Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 8, No. 3, 2026

Patent Value Score Prediction Based on BERT-XGBoost-Stacking with Late Fusion

Wenjuan Li*

School of Mathematics and Statistics, Qinghai Minzu University, Xining 810000, China
* **Corresponding author**: Wenjuan Li (Email: Alin19129@163.com)

Abstract: Predicting the value scores of high-value patents is essential for evaluating technological innovation, managing intellectual property, and promoting industrial development. However, existing methods still face challenges in effectively fusing multimodal data and achieving high prediction accuracy. To address this issue, we propose a late fusion model based on BERT-XGBoost-Stacking to improve the accuracy and robustness of patent value assessment. This approach leverages BERT to extract deep semantic features from patent texts, employs XGBoost to model structured numerical features, and optimizes the fusion strategy through the Stacking framework. Experimental results on patents in China's new energy vehicle sector demonstrate that the proposed method outperforms single-modal models in terms of Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), indicating higher prediction accuracy and stability. This study not only enriches quantitative methods for patent valuation but also provides new insights for the application of artificial intelligence in patent analysis, thereby supporting technological innovation and high-quality industrial development.

Keywords: Patent Value Prediction; BERT; Xgboost; Stacking Ensemble; Late Fusion.

1. Introduction

With the high-quality development of China's economy, innovation has become the core driving force. As an important indicator to measure the level of technological innovation, high-value patents are of great significance in promoting economic growth and industrial upgrading. Accurately identifying high-value patents is not only the basis for the government to formulate patent cultivation strategies, but also an important research topic of interest to academia. This paper aims to design a simple and efficient method for predicting patent value scores, realize multi-dimensional analysis of patent value by fusing text and numerical features, so as to improve the recognition accuracy of high-value patents.

Early research on patent value evaluation relied on indicators to proxy the value level, such as the number of patent citations and patent life. Later, the evaluation system was gradually expanded to cover various factors such as technology, law and market, and methods such as fuzzy comprehensive evaluation and principal component analysis were introduced to determine weights or reduce dimensionality to improve the systematicity and objectivity of evaluation [1-3]. In recent years, with the progress of computer technology, machine learning methods have been introduced into the field of patent value evaluation. For example, Hu Zewen et al. [4] optimized the patent indicator system by using BP neural network and MIV algorithm to improve the ability of identifying high-value patents. Sun Xiaoming et al. [5] used machine learning algorithms to predict the "unused patents" of scientific research institutions and improve the utilization rate of patent resources. However, these methods mainly focus on the external attributes of patents and fail to fully explore the deep semantic information contained in patent texts. With the rapid development of Natural Language Processing (NLP) technology, researchers have begun to use text information to evaluate patents to improve prediction accuracy. For example, Sun Ran et al. [6] extracted semantic features of patent titles based on BERT and predicted patent value by combining machine learning and deep learning models; Zhang Biao et al. [7] used Word2Vec to extract semantic features of patent specifications and constructed different machine learning models combined with principal component analysis to identify transferable patents; Fu Jiao et al. [8] combined traditional bibliographic items with semantic features of patent claims and used the CatBoost model to predict early patent value.

Deep learning has made remarkable progress in patent text feature extraction, but model fusion still faces many challenges. On the one hand, there are differences in feature space and learning mechanisms between numerical features extracted by deep learning models and those of traditional machine learning models, which may lead to incompatibility problems during fusion. In recent years, this problem has become a research hotspot. The Stacking framework proposed by Breiman et al. provides a general framework for fusing heterogeneous models, and realizes the complementary advantages of models and improves the generalization ability by introducing a meta-learner [9]. In the patent value evaluation task, existing studies have verified the positive effect of the fusion framework on prediction results by fusing multi-source structural information, providing evidence for the research paradigm of "multi-source information fusion - improving prediction performance" [10]. On the other hand, patent text data usually have high dimensions, and direct fusion of numerical features often requires a more complex model structure, thus bringing higher training and computing costs. Taking abstract semantic representation as an example, in previous studies, abstracts are often encoded into high-dimensional vectors (such as 512-dimensional semantic vectors) and then jointly modeled with other features, which increases feature redundancy and modeling complexity [11]. To address the challenges of high-dimensional and long-sequence modeling, the Transformer model proposed by Vaswani et al. uses the attention mechanism to capture global dependencies, providing a new

idea for processing large-scale text data [12]; in the research of patent value evaluation, studies have also modeled numerical data based on the Transformer structure and verify its effectiveness with a multi-indicator system, providing a reference for structural optimization and complexity control [13].

In addition, most current research on high-value patents adopts binary classification methods to divide patents into high-value and low-value categories. However, compared with simple category judgment, continuous score prediction can provide a more fine-grained characterization of value. In machine learning regression modeling, the gradient Boosting framework proposed by Friedman models continuous targets in the form of function approximation, providing classic methodological support for score prediction tasks [14].

This paper proposes a BERT-XGBoost fusion model based on the Stacking framework for patent value score prediction. This method uses BERT to extract deep semantic features of patent texts and XGBoost to model structured numerical features respectively, and fuses the prediction results of the two base models through a meta-learner to realize the prediction-level fusion (late fusion) of multimodal information. Compared with single models and feature splicing methods, this model alleviates the problem of high text dimensionality while improving the generalization ability and prediction stability of the model. Experimental results show that the fusion model achieves better performance in MSE, MAE and MAPE indicators in the new energy vehicle patent value prediction task. The research results provide an effective technical path for high-value patent evaluation and expand the application potential of artificial intelligence in the field of patent analysis.

2. Theory and Methods

2.1. Principle of the BERT Model

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language representation model proposed by Google in 2018. Based on the Transformer encoder structure, the model learns the semantic representation of text through deep bidirectional context modeling, thereby improving the model's ability to understand linguistic relationships.

BERT is pre-trained on a large-scale corpus, mainly including two tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP).

Let the input sequence be:

$$X = [x_1, x_2, \dots, x_n] \tag{1}$$

2.1.1. Masked Language Model (MLM)

In the MLM task, we randomly select some words to be replaced with a [MASK] token, and then let the model predict these words through the context:

$$P(x_i|X_{-i}) = \text{softmax}(W \cdot h_i) \tag{2}$$

where X_{-i} represents the context sequence after removing x_i, h_i is the hidden state vector calculated by the Transformer, and W is the word embedding matrix. This task enables the model to learn deep semantic relationships by using contextual information from both left and right sides.

2.1.2. Next Sentence Prediction (NSP)

In the NSP task, the model receives a sentence pair (A, B) and judges whether B is the real subsequent sentence of A. The model performs classification based on the sequence-level representation corresponding to the token:

$$P(\text{IsNext}(A, B)) = \text{softmax}(W_{cls} \cdot h_{cls}) \tag{3}$$

where h_{cls} is the output of BERT at the special token, which serves as the representation of the entire input sequence. This task helps the model learn inter-sentence relationships and the semantic structure of long texts.

Through the above pre-training tasks, BERT can obtain contextual representations with strong expressive ability, and can be fine-tuned for various downstream tasks such as text classification, similarity calculation and regression prediction.

2.2. Principle of the XGBoost Model

XGBoost (Extreme Gradient Boosting) is an efficient implementation based on Gradient Boosting Decision Tree (GBDT). Its core idea is to iteratively construct multiple decision trees through an additive model, and learn the negative gradient direction of the current model on the objective function in each iteration, so as to continuously optimize the overall performance.

Given the training samples $\{(x_i, y_i)\}_{i=1}^{n}$, the prediction form of XGBoost is:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{4}$$

where $f_k(x)$ represents the k-th regression tree.

XGBoost learns the model by minimizing the following regularized objective function:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{5}$$

where $l(y_i, \hat{y}_i)$ is the loss function (Mean Squared Error is commonly used for regression tasks), and $\Omega(f)$ is the structural regularization term of the tree.

The regularization term is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{6}$$

where T is the number of leaf nodes, w_j is the weight of the j-th leaf node, γ controls the complexity of the tree structure, and λ controls the size of leaf weights.

In each iteration, XGBoost performs a second-order Taylor expansion on the objective function, and uses the first-order and second-order gradient information to calculate the optimal split point, thereby improving training efficiency and model accuracy.

2.3. BERT and XGBoost Fusion Model Based on the Stacking Framework

Stacking (Stacked Generalization) is an ensemble learning method that improves the generalization ability of the model by combining the prediction results of multiple base learners and training a meta-learner to generate the final prediction.

In the fusion model proposed in this paper, BERT and XGBoost serve as base learners to process different types of data respectively.

2.3.1. Text Feature Modeling

For the input text x_i, the semantic representation is obtained through BERT encoding:

$$h_i = \text{BERT}(x_i) \tag{7}$$

This vector can be used for downstream task prediction, and the output of the text model is obtained:

$$\hat{y}_{text,i} = f_{text}(h_i) \tag{8}$$

2.3.2. Numerical Feature Modeling

For the numerical feature vector v_i, XGBoost is used for

modeling to obtain the prediction result:

$$\hat{y}_{vec,i} = f_{vec}(v_i) \quad (9)$$

2.3.3. Stacking Fusion

After the training of the base learners is completed, their prediction results are used as the input features of the meta-learner:

$$z_i = [\hat{y}_{text,i}, \hat{y}_{vec,i}] \quad (10)$$

The meta-learner learns the fusion weights by minimizing the loss function:

$$L_{meta} = \sum_{i=1}^{n} l(y_i, f_{meta}(z_i; W)) \quad (11)$$

where f_{meta} is the meta-model (such as linear regression or support vector machine), and W is the parameter to be learned.

To avoid information leakage, the predictions of the base learners are generated through cross-validation. This method realizes the prediction-level fusion (late fusion) of multimodal information, and has stronger generalization ability than feature splicing.

3. Empirical Analysis

3.1. Indicator Selection and Data Source

Patent value has the characteristics of uncertainty, timeliness and fuzziness, and there are many factors affecting patent value, so it is necessary to scientifically select high-value patent evaluation indicators. The recommended national standard Patent Evaluation Guidelines (National Standard No. GB/T42748—2023, hereinafter referred to as the Guidelines), compiled by the National Intellectual Property Administration in conjunction with the People's Bank of China and the State Administration of Financial Regulation, has been implemented since September 1, 2023. These Guidelines provide an important reference for indicator construction in the patent field.

Combining previous research and the latest patent indicator policies, in accordance with the characteristics of data indicators in the current patent database and following the principles of ease of acquisition and comprehensiveness, this paper constructs a patent value evaluation indicator system including 11 indicators in three dimensions: Legal stability, Technical advancement, Protection scope. The specific explanation of the indicators is shown in Table 1.

Table 1 Patent Value Indicator System

First-level Indicator	Second-level Indicator	Indicator Definition
Legal	Number of claims	The total number of claims in the patent application document, i.e., the number of specific clauses of the technical solution that the applicant wishes to be protected by the patent.
	Number of document pages	The total number of pages of the patent, i.e., the number of pages occupied by the document from the first page to the last page.
Technical	Number of IPC classifications	The number of IPC classifications assigned to the patent.
	Number of inventors	The number of people participating in the patent invention.
	Number of citations	The number of times the patent is cited by other subsequent patents.
	Number of cited times	The number of times a patent is cited by other patents.
	Number of simple patent families	A set of patents with the same priority or the same earliest application.
Economic	Number of transfers	The number of times the patent right is transferred from one person or institution to another.
	Number of licenses	The number of times the patent right holder authorizes other parties to use its patent.
	Number of pledges	The number of times the patent right is used as a guarantee for debts.
	Duration of existence	The time span from the authorization date of the patent to the expiration date (search date).

All data selected in this paper are from the Incopat patent database system, which is divided into three dimensions: technical stability, technical advancement and protection scope, and selects more than 20 indicators including patent type, number of cited times and number of patent families, being a major domestic patent database. This paper selects the invention-authorized patents in China's new energy vehicle field as the research object, with the authorization date up to the end of 2024 and the search date on February 15, 2025. Finally, 14,528 pieces of data are screened out, and their vector data and text data are obtained. Abstracts and patent claims are important data reflecting patent information, and this paper chooses to combine them as text indicators to extract more patent information. Preliminary cleaning is carried out on the data, duplicate data are deleted, and blank data are filled with 0. The patent text part is cleaned with regular expressions to remove meaningless symbols and numbers; then the jieba library is used to build a custom dictionary, stop- word list and synonym list suitable for this field to construct a patent text corpus.

Compared with simple binary classification, patent value score prediction is more in line with practical research. The indicator weights calculated by the objective weighting method will vary in different fields, and the scores calculated at this time are more in line with the actual situation. Based on this, this paper makes full use of the database indicator information, combines the scores of patents in three dimensions of technical stability, technical advancement and protection scope, and uses the CRITIC weighting method to calculate the corresponding weights, as shown in Table 2.

Table 2 CRITIC Weight Results

Indicator	Legal stability	Technical advancement	Protection scope
Weight	0.2405	0.5790	0.1805

Then the value score of each patent is calculated according to the above weights, which is used as the response variable. The final score obtained by CRITIC is in the range of 0-1, and the closer it is to 1, the higher the patent value. Finally, a data

set with 14,528 pieces of vector data and text data, with the patent score of each piece of data as the response variable, is obtained for subsequent research.

3.2. Experimental Results

Ablation experiments are conducted on the BERT-XGBoost-Stacking model constructed in this paper to verify its effect. First, the data are read in for Monte Carlo experiments, the number of iterations is set to 10, and the training set, validation set and test set are divided at a ratio of 3:1:1 each time. The specific operation steps are shown in Figure 1.

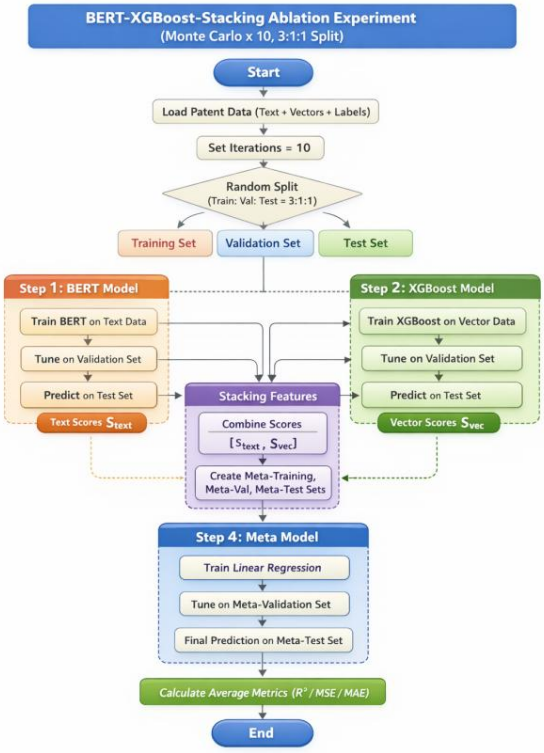

Figure 1 Steps of the ablation experiment

Two base learners, BERT and XGBoost, are constructed to model patent text and structured numerical features respectively, and parameter optimization is completed through the validation set; then the tuned models are used to predict each data set to obtain text scores and numerical scores, which are spliced into two-dimensional features as the input of the meta-learner. Finally, a linear regression meta-model is trained and optimized with the spliced features to perform fusion prediction on the test set and obtain the value score of each patent. The final prediction is made on the test set to obtain the value score of each patent under the fusion model. Error Results of the Monte Carlo Experiment are shown in Table 3.

Table 3 Error Results of the Monte Carlo Experiment

Model	MSE - Mean	MSE - Std	RMSE - Mean	RMSE - Std	MAE - Mean	MAE - Std	MAPE - Mean	MAPE - Std
BERT	0.041397	0.00569	0.203056	0.013554	0.163405	0.00961	34.738075	2.034679
XGBoost	0.026408	0.000473	0.1625	0.001455	0.143642	0.001469	28.823635	1.026704
Proposed Model	0.024260	0.000483	0.155748	0.001545	0.130988	0.001643	25.761225	0.915811

Typically, Mean Squared Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are important indicators to measure the performance of regression models. The smaller the error value, the better the prediction effect of the model. From the results in the above table, after 10 Monte Carlo experiments, the mean values of the BERT-XGBoost-Stacking model in all error indicators are significantly lower than those of using BERT or XGBoost models alone. This result fully verifies the effectiveness of the BERT-XGBoost-Stacking model, indicating that the model has a significant advantage in improving prediction accuracy.

4. Conclusions and Prospects

In the current field of data analysis and machine learning, multimodal data fusion and complex feature interaction modeling have become key technologies to improve prediction accuracy. The method proposed in this paper combines the advantages of BERT and XGBoost, and realizes the final prediction through the Stacking meta-model, constructing an efficient prediction framework. Among them, BERT is responsible for extracting deep semantic features of text data, and XGBoost focuses on processing numerical data. The synergy between the two significantly improves the

model performance. Furthermore, this paper uses linear regression as the meta-model to fuse the prediction results, ensuring the stability and accuracy of the model.

Future research can further expand the selection of meta-models, such as adopting the Generalized Additive Model (GAM) to better characterize the nonlinear interaction relationships between different features and enhance the fusion ability of text and numerical data. In addition, the model selection can also be extended to more advanced text processing models (such as T5, RoBERTa) and numerical modeling methods (such as LightGBM, CatBoost) to improve adaptability and generalization ability. The meta-model can also consider Support Vector Machine (SVM) or neural networks to further enhance the predictive performance. Overall, the research direction combining additive models and diversified model selection will have broad prospects in data fusion, feature interaction modeling and model optimization. In the future, more flexible deep learning methods such as multi-task learning can be explored to further improve the generalization ability of the model and promote its application value in fields such as financial prediction, medical diagnosis, and social media analysis.

Acknowledgments

This work was financially supported by the fund of the 2025 Postgraduate Innovation Project of Qinghai Minzu University, titled "Patent Value Score Prediction Based on BERT-XGBoost-Stacking with Late Fusion" (Project Number:07M2025001).

References

[1] Liu Yan. Review and Future Outlook of Patent Value Assessment Research [J]. Library and Information Service, 2022, 66(15): 127-139.

[2] Yu Xinmiao, Li Wenhong. Review and Outlook of Patent Quality Perception and Evaluation Systems [J]. Studies in Science of Science, 2024, 42(10): 2100-2109.

[3] Li Liming. Literature Review and Future Outlook on Patent Value Research [J]. Journal of Information Science, 2023, 42(02): 166-174.

[4] Hu Zewen, Zhou Xiji. Prediction of High-Value Patents and Analysis of Influencing Factors Based on BP Neural Network and MIV Algorithm [J]. Journal of Information Resource Management, 2023,13(6):144-155.

[5] Sun Xiaoming, Xiong Wang, Yuwen Lewei, et al. Value Assessment of Dormant Patents in Universities and Research Institutions [J/OL]. Science and Technology Progress and Countermeasures, 1-10.

[6] Sun Yaoran, An Yaolu, Li Yaogang. Patent Value Prediction Using Multi-Feature Fusion: A Case Study of 5G Technology [J]. Modern Intelligence, 2022, 42(11): 87-96.

[7] Zhang Yaobiao, Wu Yaohong, Gao Daobin, et al. Research on Identifying Transferable Patents in Universities Based on Feature Fusion [J]. Journal of Intelligence, 2022, 41(9): 159-165.

[8] Fu Yao-jiao, Liang Li-zhi. Early Prediction of Patent Value Integrating Claim Semantic Features [J]. Journal of Information Science, 2024, 43(2): 183-191.

[9] BREIMAN L. Stacked regressions [J]. Machine Learning, 1996, 24(1): 49-64.

[10] Chen Xi, Cheng Ge, Yin Zhibin. A Patent Value Evaluation Method Integrating Heterogeneous Graph Global Structure Information and Time Series [J/OL]. Journal of Intelligence, 2025-03-07.

[11] Feng Guohua, Li Lin, Liu Renhua, Deng Weiwei. Identifying High-Value Patents in Nanomedicine: Integration of Cross-Perspective and Multi-Method Approaches [J/OL]. Journal of Intelligence, 2025-05-15.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[13] Li Nian. Research on Transformer Model Improvements for Patent Value Assessment [D]. Nanjing: Nanjing Tech University, 2024.

[14] Friedman J.H. Greedy Function Approximation: A Gradient Boosting Machine [J]. Annals of Statistics, 2001, 29(5): 1189-1232.