academic literature is very lengthy, and the important information in it is widely distributed. In order to discover the development trend in the field of 5G land transportation in a large number of academic literature, it takes a huge time cost and energy to rely on manual reading and summarization. Text mining is the process of editing, organizing and analyzing the specific huge amount of text collected, in order to discover the implicit correlation features or interesting and novel patterns, and provide effective and key information for analysts or decision makers [6]

[7]; the specific process is mainly to mine unstructured or semi-structured file data, find out the rules and structures hidden in the word model of text data, and then automatically from the huge amount of data. Common and frequently occurring terms and keywords are identified, and their relevance is then explored and analyzed [8].

Besides, applying machine learning techniquesto deal with different classification problemshave advantages and disadvantages. The advantage of using artificial neural network such as Back-propagation network (BPN) that can quickly divide the words related to land transportation into land transportation and not land transportation, but it cannot dig out the “core words” and their importance as well as the set of parameters is not easy to find; on the contrary, using decision tree for classification can effectively screen and mine “core words” to achieve knowledge discovery and find trend about 5G applications in land transportation. How to mine valuable key information from a large amount of academic literature, and exploring the important development trends of 5G in various sub-fields of land transportation is the purpose of this study. Therefore, this study proposes a two text mining models to explore the application trends of 5G technology in land transportation and expect that guide to the directions of further sustainable research development from these core words.

II. LITERATURE REVIEW

2.1 The development status of 5G land transportation

As of 2019, the number of patents related to 5G in the field of Internet of Vehicles total 30,447. Additional, according to the latest research report by Counterpoint IoT Server, the global connected car market is expected to grow by 270% by 2022[5]. In domestic, the "Regulations of Shenzhen Special Economic Zone on the Administration of Intelligent Connected Vehicle" issued by the Shenzhen Municipal People's Congress directly contributed to the transition from road testing to commercialization of intelligent connected vehicles [6]. In the field of public land transportation, in 2019, with the support of the Henan Provincial Government and China Mobile's 5G technology, the Wisdom Island 5G smart bus started its first departure. This is the world's first driverless bus line operating on open roads [7]. In 2020, the Outline of the Railway First Plan for a Transportation Powerful Country in the New Era [8] clearly stated that a modern railway network will be built by 2035. The railway train will have a "Super Brain" led by technologies such as 5G communication and BeiDou navigation.

2.2 Text mining

With the vigorous development of the Internet and the development of information technology, the traditional written data has been transformed into an electronic form, and the acquisition of data is much easier than in the past, and most of the information can be obtained online. However, these text data are usually stored in semi-structured or unstructured form, and their content has no certain format. There is no common structure between each data, and no clear definition of its classification. However, these huge amounts of information can no longer be handled by manpower alone. Computer computing and processing capabilities must be used to extract the required information from tens of thousands of large amounts of data.

The text mining is the process of editing, organizing and analyzing the specific huge amount of text collected, in order to discover the implicit correlation features or interesting and novel patterns, and provide effective and key information for analysts or decision makers [9-10]; the specific process is mainly to mine unstructured or semi-structured file data, find out the rules and structures hidden in the word model of text data, and then automatically or semi-automatically from the huge amount of data. Common and frequently occurring terms and keywords are identified, and their relevance is then explored and analyzed [11]. Text mining is not just a technology, but a combination of research on multiple professional technologies, using self-talk language processing tools to extract text patterns in unstructured documents, and automatically select unknown and useful hidden information from documents, which is beneficial to The development of knowledge exploration [12].

Many domestic scholars have also conducted research in this field. Wang [13] analyzed the difficulties of Chinese word segmentation, outlined the main published word segmentation schemes and developed a word segmentation system. Yin [14] used neural network technology for automatic word segmentation. The self-learning and nonlinear processing method of neural network makes it have the characteristics of fast word segmentation, accurate word segmentation, and high acceptance of new words. Cui [15] combined artificial neural network and text mining methods, took a large number of academic documents as the research body, and realized the intelligent identification of academic document classification.

2.3Word segmentation

Text word segmentation technology is an important part of Chinese text processing. According to a certain rule and algorithm, the words in the text are segmented to form a word list. The subsequent analysis and understanding of the text is based on the word segmentation list [16]. Text segmentation is an indispensable step before TF-IDF calculation and word vector conversion.

In addition, JIEBA word segmentation is a widely used word segmentation tool with good word segmentation effect. It is an open source word segmentation tool and realizes efficient word map scanning based on prefix dictionary. For unregistered words, the HMM model based on the ability of Chinese characters to form words is adopted, and the Viterbi algorithm is used [17].

2.4Term frequency-inverse document frequency, TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a weighting technique for text mining and information gathering [18]. The TF-IDF technique is used to evaluate the importance of words to a document set. The term frequency (TF) represents the frequency of a particular word in the document, and the inverse document frequency (IDF) is used to evaluate the generality of the word to the corpus. The basic idea is that if a certain word appears more frequently in this document and rarely appears in other documents, its word can be regarded as an important representative word of the document in which it is located. The specific calculation process is as follows:

TF = Occurrences in a certain category of terms / The number of all terms in this class (1)

IDF = log(The total number of documents in the corpus / The number of documents containing the term+1) (2)

$$TF\text{ - }IDF = TF \times IDF \tag{3}$$

2.5ArtificialNeural network

Artificial neural networks(ANNs), which is usually called neural networks (NNs), is a family of models which is inspired by biological neural networks. The kind of models can have more complex structure than conventional models such SVM, random forest, etc., and they are assumed to have more powerful ability to handle the difficult problem in the situation of sufficient training data.

Many scholars for domestic have in-depth discussions in related fields. He et al. [19] use NNs combined with LDA topic model to perform sentiment analysis and topic word extraction of Shanghai Disney satisfaction survey research.

Back-propagation network (BPN) is one of the base-known multiplayered feed-forward neural networks. The back-propagation learning algorithm has reawakened the researchers in variety of scientific and engineering areas in neural networks. The back-propagation learning is a gradient-descent-type learning algorithm in which the error is propagated backward to adjust the connection weights of preceding layers and to minimize output error. The error function E_p is normally defined as the sum of squared output error,

$$E_q = \frac{1}{2}\sum_{i=1}^{N}(t_i^p - O_i^p)^2 \tag{4}$$

where t_i^p and O_i^p are the desired and actual output of the i^{th} output node for the p^{th} training pattern, respectively. N represents the total number of output nodes. Accord to the gradient-descent method, the adjustment of each weight is proportional to the negative of the gradient of E_p :

$$\Delta w_{ij} = -\eta \frac{\partial E_p}{\partial w_{ij}} \tag{5}$$

where$w_{ij}$is the connection weight between the j^{th} neuron of the $(k-l)^{th}$ layer and the i^{th} neuron of the k^{th} payer, and η is the learning rate. BPN has the abilities of self-learning, massively parallel construction, high memorial capacity, generalization, fault tolerance, robustness and noise insensitive, and among many others. It has been applied extensively in many fields, in this study, BPN will be applied to construct text classifier judgement.

2.5 Decision tree

Decision tree is a supervised machine model with simple intuition and high execution efficiency. It can be used to solve classification and regression problems, and the output results can be easily explained through the structure of the model. A decision tree is constructed into the suitable tree from the training set in terms of the divide-and-conquer strategy. The first step of building a decision tree is to select the appropriate splitting attribute as the root node and to make the corresponding branch for each possible value. The learning process of decision tree divides the search space into two or more subsets each subset can be regarded as a new and simpler learning problem and repeated recursively to divide until the stopping criterion is satisfied.

Decision trees can not only be used to make classification judgments, but the decision nodes and pruning functions on the tree can help researchers find key attributes in the data set. Therefore, this study uses decision trees to find important development trends in the field of 5G land transportation. The generation process of decision trees is as follows:

(1)The optimal condition attribute is selected as the split node of the decision tree, and the split node of the training set is divided.

(2)Combined with recursive algorithm, the subset is divided until no new nodes are generated.

(3)Determine the category of leaf nodes.

2.6 Model evaluation

When judging the accuracy of model classification results, confusion matrix is often used to assist in judging the model classification effect. Confusion matrix is an auxiliary tool for judging the classification effect of supervised learning with a specific matrix [22]), False Positive (FP), True Negative (TN) and False Negative (FN) are the basic values of confusion matrix judgment. (see Table 1)

Table 1 Confusion matrix and value

Confusion matrix		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	negative	FP	TN

Using the confusion matrix and its basic values, indicators such as precision rate, recall rate, and F_1 score can be constructed to measure the effectiveness of the classification model.

Precision rate is the result of prediction, which refers to the accuracy of the true positive class in the predicted positive sample. Its formula is:

$$precision = \frac{TP}{TP+FP} \quad (6)$$

The recall rate is to analyze the predicted samples, which refers to how many positive classes in the samples are accurately predicted. Its formula is:

$$recall = \frac{TP}{TP+FN} \quad (7)$$

The F_1 score is an indicator that combines precision and recall to judge the effect of binary classification. Its essence is the harmonic average of model precision and recall, and its value is between 0 and 1. When the F_1 score value is larger, it can be considered that the effect of the classification model is better. Its formula is:

$$F_1 = \frac{precision \times recall}{precision + recall} \times 2 \quad (8)$$

III. RESEARCH DESIGN

3.1 5G research route

This study focuses on exploring the important development trends of 5G in the field of land transportation. The main implementation steps are as follows:

(1)It is the initial text data for collecting "5G land transportation t8 field". The initial text corpus of this study was constructed by downloading academic literature related to "5G land transportation field" and integrating them into text files through PDF parsing technology.

(2)It is data pre-processing for text data. Including data cleaning, removing stop words, JIEBA word segmentation and other experimental steps, the text data is preliminarily de-noised and the text word segmentation is performed to prepare for the transformation and modelanalysis of word vectors.

(3)Data reduction. This research uses "5G technology" as the search term for retrieval, uses crawler technology to crawl the keywords of literature related to "5G technology", and performs text matching with the word-segmented data set to further improve the data quality.

(4)TF-IDF calculation. The TF-IDF algorithm is used to calculate the TF-IDF weight value of each word under each document after data dimensionality reduction and the TF-IDF weight value of each word after document aggregation.

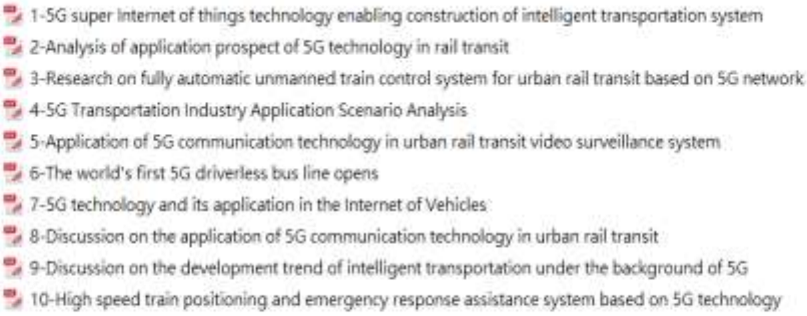
(5)It is the construction of word vectors. Because computers cannot process natural language directly. Therefore, it is necessary to use Word2Vec to convert words to word vectors to obtain word vector data that can be used for BPN model training.

(6)It is the model training data. This study mainly uses the BPN model and the decision tree model. Different models require different data conditions. Therefore, it is necessary to reconstruct the data by combining TF-IDF weights and word vectors under Word2Vec to obtain model training data that can be used in their respective models.

(7)It is to build a training model. By building a BPN model to realize a text classifier to judge whether a word belongs to the "5G land transportation field", and building a decision tree model to realize key information mining through important features of the decision tree, infer the important development trend of 5G in the land transportation field.

3.2 Data set collection

Based on related literatures such as "5G Land Transportation", this study uses text mining methods to mine important information and development trends in the field of 5G land transportation in the literature. So, this study will use "5G Land Transportation" as the search term to retrieve 70 academic literatures on 5G land transportation in CNKI in the past five years (see Figure 1) as the text data database, and construct the text data set.



1-5G super Internet of things technology enabling construction of intelligent transportation system
2-Analysis of application prospect of 5G technology in rail transit
3-Research on fully automatic unmanned train control system for urban rail transit based on 5G network
4-5G Transportation Industry Application Scenario Analysis
5-Application of 5G communication technology in urban rail transit video surveillance system
6-The world's first 5G driverless bus line opens
7-5G technology and its application in the Internet of Vehicles
8-Discussion on the application of 5G communication technology in urban rail transit
9-Discussion on the development trend of intelligent transportation under the background of 5G
10-High speed train positioning and emergency response assistance system based on 5G technology

Fig 1 Part of the PDF document dataset

3.3 Data Pre-processing

In this study, a large number of relevant documents are constructed as the initial data of the research, and the text data has text redundancy due to the influence of numbers, symbolic formulas and other factors. The quality of the data often determines the quality of the experiment and affects the results of the experiment. In order to ensure the quality of text data and prepare for subsequent research, it is necessary to clean the initial data to achieve the purpose of removing irrelevant data and achieving high-quality text data.

Next, there are still a large number of stop words in the cleaned data. If these stop words are directly brought into the model training, it will inevitably have a negative impact on the training results. Therefore, this study selected the stop words database of Harbin Institute of Technology and Chinese stop words. The mixed stop word database of the thesaurus performs stop word processing on the text data.

In the word segmentation part, this study calls the JIEBA word segmentation tool library. The precise mode in JIEBA word segmentation is selected. This mode, on the premise of ensuring the effect of word segmentation, well retains the context of the context, and ensures the effect of subsequent word vector conversion and model analysis.

After the above operations, the text dataset of this study is divided into 97879 words as shown in Table 2).

Table 2 Text dataset after data pre-processing

No.	Index_PDF file	word
0	1	Volume
11	1	Month
51	1	Things
52	1	Internet
53	1	Internet of things
⋮	⋮	⋮
163170	70	Enterprise
163171	70	Management

3.4 Data reduction and TF-IDF calculation

After word segmentation, the text dataset has formed a vocabulary that stores a large number of words, but it is found that even after the stop words are removed, there are still noise words such as "volume" and "month period", which are close to ten. The vocabulary size of 10,000 will have an excessive impact on the data dimension of the model training, which will cause difficulties for subsequent analysis.

In order to ensure the high quality of the data, we searched with "5G technology" as the search term, crawled the keywords of 1000 literatures related to "5G technology" and imported them into the TXT file. The text dataset is used for vocabulary matching, and a total of 475 words are filtered out as the final sample dataset (see Table 2).

Table 2 Partial final text data display

5G	Communication	Network	System
Traffic	Technology	Data	Development
Intelligent	City	Business	Smart
scenario	Automatic drive	subway	Carrierwave

3.5 Construct the training dataset

When the supervised learning model is trained, it is necessary to label the words in the sample training set as the expected value of the sample data.

3.5.1 BPN training dataset

In this study, the median of the TF-IDF weight value of the sample is calculated, and the words with the TF-IDF weight value greater than the median value are marked with the tag value "5G traffic", and the words with the weight value lower than the median value are marked with the tag value "Other" (See Figure 3). Finally, there are 238 words labeled "5G transportation", and 237 words labeled "other". This sample data will be used for BPN model training to construct a text classifier that can distinguish whether it is related to the 5G land transportation field.

Fig 3 Part of the BPN training data label display

Speed	5G transportation	Military	Others
Car	5G transportation	Cybertimes	Others
Route	5G transportation	Network application	Others
Way	5G transportation	Publish	Others
Trend	5G transportation	Internet Adoption	Others
Empowerment	5G transportation	Network system	Others
Road	5G transportation	Try	Others
Automobile	5G transportation	Enhancement	Others
Track traffic	5G transportation	Logic	Others
Interconnection	5G transportation	China Telecom.	Others
Traffic	5G transportation	Economic benefits	Others
Artificial intelligence	5G transportation	Integration	Others
Transportation	5G transportation	China Mobile	Others
Ops	5G transportation	China Unicom.	Others
Auxiliary	5G transportation	Cost-Saving And Profit-Increasing	Others
Algorithm	5G transportation	reduce distractions	Others
Dynamic	5G transportation	Array	Others
Real time	5G transportation	Topic	Others
Index	5G transportation	Cloudification	Others
Digital	5G transportation	Cloud	Others
Surface	5G transportation	Crossing	Others
Driverless cars	5G transportation	Medicine	Others

3.5.2 Decision tree training dataset

While BPN can be used for text mining, it cannot be used to discover key trends in 5G land transportation due to the black-box testing nature of neural networks. Therefore, this study will use the decision tree model to mine valuable information.

The data set of BPN is based on a word-label as the basic unit. In order to use the decision tree to explore the development trend of 5G in the field of land transportation, we need to reconstruct the data set based on the sample data set. In this study, 70 academic documents related to 5G land transportation are labeled with the value of "5G transportation", and an additional 40 academic documents related to 5G in other fields (including 5G medical care, 5G economics, 5G military, etc.) are selected and labeled with the value "Other". Then calculate the TF-IDF weight value of each document vocabulary separately, connect all documents, fill the empty value with 0, and construct a document with a label as the basic unit and vocabulary as a conditional attribute. Decision tree training dataset to use for decision tree training.

IV. EMPIRICAL ANALYSIS

4.1 Construct WVBP model

In this study, the TF-IDF algorithm is used to obtain the class label of the vocabulary, and based on the training data of BPN, a classifier that can judge whether the vocabulary belongs to the 5G land transportation field is constructed.

Disorder the data order in the training data of BPN, select 70% of the data as the training set, and 30% of the data as the test set, then build the BPN model (see Table 3 for specific model parameters), and verify the test set the F1 score.

Table 4 BPN model parameters

Layers	Number of Neuron	Activation Function	Learning rate
Input	385	$f_{relu}(x^{m}) = \max(0,$	0.01
Hidden	28	$f_{relu}(z^{m}) = \max(0, z^{m}$	
Output	2	$f_{softmax}(y^{m}) = \frac{e^{y^{m}}}{\sum_{j} e^{y_j^{m}}}$	

4.2Construct TIDT model

Decision tree can not only be used to solve the implementation problem of binary classification, but its functions such as nodes and pruning can help researchers find the optimal features of data. Therefore, this study uses the characteristics of decision trees to reconstruct the data set, taking documents as the basic unit of training, and vocabulary as the conditional attribute of training. Build three decision trees under the three algorithms ID3, C5.0 and CHAID, and summarize the model performance under the three decision trees. The is to find the optimal characteristics (i.e., core words) of the decision tree to infer the use of 5G in land transportation.

4.3 Research results

4.3.1 WVBP model results

The constructed model structure and the prepared data set were trained for 30 times, and the average F_1 score of the WVBP model was calculated through the precision rate and recall rate of the WVBP model is 0.92 (See in Figure5). It can be seen that WVBP model has a good effect on the classification of vocabulary in the field of 5G land transportation, and the classifier can be used to judge whether a vocabulary belongs to the key vocabulary in the field of 5G land transportation.

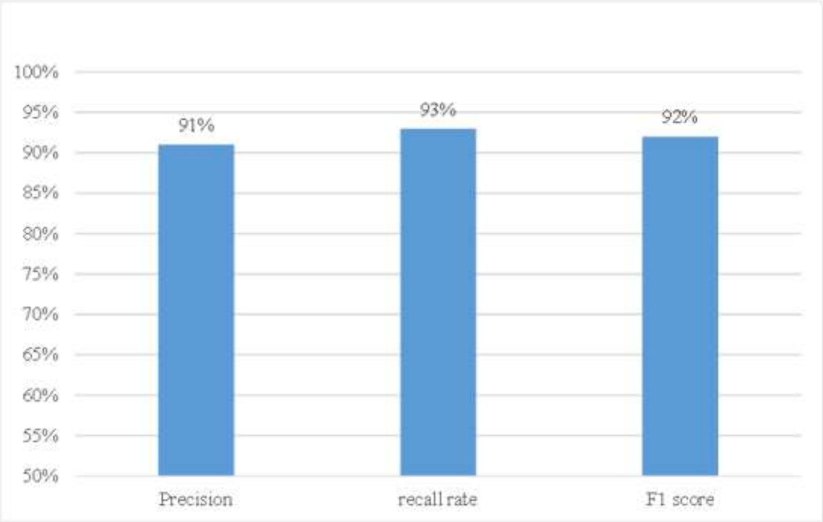


Fig 1 WVBP model evaluation results

4.3.2 TIDT model results

The training set was imported into the TIDT model for training, and the training was performed 30 times. The average F_1 score was calculated to be 0.92 (See in Figure 6)

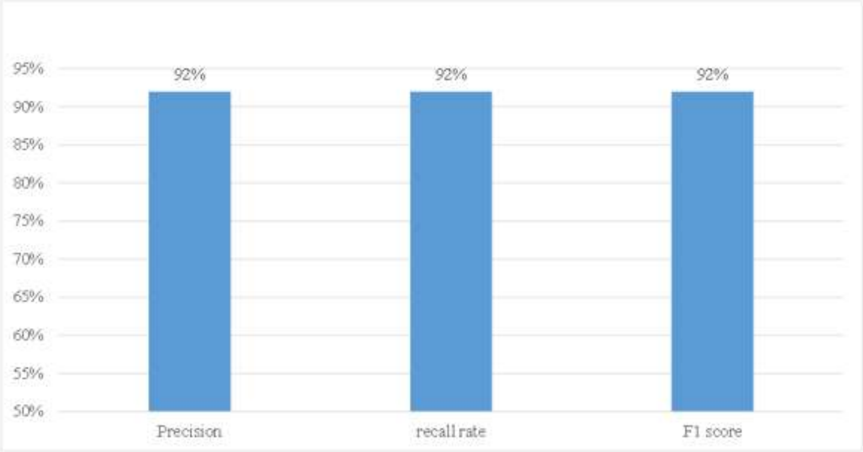


Fig 2TIDT model evaluation results

Comprehensive analysis shows that the words 'vehicle', 'dispatching', 'rail transit', and 'autonomous driving' that appear under the three different algorithms of ID3, C5.0 and CHAID are the core words of 5G in the field of land transportation (see Table 4). By searching the Internet for "5G+transportation +vehicle\ dispatching\rail transportation\ autonomous driving", the development trend of 5G in the field of land transportation is summarized (see Table 5 below).

V. CONCLUSIONS

With the rapid development of 5G technology, its characteristics of high speed, low latency and large bandwidth have injected fresh vitality into many fields, among which the development of 5G land transportation will become a key factor in changing people's lives and travel. This research uses the academic literature in the field of 5G land transportation as a dataset, constructs a text classifier in the field of 5G land transportation, excavates important valuable information in this field, and explores and discusses the development of 5G in land transportation.

According to the proposed two models which are WVBP and TIDT to construct a text classifier in 5G land transportation based on key words as the basic unit of judgment, and both models'F_1 score reaches 0.92. Also, it is found that 'vehicle', 'dispatching', 'rail transit', and 'autonomous driving' are the core words in the field of 5G land transportation, and from the core words, "5G vehicle terminal technology", "5G urban public transport" can be obtained. "Intelligent dispatching in the field", "Intellgent, automated and standardized rail transit", "Autonomous driving 5G cloud computing" are important development trends. Therefore, we suggest that the further research would underlying these core words for the expansion of study in 5G land transportation.

Table 5 The key words of important development trends of 5G in the fields of land transportation

Algorithm	Word				
ID3	Vehicle	Scheduling	Track traffic	Automatic drive	
C5.0	Vehicle	Track traffic	Automatic drive	Scheduling	Wireless
CHAID	Vehicle	Scheduling	Track traffic	Automatic drive	Multiple access
Summary	Vehicle	Scheduling	Track traffic	Automatic drive	

Table 5 Ranking of important development trends of 5G in land transportation

Ranking	Word	Main technical fields
1	Vehicle	5G vehicle terminal technology
2	Scheduling	5G Intelligent Dispatching in the Field of Urban Public Transport
3	Track traffic	Intelligent, automated and standardized rail transit
4	Automatic drive	5G cloud computing for autonomous driving

REFERENCES

[1]. Global Electronics China(2019) The world's first 5G driverless bus line was opened. Global Electronics China, 25(05): 4-5.

[2]. Railway Quality Control(2020) Outline of the Railway First Plan for a Transportation Powerful Country in the New Era. Railway Quality Control, 48(9): 1-6+24.

[3]. Wu WX, Liu RT (2019) Overview of global 5G development.Digital Communication World, 15(5): 25-26.

[4]. Xu L (2023) Application of 5G Public-Private Network for Shenzhen Intercity Railway. Railway Transport and Economy, 45(7): 119-125.

[5]. Huawei Technologies Co., Ltd (2023) Position paper 5G Applications.[2023-09-02].https://www-file.huawei.com/-/media/corporate/pdf/mbb/5g-unlocks-a-world-of-opportunities-cn.pdf?la=zh.

[6]. Blake C (2011) Text mining. Annual Review of Information Science and Technology, 45(1): 121-155.

[7]. Sullivan D (2001) Document warehousing and text mining: Techniques for improving business operations, marketing, and sales. New York: Wiley.

[8]. Aggarwal CC (2015) Data mining: the textbook. New York: Springer.

[9]. Delen D, Crossland MD (2008) Seeding the survey and analysis of research literature with text mining. Expert Systems with Applications,34: 1707-1720.

[10]. Wang YC, Li J (1989) Automatic identification of words in Chinese literature and documents. Journal of ShangHai Jiaotong University, 23(2): 83-88.

[11]. Yin F(1998) Design and analysis of Chinese automatic segmenting system based on neural network. Journal of the China Society for Scientific and Technical Information, 17(01): 41-50.

[12]. Cui P C (2019) Research on intelligent recognition method for academic literature contents based on text mining. Beijing: Beijing Jiaotong University.

[13]. Shi FG(2020) Research on Chinese text segmentation and its visualization technology. Modern Computer, 37(12): 131-138.

[14]. Shi FG(2020) based on jieba Chinese word segmentation for implementation of Chinese text corpus preprocessing module. Computer Knowledge and Technology, 16(14): 248-251.

[15]. Wu YL, Zhao SL, Li CJ, Wei ND, Wang ZY (2017) Text classification method based on TF-IDF and cosine similarity. Journal of Chinese Information Processing, 31(5): 138-145.

[16]. He Y, Wei CQ, Lu YH (2016) Text Sentiment Analysis Based on Deep Neural Network and Topic Model—Taking Shanghai Disney Scenic Spot Tourist Satisfaction Survey as an Example. Statistical Theory and Practice, 35(12): 17-21.

[17]. Zhang J, Yan K, Ma X (2022) Analysis of complex spam filtering algorithm based on neural network. Journal of Computer Applications, 42(3): 770-777.

[18]. Wang SB (2021) Research on short text classification based on deep neural network. Dalian: Dalian University of Technology.

[19]. Yang XL (2021) Survey for performance measure index of classification learning algorithm. Computer Science, 48(8): 209-219.