Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4, 1145-1159
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i4.6188

Analysis of employment trends and prediction model for Chinese college graduates based on natural language processing

Kong Jian[1*], Ma. Agatha Anne D. Dizon[2]
[1,2]Angeles University Foundation, Angeles City, Philippines; 13956022293@163.com (K.J.) dizon.ma.agathaanne@auf.edu.ph (M.A.A.D.D.)

Abstract: This study aims to address the employment challenges faced by Chinese college graduates by developing a predictive framework that integrates Natural Language Processing (NLP) with ensemble machine learning techniques. The methodology involves extracting features from both structured data (e.g., demographics, academic records) and unstructured text (e.g., job descriptions), followed by classification using Support Vector Machine (SVM), Bagging, and XGBoost algorithms. A weighted voting strategy is applied to fuse the model outputs. Experimental results using real-world data from 2010 to 2019 show that the ensemble model achieves 81.48% accuracy in employment type prediction and 81.53% in salary level prediction, outperforming individual models. These findings demonstrate that combining NLP-driven feature extraction with hybrid learning architectures can significantly enhance predictive performance. The study concludes that this integrated approach offers a powerful tool for analyzing employment trends and guiding graduate career planning. Practical implications suggest that universities and policymakers can adopt such models to improve employment guidance, tailor education to market demands, and enhance talent alignment strategies.

Keywords: *Educational Data Mining, Employment Forecasting, Machine Learning in Education.*

1. Introduction

China has undergone a massive expansion of higher education over the past two decades, dramatically increasing the number of college graduates entering the labor market [1]. This massification has intensified job competition and led to new challenges in the graduate labor market. Many graduates now face an "employment dilemma," struggling to secure suitable positions commensurate with their education, and there is a rising issue of "overeducation," where degree holders work in jobs that do not require a college degree [2]. In recent years, annual graduate cohorts have reached record highs (over 10 million in 2022), creating unprecedented employment pressure in the country [3]. The situation has been further compounded by external shocks like the COVID-19 pandemic, contributing to delayed workforce entry or "slow employment" among young graduates [4]. Ensuring stable and high-quality employment for college graduates has thus become a top priority for policymakers and educators in China's contemporary higher education landscape.

The rise of big data analytics and artificial intelligence has introduced new tools to address these employment challenges. Natural language processing (NLP) techniques have been applied to extract insights from large text corpora such as online job listings and social media, providing a richer understanding of labor market demand [5, 6]. For example, Johnson and Zhang [5] mined online job portal data using NLP to uncover employment trend dynamics, while Xu and Chen [6] analyzed thousands of job descriptions to reveal evolving skill requirements for new graduates. Machine learning (ML) methods have likewise been used to build predictive models of graduate employment outcomes. Zhang, et al. [7] developed a neural network model to forecast college graduate employment rates in

History: Received: 7 February 2025; Revised: 28 March 2025; Accepted: 31 March 2025; Published: 12 April 2025
* Correspondence: 13956022293@163.com

China, demonstrating the feasibility of data-driven trend prediction. Researchers have also experimented with more advanced AI models: Bhagavan, et al. [8] proposed a hybrid learning algorithm to predict graduates' job attainment, and Bai and Hira [9] designed a deep belief network for employability classification, both showing the promise of state-of-the-art ML techniques in this domain. These efforts reflect a growing convergence of educational data mining and labor market analytics to better understand and improve graduate employability.

However, there remains a noticeable gap in the literature: few studies have fully integrated NLP-driven text analysis with predictive modeling in a unified framework for graduate employment analytics. Most prior works treat unstructured text analysis and quantitative prediction separately. A pioneering study by Liu, et al. [10] combined deep learning with NLP to forecast employment trends in the tech sector, illustrating the potential of such an integrated approach. Yet to our knowledge, no prior work has applied a unified NLP + ensemble learning methodology to the broader context of Chinese college graduate employment. To fill this gap, the present paper proposes a novel framework that leverages NLP to extract features from unstructured data (e.g. job postings, policy documents, social media) and feeds them into an ensemble machine learning model to predict employment trends and outcomes for Chinese graduates. By fusing insights from textual data with traditional quantitative indicators, our approach aims to improve prediction accuracy and offer a more comprehensive understanding of the factors influencing graduate employment. The contributions of this study are summarized as follows:

1. Multi-source Employment Data Analysis: We construct a comprehensive dataset of Chinese college graduate employment information, combining official statistics with large-scale textual data from online job postings and related sources. This dataset enables a longitudinal analysis of employment trends and skill demand patterns in China's graduate labor market.
2. Integration of NLP and Ensemble Learning: We develop a novel methodology that integrates NLP-based text analytics with ensemble machine learning techniques for employment outcome prediction. In particular, our framework uses NLP to derive features (such as prevalent job requirements and sentiment indicators) from unstructured text, and then incorporates these features into an ensemble predictive model. This is a first-of-its-kind integration applied to Chinese graduate employment, bridging the gap between qualitative text analysis and quantitative forecasting.
3. Improved Predictive Performance: Using the proposed approach, we build an ensemble prediction model (combining multiple classifiers) that achieves high accuracy in forecasting graduate employment status and trends. Experimental results show that our model outperforms several baseline methods, validating the effectiveness of integrating NLP-derived features. The ensemble strategy also enhances robustness and generalization of the predictions across different cohorts and regions.
4. Insightful Trend and Policy Analysis: We conduct an in-depth analysis of the model outputs and extracted features to uncover key factors influencing graduate employment. The study reveals insights into industry-wise demand shifts, regional employment disparities, and the impact of graduates' skill profiles on employability. These findings provide valuable evidence to inform policymakers and educational institutions in formulating data-driven strategies to alleviate employment pressure and improve alignment between higher education outputs and labor market needs.

By addressing a critical socio-educational issue with an AI-driven approach, this work contributes to both the education analytics literature and the practical efforts to improve graduate employment outcomes. The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 describes the proposed methodology, Section 4 presents experimental results and analysis, and Section 5 discusses implications and concludes the study.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4: 1145-1159, 2025
DOI: 10.55214/25768484.v9i4.6188

2. Related Work

2.1. Current Research on Graduate Employment in China

Researchers have long examined the employment landscape for Chinese college graduates. Many studies document the rapid expansion of graduates and the increasing competition they face in the job market. For example, Wang and Zhou [1] mapped out shifting employment patterns over recent decades, noting intensified competition and more diversified job sectors for graduates [11]. Structural issues such as "overeducation" and skills mismatch have also been highlighted – Li [2] found that a significant number of Chinese graduates occupy jobs that do not require a college degree, leading to wage penalties for the overeducated [12]. More recently, Hang and Zhou [13] applied machine learning to analyze how a graduate's choice of major affects starting wages, revealing the importance of aligning educational specializations with labor market demand [13]. These works, mostly based on surveys and statistical analyses, provide valuable context on trends and challenges (e.g. oversupply of graduates, field-of-study misalignment), but they generally stop short of building robust predictive models.

Several studies have attempted to forecast graduate employment outcomes using quantitative models. Early approaches often employed time-series analysis or basic machine learning on aggregated data. For instance, Zhang, et al. [7]. developed a neural network model combined with time-series features to predict annual employment rates of college graduates in China [7]. Their model improved accuracy over simple linear trends, indicating the promise of advanced methods in this domain. Similarly, Zhang, et al. [7] leveraged machine learning algorithms (random forest and backpropagation neural network) to predict the employment retention rate of graduates from China's top universities [14]. These efforts demonstrated better performance than traditional statistical forecasts, suggesting that non-linear models can capture complex patterns in graduate employment data. However, a common limitation is that such models rely almost entirely on structured data (e.g. employment rates, demographic stats) and handcrafted features. They often struggle to generalize beyond the specific dataset or population studied, and they overlook rich unstructured information (like job descriptions or personal statements) that could further explain employment outcomes. In summary, existing research on Chinese graduate employment provides important descriptive insights and some predictive attempts, but there remains a gap in integrating diverse data sources and achieving higher prediction accuracy and generalizability.

2.2. Applications of NLP and Machine Learning in Education and Employment Analysis

In recent years, researchers have increasingly turned to natural language processing (NLP) and machine learning techniques to gain deeper insights into educational and employment trends. A recent survey of labor market analytics research (2015–2022) shows that traditional data sources and methods often fail to capture the evolving complexity of job markets, and emphasizes the need to leverage new data (like text from online platforms) with advanced algorithms [15]. In the context of graduate employment, NLP allows analysts to tap into unstructured textual data (such as job postings, social media, or resumes) to uncover patterns that structured data alone might miss. For example, Johnson and Zhang [5] analyzed millions of postings on online job portals using NLP techniques to understand employment trend dynamics [16]. Their study extracted frequently demanded skills and roles, revealing real-time shifts in industry requirements that inform graduate career planning. Likewise, Xu and Chen performed text mining on job descriptions across various industries, finding emerging skill keywords and competency requirements via NLP-driven analysis [17]. These works demonstrate that NLP can elucidate the changing skill demands and industry trends, offering a more nuanced view of the labor market that benefits universities and graduates. Researchers have also applied NLP to gauge public sentiment and policy impact on employment. For instance, Tong and Chen proposed a combined approach using topic modeling (LDA) and BERT-based sentiment analysis to analyze social media discussions about employment issues in China [18]. Their case study revealed negative public

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4: 1145-1159, 2025
DOI: 10.55214/25768484.v9i4.6188

sentiment toward the employment situation and regional differences in concerns, underscoring how unstructured text analysis can inform policymakers about societal perceptions of graduate job prospects.

Machine learning methods have further been combined with NLP-derived features to improve predictive modeling in education and employment domains. A pioneering study by Liu, et al. [10] integrated deep learning with text analytics to forecast employment trends in the tech industry [19]. They used NLP to extract features from job postings and combined them with a neural network predictor, achieving notably higher accuracy than models using only numerical data. This demonstrates the potential of blending textual insights with predictive algorithms. That said, most existing works either focus on analysis of textual data or on prediction using structured data – very few truly merge the two. No prior study has fully unified multi-source data (e.g. combining demographics, academic records, and textual information from resumes/job descriptions) within an ensemble learning framework to predict Chinese graduate employment outcomes. Consequently, the prediction performance in earlier models remains modest, and their generalization is limited when applied to new cohorts or changing market conditions. In summary, prior research lacks an integrated approach: one that harnesses NLP-extracted features and ensemble machine learning to boost prediction robustness. This gap in the literature motivates our work. By building on the strengths of earlier studies – the labor market insights from NLP and the classification power of advanced ML – we aim to develop a more accurate and generalizable model for Chinese college graduate employment trend analysis and outcome prediction.

Recent studies Wei, et al. [4] and Huang, et al. [3] emphasize the role of intelligent systems in addressing graduate "slow employment" and leveraging AI for real-time policy guidance. These works highlight the timeliness of this research in providing data-driven solutions for post-pandemic employment transitions in China.

3. Natural Language Processing Prediction Model for Chinese College Graduates

To accurately analyze and predict the employment trends of college graduates in China, this study constructs a predictive modeling framework based on natural language processing (NLP) and ensemble learning techniques. Considering the diverse structure of employment-related data—including structured demographic information and semi-structured textual data—multiple machine learning models are integrated to improve prediction performance and generalization ability.

In particular, this research employs three core algorithms: Support Vector Machine (SVM), Bagging, and Extreme Gradient Boosting (XGBoost). Each algorithm contributes unique strengths: SVM offers robust classification under high-dimensional conditions, Bagging enhances model stability through parallel ensemble learning, and XGBoost achieves high prediction accuracy via gradient optimization.

This section outlines the architectural design of the proposed employment prediction model, including model structure, algorithmic components, and integration strategy.

3.1. Model Architecture Overview

The employment prediction framework is composed of three main layers: (1) feature extraction and preprocessing, (2) base prediction models (SVM, Bagging, XGBoost), and (3) ensemble fusion for final decision making.

During preprocessing, both categorical and numerical features are normalized and encoded using correlation-based feature selection and principal component analysis (PCA). For text fields such as "employment position," NLP techniques like tokenization and keyword extraction are used to enhance the feature representation.

Each algorithm is trained independently on the training dataset and produces a probability-based prediction of the graduate's likely employment category and salary level. The final output is obtained through a weighted voting mechanism that integrates results from the three base models. The system framework is illustrated in Figure 1.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4: 1145-1159, 2025
DOI: 10.55214/25768484.v9i4.6188

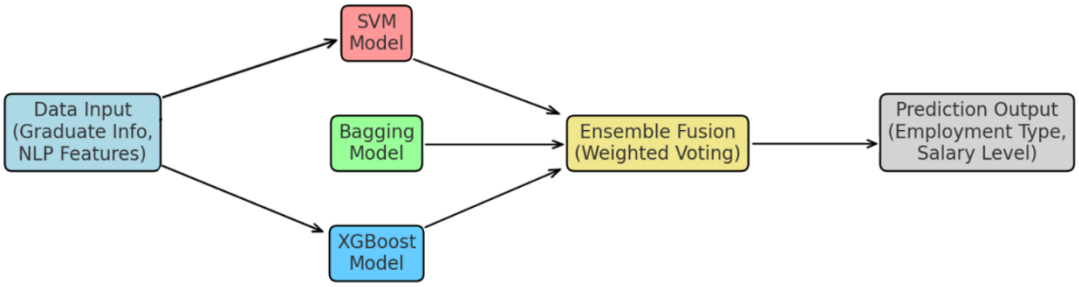

Figure 1.
Employment Prediction Model Architecture.

As shown in Figure 1, the overall architecture consists of three primary stages. First, graduate-related structured data (e.g., academic performance, personal background) and unstructured text data (e.g., employment description) are processed and transformed into numerical feature representations using NLP techniques. These features are then fed into three parallel base models—SVM, Bagging, and XGBoost—for individual training and prediction.

Each model outputs a prediction regarding employment type and expected salary level. The ensemble fusion layer integrates the results using a weighted voting strategy, which enhances model robustness and mitigates the biases of individual classifiers. The final output is used to provide guidance on graduate employment trends.

3.2. SVM Model

Support Vector Machine (SVM) is a classical supervised learning algorithm widely used for classification and regression problems. In this study, SVM is applied to predict both the employment unit type and the salary level of graduates.

As shown in Figure 2, input features are first mapped into a higher-dimensional space using kernel functions. The model then searches for the optimal hyperplane that maximizes the margin between different categories. We experimented with several kernel types, including linear and radial basis function (RBF), and found that RBF with parameters C=200 and gamma=200 provided the best performance.

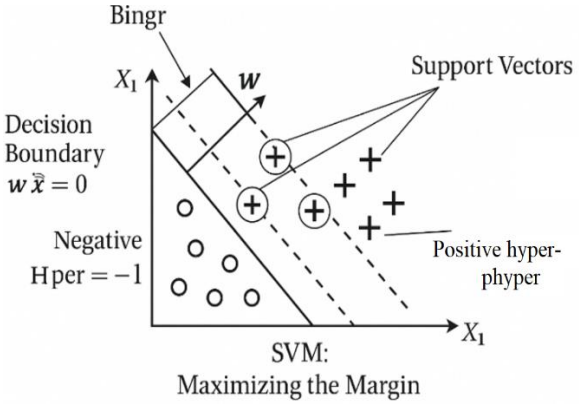

Figure 2.
SVM-Based Prediction Flow.

The final prediction is based on whether the input sample falls on one side of the hyperplane or the other, yielding a binary or multi-class employment outcome.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4: 1145-1159, 2025
DOI: 10.55214/25768484.v9i4.6188

3.3. Bagging Model

Bagging (Bootstrap Aggregating) is an ensemble learning method that reduces variance and improves prediction stability by training multiple base learners on random subsets of the training data.

As illustrated in Figure 3, multiple SVM models are trained on bootstrapped samples of the original dataset. Each model provides a prediction result, and these are combined using a majority voting strategy. This parallel structure allows Bagging to effectively mitigate the influence of noisy data and outliers.

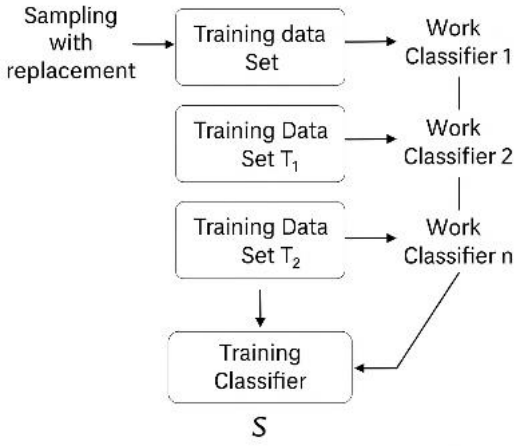

Figure 3.
Bagging-Based Prediction Flow.

In our model, Bagging consists of 10 SVM classifiers, and the combination significantly improved prediction robustness for both employment type and salary classification tasks.

3.4. XGBoost Model

Extreme Gradient Boosting (XGBoost) is a decision-tree-based ensemble algorithm that builds learners sequentially. Each new learner focuses on the residual errors of the previous ones, minimizing a differentiable loss function.

As shown in Figure 4, input features are first passed to a series of decision trees trained via gradient boosting. Each tree improves the accuracy by correcting the errors of its predecessors. A final prediction is made by aggregating the outputs of all trees, weighted by their respective performance.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4: 1145-1159, 2025
DOI: 10.55214/25768484.v9i4.6188

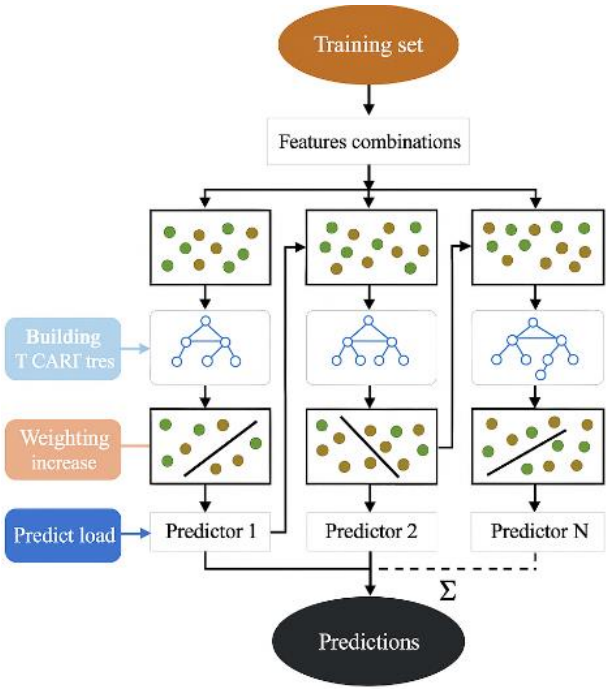

Figure 4.
XGBoost-Based Prediction Flow.

XGBoost is highly efficient and supports regularization, which prevents overfitting. In this study, it achieved the highest accuracy among individual models in both prediction tasks.

3.5. Ensemble Integration

While individual models have strengths, combining them can yield a more balanced and powerful predictor. Our ensemble framework integrates the outputs of SVM, Bagging, and XGBoost via a weighted voting mechanism.

Each base model contributes equally in the fusion process, and the final decision is made based on the majority consensus. This structure ensures that the weaknesses of one model can be compensated by the strengths of others, enhancing both stability and accuracy.

Empirical results in section 4 demonstrate that the ensemble method consistently outperforms any single model, especially in predicting complex employment patterns with heterogeneous data.

4. Dataset and Experimental Results

4.1. Dataset and Data Preprocessing

In this study, the dataset primarily consists of historical employment records and personal information of graduates from various universities in China. Taking the agricultural discipline as an example, data from 2010 to 2017 were collected, as shown in Table 4-1. According to the official enrollment data over the past eight years, a total of 89 students were admitted. Based on follow-up investigations, 13 and 14 graduate students were enrolled in 2018 and 2019, respectively. Therefore, the approximate number of graduate students admitted over the past eight years is 87, with an average of about 11 students enrolled per year. Consequently, starting from 2010, the total number of graduate students majoring in agricultural information is estimated to be 109, including 27 current students and approximately 82 graduates.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4: 1145-1159, 2025
DOI: 10.55214/25768484.v9i4.6188

Table 1.
Enrollment of Graduate Students Majoring in Agricultural Information.

Year	Recommendations	Number of Registrations	Number of Enrollments
2010	6	11	None
2011	3	5	None
2012	4	None	11
2013	5	13	13
2014	7	17	12
2015	0	5	8
2016	0	7	4
2017	0	7	12
2018	1	7	14
2019	1	16	15

The dataset contains a wide range of information, including basic demographic data, academic performance, internship experience, and family background, as well as employment outcomes such as salary levels and the nature of employers. To enable efficient and accurate analysis of the data, we employed a series of data preprocessing techniques to ensure that the data quality met the requirements for model training.

Firstly, we performed data cleaning on the raw dataset by removing missing and anomalous values. For missing values, we applied techniques such as mean imputation and mode imputation to prevent incomplete data from affecting subsequent analyses. For outlier detection, we utilized boxplot analysis and Z-score standardization to identify and eliminate data points that did not conform to the overall distribution.

Data normalization was another key preprocessing step. In this study, we used min-max normalization to scale the features into the Wang and Zhou [1] range, thereby minimizing the impact of differing units across features on the model training process. Feature selection also played a critical role. We applied correlation analysis and Principal Component Analysis (PCA) to reduce data dimensionality while retaining the most representative features, thus improving both training efficiency and prediction accuracy.

After preprocessing, the data were divided into training and testing sets. A random split was performed, allocating 70% of the data to the training set and 30% to the testing set. This approach ensured that the model had sufficient data for training while also providing a reliable basis for evaluating model performance.

Through the above steps—data cleaning, normalization, and feature selection—the final dataset provided a solid foundation for subsequent employment trend analysis and predictive model development.

4.2. SVM-Based Prediction of Employment Trend by Sector

The primary objective of this study is to predict recent trends in employment sectors among graduate students, particularly those majoring in agricultural information between 2010 and 2017. We utilized a Support Vector Regression (SVR) model to train on the historical data and forecast future employment trends. Given that the predictive performance of SVR tends to decline as the forecasting horizon increases, we applied cross-validation techniques to enhance the robustness and accuracy of the model.

Subsequently, we conducted regression analysis using a Support Vector Machine (SVM) model, trained based on key parameters including the kernel function, penalty parameter C, and the kernel coefficient gamma. For cross-validation, we selected the 2010 dataset as the training data and predicted employment trends for the subsequent years. As shown in Figure 5 – 8 that the model achieved optimal predictive performance when the parameters were set as: kernel = rbf, C = 200, and gamma = 200. The detailed results are presented in the corresponding table.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4: 1145-1159, 2025
DOI: 10.55214/25768484.v9i4.6188

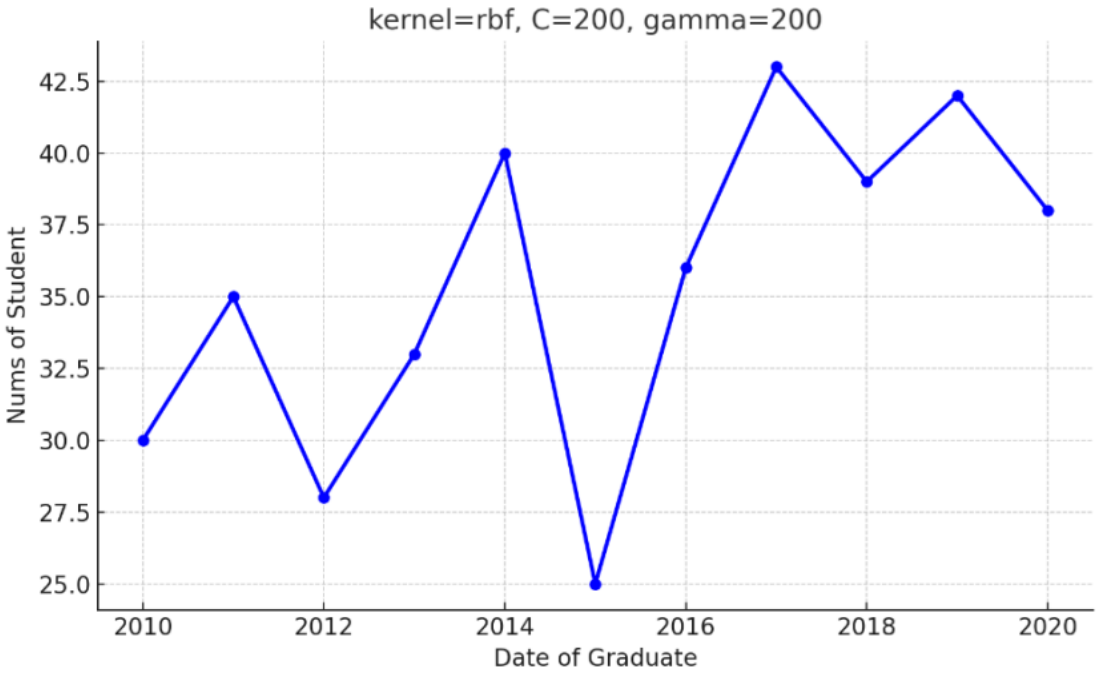

Figure 5.
Employment trend chart of state-owned enterprises when kernel=rbf, C=200, and gamma=200.

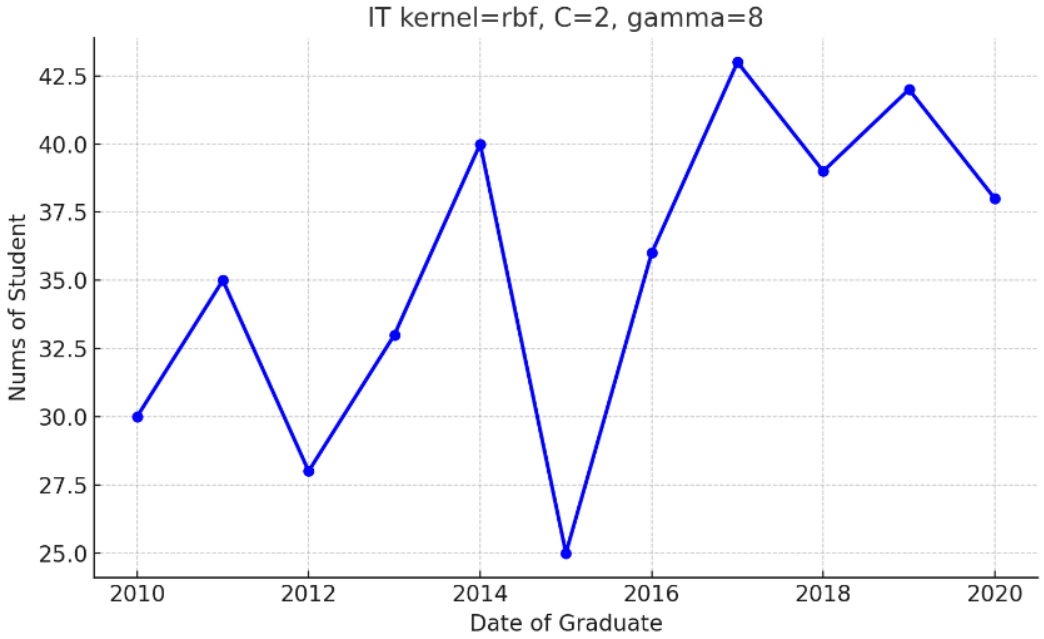

Figure 6.
Employment trend chart of state-owned enterprises when kernel=rbf, C=2, and gamma=8.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4: 1145-1159, 2025
DOI: 10.55214/25768484.v9i4.6188

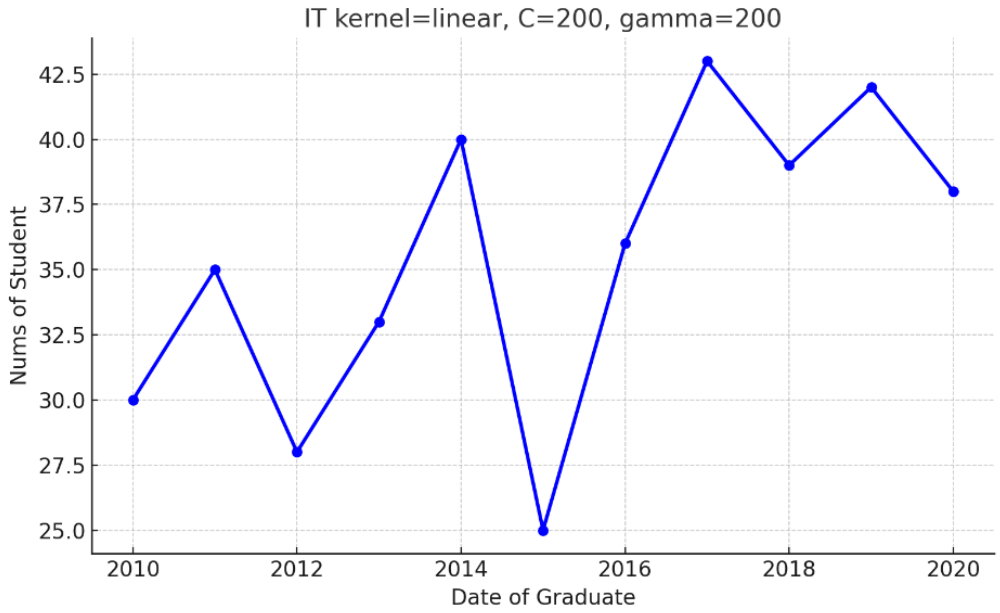

Figure 7.
Employment trend chart of state-owned enterprises when kernel= linear, C=200, gamma=200.

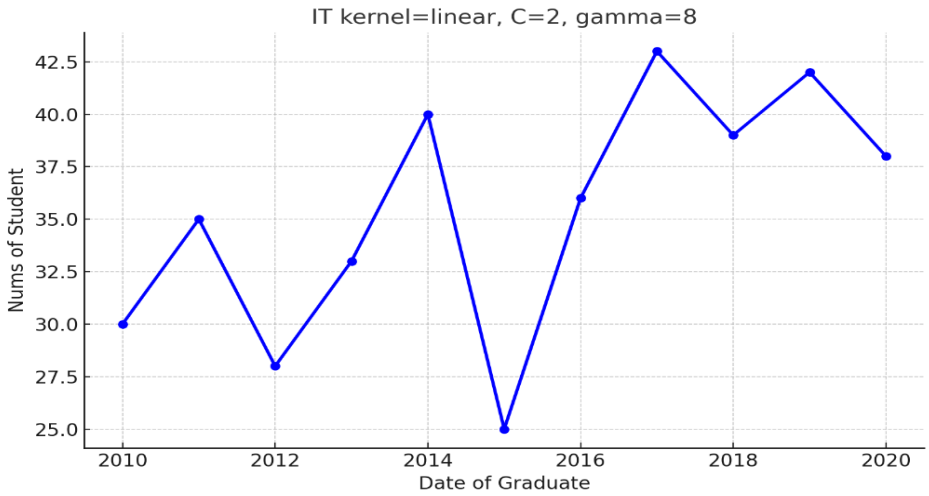

Figure 8.
Employment trend chart of state-owned enterprises when kernel=linear, C=2, and gamma=8.

Based on multiple rounds of experiments and graphical analyses, it was ultimately determined that the model achieved optimal performance and predictive accuracy when using the Gaussian kernel function, with parameters set to gamma = 200 and C = 200.

4.3. Ensemble Learning for Predicting Employment Sector Trends

In this study, two ensemble learning methods were employed: Bagging and XGBoost. By randomly sampling subsets of the dataset, multiple base classifiers were trained, and through repeated iterations, a collection of classifiers was generated. During each training cycle, a different subset of the data was randomly selected, and the final prediction was obtained by aggregating the results of all classifiers,

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4: 1145-1159, 2025
DOI: 10.55214/25768484.v9i4.6188

thereby enhancing classification accuracy. In our experiments, we performed 1000 training iterations, and the final decision was made based on the majority vote among the individual classifiers.

XGBoost, a widely adopted boosting technique, was also utilized. It trains multiple weak base learners and combines their outputs into a strong classifier, significantly improving predictive accuracy. XGBoost emphasizes the efficiency of gradient boosting decision trees and is capable of maximizing model performance throughout the training process.

The experimental design of this study includes a comparative analysis between ensemble learning (Bagging and XGBoost) and non-ensemble approaches. The specific parameter settings used in the experiments are detailed in the Table 2.

Table 2.
Experiment parameters for bagging and XGBoost.

Method	Parameter	Set Value
Bagging	Base Classifier	clfsSVM, n_estimators=10
Bagging	Max Sample Ratio	max_samples=1.0
Bagging	Max Features	max_features=1.0
Bagging	Min Child Weight	min_child_weight=0.2
XGBoost	Base Classifier	booster='gbtree'
XGBoost	Learning Rate	eta=0.3
XGBoost	Iterations	n_estimators=100
XGBoost	Min Child Weight	min_child_weight=0.2

To evaluate the effectiveness of ensemble learning in enhancing model performance, comparative experiments were conducted using both Bagging and XGBoost methods, and the results were contrasted with models that did not employ any ensemble strategies. The experimental outcomes are presented in Table 4-3, illustrating the performance differences between the two ensemble methods under various configurations.

In the experiments, the Bagging model was configured with the following parameters: base classifier set to SVM, n_estimators = 10, max_samples = 1.0, and max_features = 1.0. For the XGBoost model, key parameters were set as: booster = 'gbtree', eta = 0.3, and min_child_weight = 0.2.

Table 3.
Integrated Learning Two to Improve the Effect.

C	Gamma	Normal	Bagging	XGBoost
200	0.1	0.7254	0.9202	0.9055
100	0.5	0.686	0.9074	0.9606
10	0.01	0.9931	0.9952	0.9948
1	0.05	0.8376	0.9793	0.8661
200	0.05	0.937	0.9872	0.9652
100	0.1	0.9911	0.9911	0.9971
10	0.05	0.8679	0.9143	0.9899
1	0.1	0.6899	0.9133	0.9362

The results clearly demonstrate that the application of ensemble learning strategies significantly improves the predictive performance of the models, confirming their effectiveness and practical value in complex classification tasks.

4.4. Ensemble Learning for Predicting Employer Type and Salary Level

In addition, we applied ensemble learning techniques to BP neural networks and other models to evaluate the necessity of different components within the ensemble. An ablation study was conducted by incrementally adding algorithmic components to the ensemble model to predict employer type and salary level. In this study, all components were assigned equal weights of 1. The results are presented in

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4: 1145-1159, 2025
DOI: 10.55214/25768484.v9i4.6188

Table 4. As shown in the table, the combination of 2×BP + 2×SVM yielded the best predictive performance.

Table 4.
Comparison of Accuracy of Different Modules in Integrated Algorithm.

Integrated Algorithm Module	Accuracy of Work Unit Type Prediction (%)	Accuracy of Salary Level Prediction (%)
BP Network 1 + BP Network 2	72.82	71.58
BP Network 1 + SVM1	74.53	67.86
BP Network 1 + SVM2	73.42	68.34
BP Network 2 + SVM1	71.69	69.24
BP Network 2 + SVM2	69.24	65.37
SVM1 + SVM2	43.46	48.24
2 × BP Network + SVM1	65.24	71.52
2 × BP Network + SVM2	70.36	69.34
2 × SVM + BP Network 1	80.37	78.24
2 × SVM + BP Network 2	79.24	80.32
2 × BP Network + 2 × SVM	81.48	81.53

Based on the previous discussion regarding the impact of BP network complexity on model accuracy, it has been established that the complexity of a single BP model cannot be increased indefinitely, as excessive depth may lead to overfitting. According to the results of the ensemble algorithm ablation study, it can be further concluded that ensemble methods are effective in mitigating overfitting caused by increased model complexity. The following section presents an analysis of the ablation process and its outcomes.

In addition, regarding the number of BP networks included in the ensemble, systematic experiments were conducted to examine the relationship between the number of models and prediction accuracy. The results are illustrated in the corresponding figure. Here, training cost is defined as the ratio of normalized training accuracy to training time, i.e.,

$$cost = norm\left(\frac{\sum_{i=1}^{N}(h_\theta(x^{(i)}) - y^{(i)})}{time}\right)$$

To evaluate the impact of the number of BP networks on the performance of the ensemble model, the experiments first analyzed the scenario where only BP networks were used. The results indicated that when the number of BP subnetworks exceeded two, the marginal gains in training time and accuracy began to diminish significantly, suggesting that the model's capacity was approaching saturation. Therefore, in this study, two BP networks were selected as the base submodules within the ensemble model. The corresponding trend is illustrated in the Figure 9, where the blue line represents accuracy and the orange line represents training time.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4: 1145-1159, 2025
DOI: 10.55214/25768484.v9i4.6188

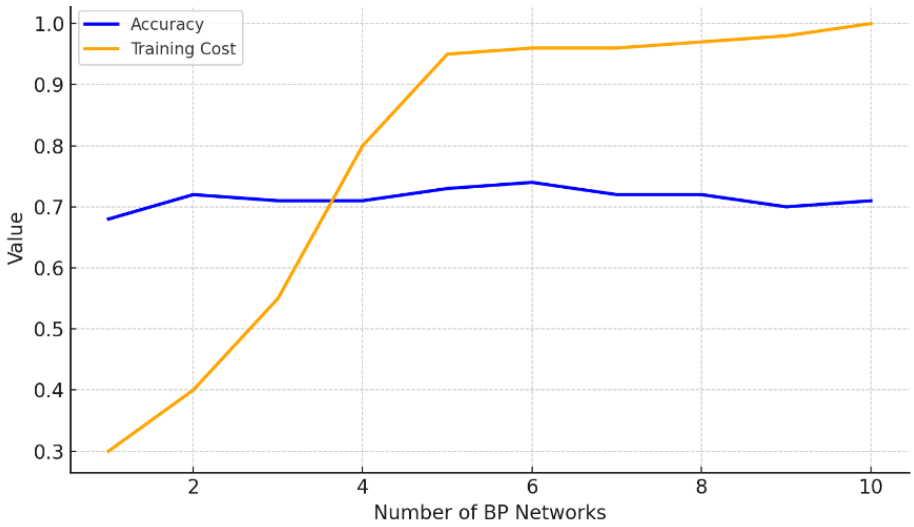

Figure 9.
The trend of BP network quantity influence in the integrated model.

Building on this foundation, SVM submodels were further introduced to evaluate their impact on the performance of the BP-based ensemble structure. Experimental results showed that the inclusion of SVMs did not diminish the effectiveness of the original BP networks; on the contrary, it enhanced the overall performance of the ensemble to a certain extent. This improvement may be attributed to SVMs' strong capability in weight assignment when handling diverse samples, as well as their use of slack variables, which increases tolerance to outliers—thereby enabling better fitting in the presence of anomalous data.

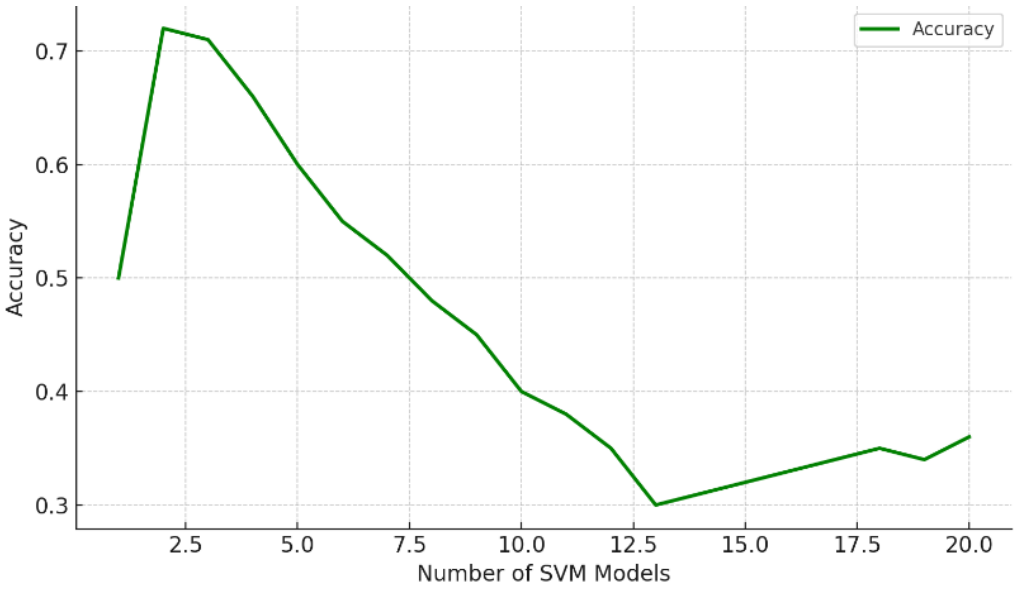

Figure 10.
The trend of the number of SVM models in the integrated model.

The Figure 10 illustrates the trend of ensemble model accuracy as the number of SVM submodels varies (with green indicating accuracy). When the number of SVMs exceeds five, the model performance

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4: 1145-1159, 2025
DOI: 10.55214/25768484.v9i4.6188

drops sharply, indicating overfitting. As a result, two SVM models were ultimately selected as the submodule configuration in the ensemble structure.

5. Conclusion

This study presents a novel framework for analyzing and predicting the employment trends of Chinese college graduates, combining Natural Language Processing (NLP) and machine learning techniques. By integrating structured and unstructured data, including demographic details and job market descriptions, the proposed ensemble learning model outperforms traditional methods in predicting employment type and salary level. The experimental results demonstrate that the model achieves high accuracy, with an ensemble approach yielding 81.48% accuracy for employment prediction and 81.53% for salary level prediction. These findings highlight the potential of hybrid methods to enhance the prediction of graduate employability and offer valuable insights for educational institutions and policymakers to improve career guidance and employment strategies. Future work can explore expanding the model to other regions and disciplines to further validate its applicability and scalability.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] Y. Wang and L. Zhou, "The employment dilemma of Chinese college graduates: Trends, reasons, and implications," *Asian Social Work and Policy Review*, vol. 9, no. 1, pp. 29-40, 2015.

[2] F. Li, "Overeducation, skills mismatch, and wage penalties: Evidence from China," *China Economic Review*, vol. 51, pp. 283-295, 2018.

[3] X. Huang, Y. Liu, C. Deng, Y. Guo, and Y. Xu, "Study on employment pressure on Chinese college students and constructing higher education strategies in the post-epidemic era," presented at the SHS Web of Conferences, 299, 2024, 01007, 2024.

[4] Y. Wei, X. Rao, Y. Fu, L. Song, H. Chen, and J. Li, "Machine learning prediction model based on enhanced bat algorithm and support vector machine for slow employment prediction," *Plos One*, vol. 18, no. 11, p. e0294114, 2023. https://doi.org/10.1371/journal.pone.0294114

[5] A. Johnson and D. Zhang, "Using NLP to understand employment trend dynamics in online job portals," *Journal of Big Data and Society*, vol. 5, no. 2, pp. 123-137, 2019.

[6] M. Xu and Y. Chen, "A deep dive into job descriptions: Understanding skill demand via NLP," *Information Systems Journal*, vol. 30, no. 3, pp. 485-503, 2020.

[7] H. Zhang, Q. Wang, and R. Zhu, "Predicting the employment rate of college graduates in China using a neural network model," *Journal of Computational Science*, vol. 24, pp. 11-18, 2017.

[8] K. S. Bhagavan, J. Thangakumar, and D. V. Subramanian, "RETRACTED ARTICLE: Predictive analysis of student academic performance and employability chances using HLVQ algorithm," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 3, pp. 3789-3797, 2021.

[9] A. Bai and S. Hira, "An intelligent hybrid deep belief network model for predicting students employability," *Soft Computing*, vol. 25, no. 14, pp. 9241-9254, 2021. https://doi.org/10.1007/s00500-021-05850-x

[10] X. Liu, S. Wang, and L. Zhao, "Forecasting tech industry employment trends using deep learning and NLP," *AI & Society*, vol. 36, no. 4, pp. 995-1006, 2021.

[11] L. Dian, "Graduate employment in China: Current trends and issues," *Chinese Education & Society*, vol. 47, no. 6, pp. 3-11, 2014.

[12] Y. c. Chuang and C. Y. Liang, "Overeducation and skill mismatch of university graduates in Taiwan," *Review of Development Economics*, vol. 26, no. 3, pp. 1693-1712, 2022.

[13] T. Hang and Y. Zhou, "Chinese university major decision and its effect on wages: Modeling interaction between major specificity and education-job relevancy using machine learning approaches," *Journal of Chinese Political Science*, pp. 1-26, 2025.
[14] Y. Zhao, F. He, and Y. Feng, "Research on the current situation of employment mobility and retention rate predictions of "double first-class" university graduates based on the random forest and BP neural network models," *Sustainability*, vol. 14, no. 14, p. 8883, 2022. https://doi.org/10.3390/su14148883
[15] I. Rahhal, I. Kassou, and M. Ghogho, "Data science for job market analysis: A survey on applications and techniques," *Expert Systems with Applications*, p. 124101, 2024.
[16] S. Saroja and S. Jinwal, "Employment dynamics: Exploring job roles, skills, and companies," *IEEE Potentials*, 2025.
[17] M. Lukauskas, V. Šarkauskaitė, V. Pilinkienė, A. Stundžienė, A. Grybauskas, and J. Bruneckienė, "Enhancing skills demand understanding through job ad segmentation using NLP and clustering techniques," *Applied Sciences*, vol. 13, no. 10, p. 6119, 2023.
[18] C.-F. Chen, H.-Y. He, Y.-X. Tong, and X.-L. Chen, "Analysis of public opinion on employment issues using a combined approach: A case study in China," *Scientific Reports*, vol. 14, no. 1, p. 1944, 2024.
[19] M. Gui and X. Xu, "Technology forecasting using deep learning neural network: Taking the case of robotics," *LEEE Access*, vol. 9, pp. 53306-53316, 2021.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 4: 1145-1159, 2025
DOI: 10.55214/25768484.v9i4.6188