Modeling for Car Quality Complaint Classification based on Machine Learning

Chen Xiao Yu[1], Hou Xia[2]
Computer School
Beijing Information Science & Technology University
Beijing, China

Zhang Xiao Min[3]
Academy of Agricultural Planning and Engineering
Ministry of Agriculture and Rural Affairs, Beijing, China

Song Ying[4]
Computer School
Beijing Information Science & Technology University
Beijing, China

***Abstract*—Cars play an important role in many aspects of people's social life, and the effective handling of car quality complaints is of great significance to the proper running of cars and the reputation maintenance of car brands; effective classification of car quality complaint texts is the basis of the efficient handling of corresponding quality complaints, while relying on manual classification has disadvantages such as heavy workload, experience dependence, and error proneness; machine learning methods have been quite widely used in the automatic classification modeling for different types of natural language texts. It is of great practical significance to construct the automatic classification model of car quality complaints based on machine learning. Based on the characteristics of car quality complaint texts, this study vectorized the texts after word segmentation, performed feature selection and dimension reduction based on correlation analysis, and combined the progressive model training method and support vector machine to construct the classification model; in model reliability analysis, it was evaluated based on the effect of data amount on the modeling and the effect of text length on the prediction probability distribution. The results show that based on the method in this study, effective automatic classification model of car quality complaint texts could be constructed.**

***Keywords*—Car; quality complaint; natural language text; classification modeling; machine learning**

I. Introduction

The studies on text classification are quite extensive, but there are few related studies on complaint text, and the applicability of classification methods is closely related to text characteristics. The composition and structure of cars are relatively complex; during the long-term use of cars, quality problems might gradually appear, reasonable handling for the quality problems has important effect on the normal operation of cars and the maintenance of user experience, which is also an important decision-making influence factor for people choosing car brand and car product.

Machine learning has been quite extensively applied to natural language text classification in recent years [1-4]. Text classification based on machine learning mainly involves two core links: text vectorization and classification modeling. The methods used in text vectorization mainly include the methods based on word frequency [5-7], the methods based on distributed static word vectors [8-11], and the methods based on distributed dynamic word vectors [12-13]. The methods used in the classification modeling mainly include classical machine learning methods [14], various neural networks [15-19], ensemble learning [20] and so on.

II. Technical Route

The technical route of this study includes seven parts, including data sorting, data characteristic analysis, word segmentation, feature extraction, classification modeling, model reliability analysis, summary and prospect.

The part of data sorting includes the acquisition of basic data, and the construction of research dataset based on the text characteristics and research purposes. The part of data characteristic analysis conducts a comprehensive overview of the dataset mainly from the aspects of data distribution, text length characteristics, the distribution of car type, the distribution of purchase time, and the distribution of car brand.

The research object of this study is Chinese text and the study involves word segmentation. The word segmentation part in the technical routes include using Jieba for word segmentation, removal of stop words, word frequency distribution analysis, classification feature word analysis, etc. The removal of stop words aims mainly at removing function words which have little significance for classification, such as the connectives in complaint texts. Word frequency distribution analysis mainly analyze the discrimination and contribution potential of high-frequency words in the classification of car quality complaints, from the perspective of the word frequency distribution of global high-frequency words in different categories. Classification feature word analysis mainly analyzes the characteristics of high-frequency words in each category after removing stop words, and conducts preliminary data status analysis.

The feature extraction part mainly involves three links: text vectorization, feature correlation analysis, and feature selection. In the text vectorization process, the text data is converted to vector form based on bag-of-word method, which doesn't include stop words. In the feature correlation analysis link, the frequency correlation of word features is analyzed through correlation matrix constructing, and the feature selection is performed by removing highly correlated word features to reduce vector dimension so as to improve the efficiency of modeling and classification.

Funding Project: Promote the Classification Development of College-Postgraduate Education Reform of the Computer School (5112211037).

In the classification modeling part, the progressive strategy is used, the proportion of the features used in modeling is gradually increased in multiple stages; the modeling effects under different proportions of features are compared to obtain the optimal modeling feature quantity. The meaning of the progressive strategy is that too few features might don't contain enough necessary information for building an effective classification model, at the same time, too many features might confuse the core information and reduce the classification ability of the model, furthermore, too many modeling features would also result in negative effects on the efficiency of modeling and classification. The classification modeling uses the method of support vector machine, and the evaluation of modeling effect is analyzed from two aspects: the classification quality on the whole dataset and the quality on different categories of complaint texts.

The model reliability analysis part includes three aspects: the effects of data amount on the overall modeling indexes, the effects of data amount on the classification effect in each category, and the effects of text length on the probability distribution of the classification prediction. The amount of training data commonly has an important effect on the reliability of the model, too little data might don't be enough to train a reliable and stable model, and the model's predict ability to new data outside the research dataset might be insufficient or unstable, generally, based on more data, more stable model could be obtained; at the same time, after the amount of data reaches a certain threshold, the continued increase of the data amount commonly no longer has significant effect on the stability of the model. For different types of texts, the amount of data required for classification model training commonly varies. This study analyzes the effect of data amount on the classification modeling of car quality complaint texts by incrementally adding of data and comparing multiple rounds of model training; the evaluation and analysis are carried out from two aspects: the effect of data amount on the overall indexes of modeling, and the effect of data amount on the classification effect in each category. In addition, the text length might have effect on the text classification prediction effect, and the discrimination of the classification prediction probability distribution could reflect to some extent the reliability of the classification prediction, therefore, in this study, the effect of text length on the probability distribution of classification prediction is regarded as another aspect of the reliability evaluation.

At the end of the study, the results are summarized to obtain effective conclusions, and the deficiencies of the study are analyzed, so as to provide reference for subsequent related research and application.

The technical route of this study is shown in Fig. 1.

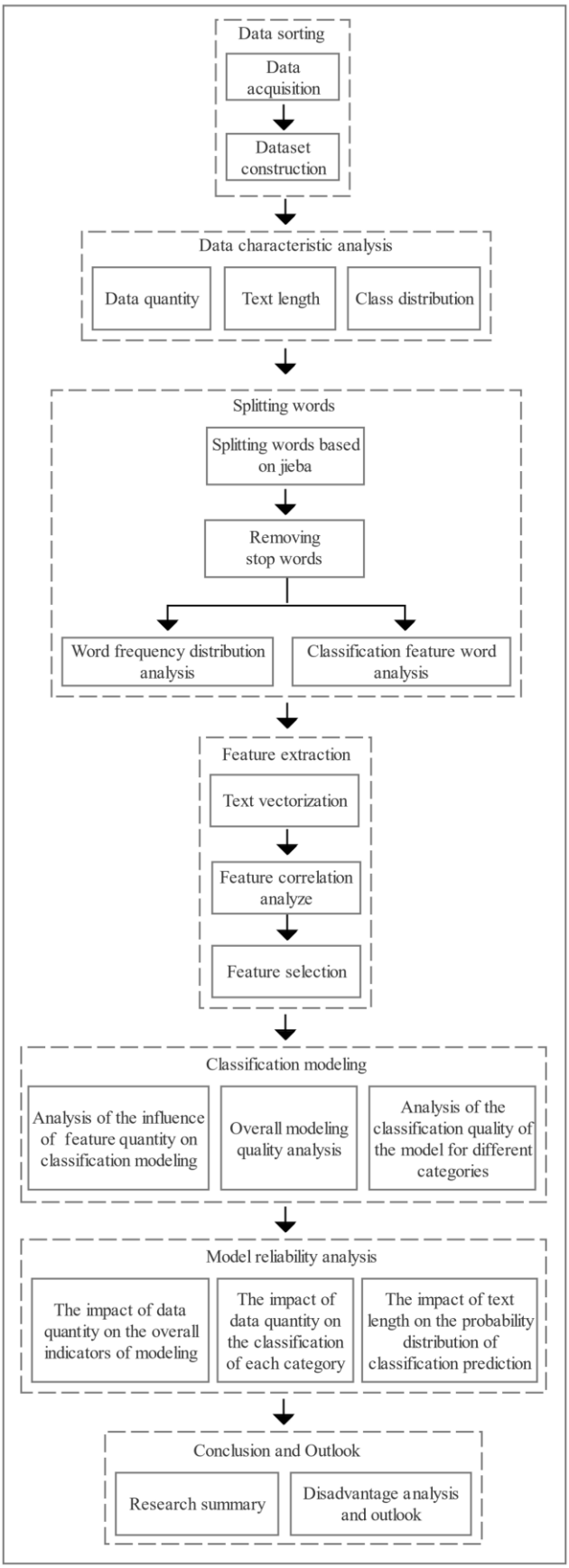

Fig. 1. Technical Route.

III. DATA

The research data of this study comes from the Beijing Car Quality Net Information Technology Limited Company. The dataset of this paper includes 8 categories of car quality complaint text data, including engine/electric motor, transmission, clutch, steering system, braking system, tires, front and rear axles and suspension system, car body accessories and electrical appliances. The data amount is 2400, and for every category, the data amount is 300. The data amount and text length characteristics are shown in Table I.

The car quality complaint texts involve attribute labels such as car type, purchase time, car brand, etc., and the attribute differences might influence the classification model training and the texts classification prediction. The data distribution of the dataset used in this paper in terms of car type, purchase time, and car brand is shown in Fig. 2.

TABLE I. DATA DESCRIPTION

No.	Category	Data amount	Average length of text	Maximum length of text	Minimum length of text	Text length standard deviation
1	Engine / electric motor	300	18.0267	27	14	1.8194
2	Transmission	300	17.8567	25	14	1.8332
3	Clutch	300	17.9000	24	15	1.7436
4	Steering system	300	17.7767	23	15	1.7214
5	Braking system	300	17.9233	23	14	1.7323
6	Tires	300	17.6500	24	15	1.7216
7	Front and rear axles and suspension system	300	17.9467	23	15	1.7571
8	Car body accessories and electrical appliances	300	18.0767	26	14	2.0845
9	All	2400	17.8946	27	14	1.8071

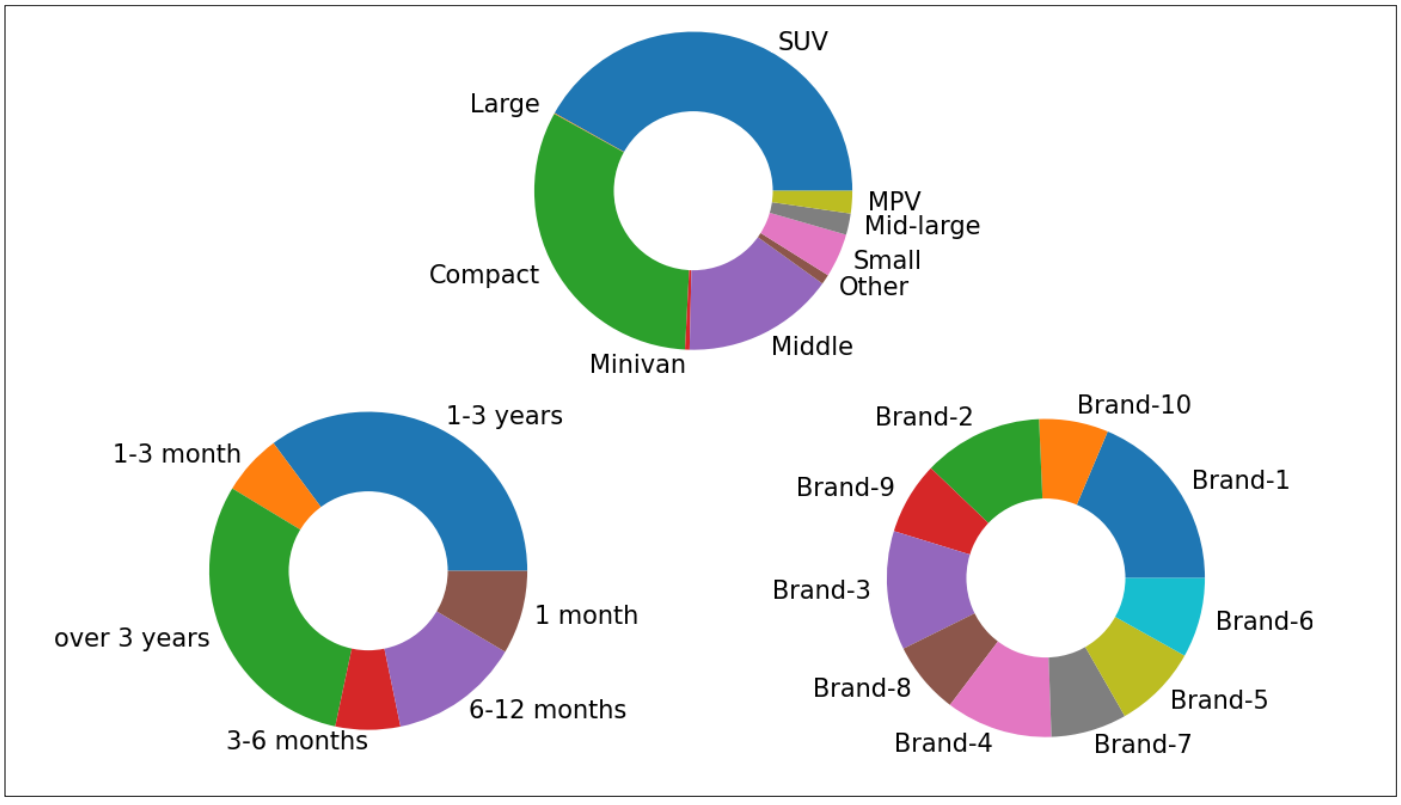

Fig. 2. Data Distribution Characteristics.

IV. Splitting Words

The research data of this study is Chinese text, and it is necessary to divide texts into words. This study uses the Jieba word segmentation tool which is widely used in the field of Chinese word segmentation to separate the words; the statistics and analysis for word segmentation results are carried out from the aspects of category, number of characters, number of separated words, number of unique words, repetition rate, etc. The word segmentation results are shown in Table II.

Fig. 3 depicts the word frequency distribution of the global high frequency words in different categories. The difference of the frequency distribution in different categories of the global high frequency words is an important reference factor for the evaluation of the potential classification discrimination contribution ability of these words. If there are widely significant differences in the word frequency distribution of global high-frequency words in different categories, the method based on word frequency might have well applicability for corresponding text classification modeling scene.

After the word segmentation and the removal of stop words, the top 20 high-frequency feature words of each category are shown in Table III.

TABLE II. Word Segmentation Results

No.	Category	Number of characters	Number of separated words	Number of unique words	Repetition rate
1	Engine / electric motor	5408	2576	584	0.7733
2	Transmission	5357	2457	499	0.7969
3	Clutch	5370	2512	413	0.8356
4	Steering system	5333	2593	479	0.8153
5	Braking system	5377	2610	495	0.8103
6	Tires	5295	2655	285	0.8927
7	Front and rear axles and suspension system	5384	2615	477	0.8176
8	Car body accessories and electrical appliances	5423	2603	766	0.7057
9	All	42947	20621	1866	0.9095

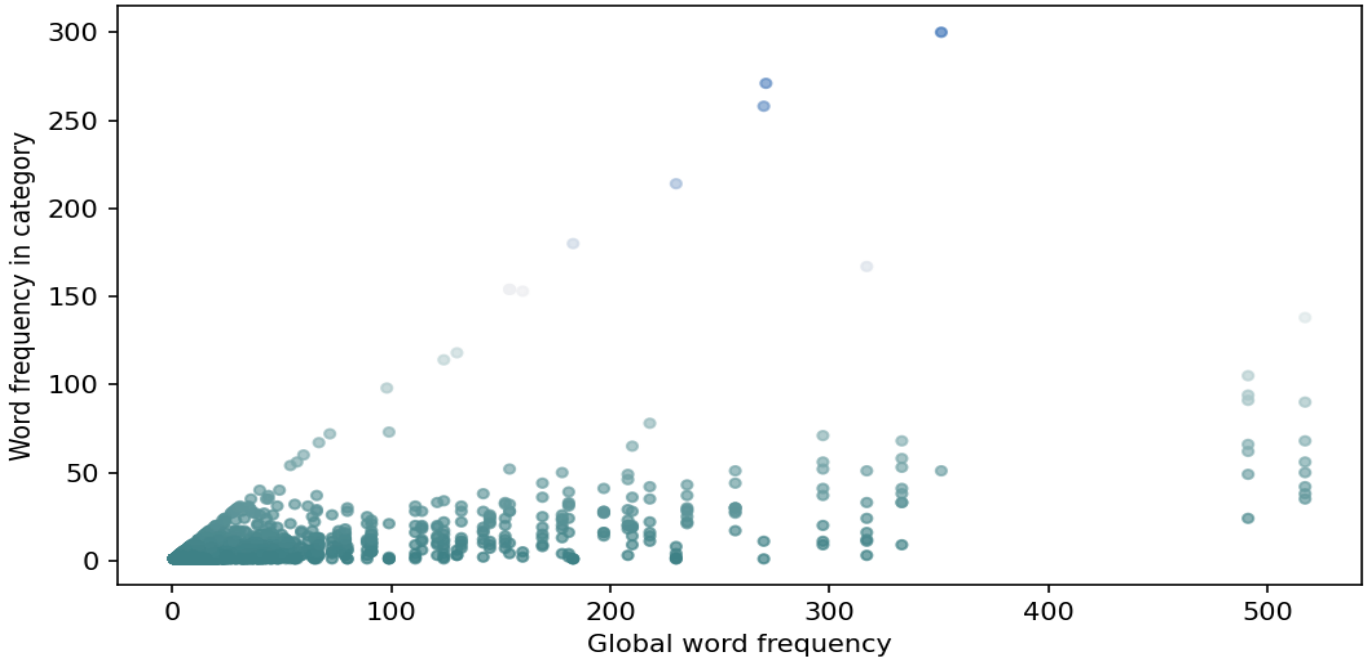

Fig. 3. Word Frequency Distribution.

TABLE III. CATEGORY HIGH FREQUENCY FEATURE WORDS

No.	Engine / electric motor	Transmission	Clutch	Steering system	Braking system	Tires	Front and rear axles and suspension system	Car body accessories and electrical appliances
1	Engine	Gearbox	Clutch	Turn	Brake	Tire	Factory	*
2	Resolve	Fault	Factory	Steering wheel	Factory	*	*	*
3	*	Factory	Hope	Factory	Resolve	*	Eat tires	Rust
4	Engine oil	*	Cause	Resolve	*	*	Driving	Resolve
5	Abnormal noise	Driving	Factory	*	Fault	*	Partial wear	Battery
6	Burn	*	When	Driving	Driving	*	Shock absorber	Power outage
7	*	Resolve	Resolve	Hope	*	Peeling	Oil spill	Hope
8	Fault	When	Driving	When	When	Cracked	*	Factory
9	*	Setback	Shift	Factory	Malfunction	Skin	Tire	*
10	Start up	Shift	*	Stuck	Handbrake	*	*	Start up
11	Hope	Oil spill	Oil spill	*	ABS	Change	Resolve	Body
12	Factory	When	Start	When	Hope	*	Rear wheel	Cause
13	*	Speed up	Jitter	Caton	Electronic	Affect	Change	Change
14	Particles	*	*	Not yet	Jitter	Factory	Chassis	*
15	Blockage	Electromechanical	*	Steering machine	*	Hope	*	Factory
16	*	Unit	*	*	Wear	*	Hope	*
17	*	*	Invalid	Direction	*	Drum kit	*	Fault
18	Oil spill	*	*	*	Factory	*	Cause	*
19	Catch	*	*	Help	*	*	Factory	Cracked
20	Device	Factory	Pedal	Affect	Cause	*	*	*

V. FEATURE EXTRACTION AND CLASSIFICATION MODELING

This study uses bag-of-word method which is based on word frequency for text vectorization; the correlation of features is analyzed based on correlation matrix; and feature selection is performed based on feature correlation to reduce the dimension of text vectors and improve the efficiency of classification model training and text classification. The correlation heatmap of the global high-frequency words after removing stop words is shown in Fig. 4. Due to space limitations, Fig. 4 only shows the relevance of the top 15 high-frequency words in the global word frequency.

This study uses a progressive feature selection strategy to incrementally set the feature usage ratio; the classification model is trained based on the SVM method, and the modeling results are evaluated and analyzed from the perspectives of overall accuracy, overall recall, overall F value, and F value of each category. The progressive feature selection strategy is beneficial to obtain a reasonable threshold of the model feature quantity, if too few features are used, the modeling effect might be adversely affected due to insufficient information, while if too many features are used, the model quality, model training efficiency, and classification prediction efficiency might be adversely affected due to the introduction of non-core information confusion. The training results of the classification model are shown in Table IV.

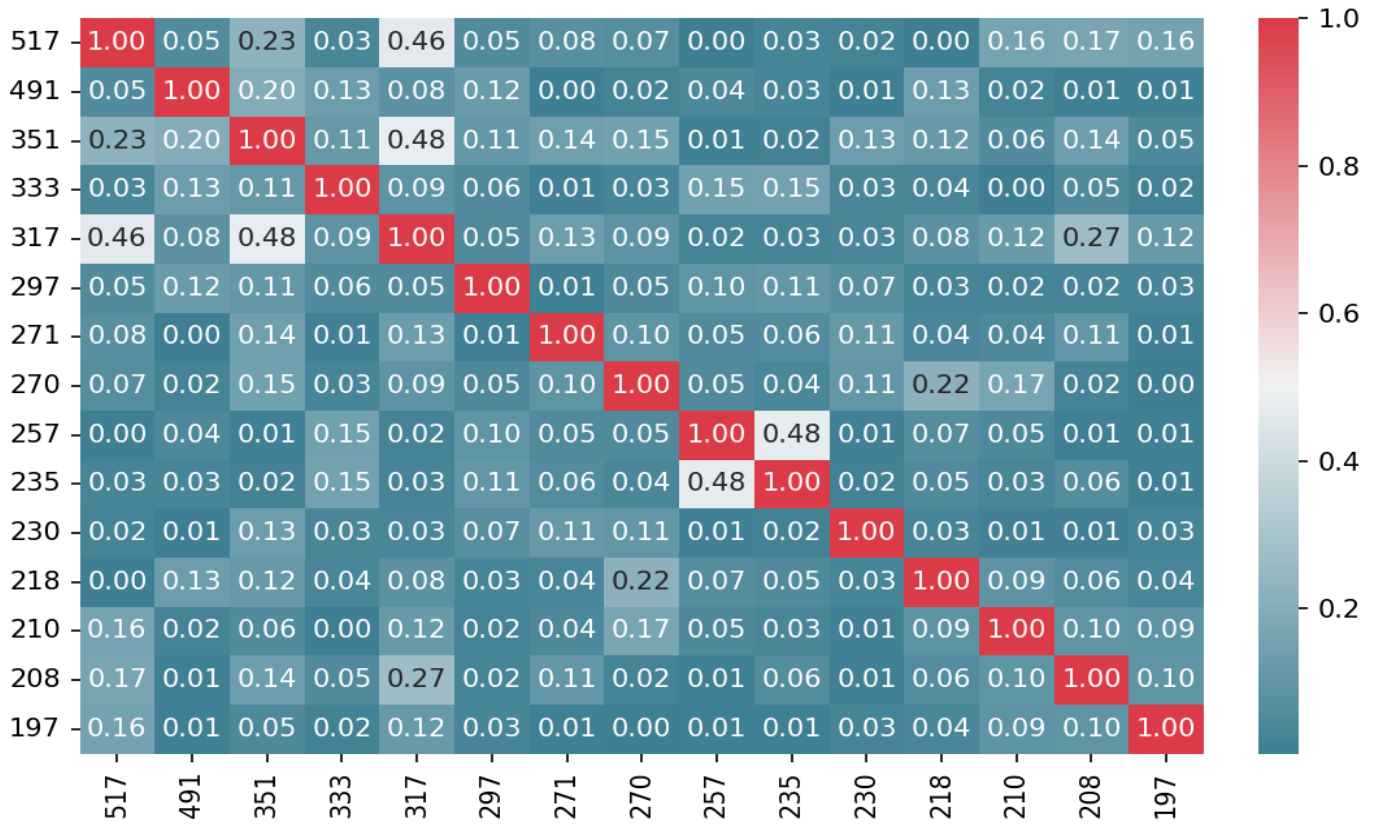

Fig. 4. Correlation Matrix Heatmap.

TABLE IV. CLASSIFICATION MODEL TRAINING

No.	Feature selection ratio	Feature amount	Accuracy	Precision	Recall	F1-score	Highestf1-scoreofeachcategory	Lowestf1-scoreofeachcategory
1	10%	156	0.8958	0.8997	0.8958	0.8968	1.0000	0.7500
2	20%	313	0.8958	0.8997	0.8958	0.8968	1.0000	0.7500
3	30%	469	0.8958	0.8997	0.8958	0.8968	1.0000	0.7500
4	40%	626	0.8958	0.8997	0.8958	0.8968	1.0000	0.7500
5	50%	782	0.8958	0.8997	0.8958	0.8968	1.0000	0.7500
6	55%	860	0.9375	0.9422	0.9375	0.9380	1.0000	0.8788
7	60%	938	0.9375	0.9443	0.9375	0.9384	1.0000	0.8657
8	65%	1017	0.9292	0.9340	0.9292	0.9300	1.0000	0.8485
9	70%	1095	0.9250	0.9296	0.9250	0.9259	1.0000	0.8485
10	75%	1173	0.9250	0.9296	0.9250	0.9259	1.0000	0.8485
11	80%	1251	0.9250	0.9296	0.9250	0.9259	1.0000	0.8485
12	85%	1329	0.9250	0.9296	0.9250	0.9259	1.0000	0.8485
13	90%	1408	0.9167	0.9220	0.9167	0.9178	1.0000	0.8358
14	95%	1486	0.9167	0.9208	0.9167	0.9176	1.0000	0.8475
15	100%	1564	0.9167	0.9238	0.9167	0.9181	1.0000	0.8235

VI. Model Reliability Analysis

Model reliability analysis is of great significance to the evaluation of model quality. This study analyzes the reliability of the model from two aspects: the effect of data amount on the classification modeling and the effect of text length on the classification prediction. The training data amount commonly has direct effect on the reliability and stability of text classification model, too little data might lead to limited applicability of the trained model and unstable prediction ability for new data, after the amount of model training data reaches a certain value, the effect of incremental data on the model training effect is commonly no longer significant.

Based on incremental data setting, this study compared multiple rounds of text classification model training, and the result parameters are shown in Table V. Text length is an important factor in text classification model training and classification prediction, the difference of the probability distribution in the classification prediction for different lengths of texts is another effective measure of the reliability of the classification model. In this study, the first 8 texts and the last 8 texts in the global ranking of text length are selected to analyze the probability distribution in the classification prediction; the results are shown in Table VI.

TABLE V. The Effect of Data Amount on Model Training

No.	Data using ratio	Data amount	Accuracy	Precision	Recall	F1-score	Highest f1-score of each category	Lowest f1-score of each category
1	10%	240	0.8333	0.8438	0.8333	0.8321	1.0000	0.3333
2	20%	480	0.8333	0.8438	0.8333	0.8321	1.0000	0.3333
3	30%	720	0.8333	0.8438	0.8333	0.8321	1.0000	0.3333
4	40%	960	0.8333	0.8438	0.8333	0.8321	1.0000	0.3333
5	50%	1200	0.8333	0.8438	0.8333	0.8321	1.0000	0.3333
6	55%	1320	0.9015	0.9080	0.9017	0.8994	1.0000	0.7333
7	60%	1440	0.9236	0.9272	0.9236	0.9230	1.0000	0.8333
8	65%	1560	0.9231	0.9304	0.9230	0.9230	1.0000	0.8095
9	70%	1680	0.9405	0.9420	0.9405	0.9399	1.0000	0.8571
10	75%	1800	0.9333	0.9338	0.9331	0.9318	1.0000	0.8372
11	80%	1920	0.9323	0.9330	0.9323	0.9322	1.0000	0.8085
12	85%	2040	0.9265	0.9272	0.9262	0.9255	1.0000	0.8302
13	90%	2160	0.9398	0.9428	0.9398	0.9401	1.0000	0.8627
14	95%	2280	0.9386	0.9445	0.9383	0.9393	0.9825	0.8923
15	100%	2400	0.9375	0.9443	0.9375	0.9384	1.0000	0.8657

TABLE VI. The Prediction Probability Distribution of the First 8 and Last 8 Texts in the Global Ranking of Text Length

No.	Engine / electric motor	Transmission	Clutch	Steering system	Braking system	Tires	Front and rear axles and suspension system	Car body accessories and electrical appliances
F-1	0.9236	0.0212	0.0017	0.0030	0.0034	0.0027	0.0032	0.0411
F-2	0.9934	0.0018	0.0002	0.0002	0.0023	0.0004	0.0004	0.0012
F-3	0.0029	0.0051	0.0015	0.0013	0.0027	0.0019	0.0032	0.9813
F-4	0.0102	0.9530	0.0227	0.0019	0.0007	0.0005	0.0024	0.0086
F-5	0.0433	0.0225	0.0040	0.0063	0.0133	0.0037	0.0259	0.8810
F-6	0.9988	0.0002	0.0003	0.0001	0.0001	0.0003	0.0000	0.0001
F-7	0.0247	0.9491	0.0161	0.0004	0.0005	0.0012	0.0027	0.0052
F-8	0.0066	0.0006	0.9775	0.0017	0.0037	0.0023	0.0027	0.0050
L-1	0.9789	0.0009	0.0049	0.0008	0.0042	0.0007	0.0013	0.0083
L-2	0.9420	0.0015	0.0008	0.0022	0.0070	0.0022	0.0044	0.0400
L-3	0.9352	0.0102	0.0060	0.0015	0.0030	0.0023	0.0092	0.0325
L-4	0.0247	0.9403	0.0108	0.0026	0.0035	0.0024	0.0044	0.0113
L-5	0.0047	0.9879	0.0034	0.0003	0.0005	0.0010	0.0007	0.0015
L-6	0.0015	0.9848	0.0101	0.0003	0.0005	0.0003	0.0010	0.0015
L-7	0.0005	0.0001	0.0003	0.0005	0.9946	0.0008	0.0008	0.0023
L-8	0.0231	0.0044	0.0018	0.0023	0.0058	0.0024	0.0029	0.9575

VII. Conclusion and Outlook

This study focuses on the automatic classification of car quality complaint, the research content mainly includes data characteristic analysis, word segmentation, feature extraction, classification modeling, and model reliability analysis. The research results show that based on the method combining the text vectorization based on word frequency, the feature selection and dimensionality reduction based on correlation analysis, and the feature increase SVM model training, the effective classification model for car quality complaints texts could be obtained; in this study, the best modeling effect is obtained when using 938 features for model training, among the global indexes, the accuracy, recall, and f1-score reach 0.9375, 0.9375, and 0.9384 respectively, the highest f1-score of each category is 1.0000, and the lowest is 0.8657; in the model reliability evaluating based on incremental data amount, after the data using ratio reaches 75%, the training effect is almost stable, the classification prediction probability distribution analysis based on the global long texts and global short texts shows that the classification probability values obtained from the model shows a high degree of discrimination overall.

In general, based on the method of this study, effective modeling for the automatic classification of car quality complaint texts could be realized; at the same time, the research content of this study belongs to theoretical research which has not been applied to practice, it is expected that this study could provide effective reference for subsequent research and practical application.

References

[1] Zhu Fang Peng, Wang Xiao Feng, Text classification for ship industry news [J], Journal of Electronic Measurement and Instrumentation, 2020, 34 (01): 149-155.

[2] Zhao Ming, Du Hui Fang, Dong Cui Cui, Chen Chang Song, Diet health text classification based on word2vec and LSTM [J], Transactions of the Chinese Society for Agricultural Machinery, 2017, 48 (10): 202-208.

[3] Bao Xiang, Liu Gui Feng, Yang Guo Li, Patent text classification method based on multi-instance Learning [J], Information Studies: Theory & Application, 2018, 41 (11): 144-148.

[4] Wen Chao Dong, Zeng Cheng, Ren Jun Wei, Zhang Yan, Patent text classification based on ALBERT and bidirectional gated recurrent unit [J], Journal of Computer Applications, 2021, 41 (02): 407-412.

[5] Hu Jing, Liu Wei, Ma Kai, Text categorization of hypertension medical records based on machine learning [J]. Science Technology and Engineering, 2019, 19 (33): 296-301.

[6] Yu Hang, Li Hong Lian, Lü Xue Qiang, Text classification of NPC report contents [J], Computer Engineering and Design, 2021, 42 (06): 1772-1778.

[7] Wang Xiang Xiang, Fang Hui, Chen Chong Cheng, Classification technique of cultural tourism text based on naive Bayes [J]. Journal of Fuzhou University (Natural Science Edition), 2018, 46 (05): 644-649.

[8] Zhou Qing Hua, Li Xiao Li, Research on short text classification method of railway signal equipment fault based on MCNN [J]. Journal of Railway Science and Engineering, 2019, 16 (11): 2859-2865.

[9] Feng Shuai, Xu Tong Yu, Zhou Yun Cheng, Zhao Dong Xue, Jin Ning, et al. Rice knowledge text classification based on deep convolution neural network [J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52 (03): 257-264.

[10] Niu Zhen Dong, Shi Peng Fei, Zhu Yi Fan, Zhang Si Fan, Research on classification of commodity ultra-short text based on deep random forest [J]. Transactions of Beijing Institute of Technology, 2021, 41 (12): 1277-1285.

[11] Zhang Yu, Liu Kai Feng, Zhang Quan Xin, Wang Yan Ge, Gao Kai Long, A combined-convolutional neural network for Chinese news text classification [J]. Acta Electronica Sinica, 2021, 49 (06): 1059-1067.

[12] Li Ke Yue, Chen Yi, Niu Shao Zhang, Social E-commerce text classification algorithm based on BERT [J], Computer Science, 2021, 48 (02): 87-92.

[13] Tian Yuan, Yuan Ye, Liu Hai Bin, Man Zhi Bo, Mao Cun Li, BERT pre-trained language model for defective text classification of power grid equipment [J]. Journal of Nanjing University of Science and Technology, 2020, 44 (04): 446-453.

[14] Zhao Yan, Li Xiao Hui, Zhou Yun Cheng, Zhang Yue. A study on agricultural text classification method based on naive bayesian [J]. Water Saving Irrigation,2018(02):98-102.

[15] Chen Ping, Kuang Yao, Hu Jing Yi, Wang Xiang yang, Cai Jing. Text categorization method with enhanced domain features in power audit field [J]. Journal of Computer Applications, 2020, 40 (S1): 109-112.

[16] Liu Zi Quan, Wang Hui Fang, Cao Jing, Qiu Jian, A classification model of power equipment defect texts based on convolutional neural network [J]. Power System Technology, 2018, 42 (02): 644-651.

[17] Ge Xiao Wei, Li Kai Xia, Chen Ming, Text classification of nursing adverse events based on CNN-SVM [J]. Computer Engineering & Science, 2020, 42 (01): 161-166.

[18] Wang Meng Xuan, Zhang Sheng, Wang Yue, Lei Ting, Du Wen, Research and application of improved CRNN model in classification of alarm texts [J]. Journal of Applied Sciences, 2020, 38 (03): 388-400.

[19] Wang Si Di, Hu Guang Wei, Yang Si Yu, Shi Yun, Automatic transferring government website e-mails based on text classification [J]. Data Analysis and Knowledge Discovery, 2020, 4 (06): 51-59.

[20] Zhang Bo, Sun Yi, Li Meng Ying, Zheng Fu Qi, Zhang Yi Jia, et al. Medical text classification based on transfer learning and deep learning [J]. Journal of Shanxi University (Natural Science Edition), 2020, 43 (04): 947-954.