Big Data Analysis and Forecast of Employment Position Requirements for College Students

https://doi.org/10.3991/ijet.v18i04.38245

Yong Wei[1(✉)], Yinle Zheng[2], Ning Li[3]
[1]Admissions and Employment Division, Hebei Vocational University of Technology and Engineering, Xingtai, China
[2]Department of Clothing Engineering, Hebei Vocational University of Technology and Engineering, Xingtai, China
[3]Department of Architectural Engineering, Hebei Vocational University of Technology and Engineering, Xingtai, China
weiyong19760423@126.com

Abstract—With the help of natural language processing and machine learning, we can analyze the information of online recruitment text posted by employer companies and dig the requirement features of these employment positions, and this is a meaningful work with practical value. However, existing methods for analyzing online recruitment information are too simple to withstand the mass data on the Internet, so this paper aims to study the analysis and forecast of employment position requirements for college students based on big data analysis. At first, the recruitment information of companies is preprocessed, keywords in the recruitment text are extracted by the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm in the Python Chinese word segmentation toolkit, words in the recruitment text are segmented using the open source tool Word2Vec, and the uncounted words in the recruitment text are identified based on the Conditional Random Field (CRF) model. Then, this paper compares the occurrence probability of skills learnt by college students in a certain company employment position with the occurrence probability of the skills in all employment positions on the website, so as to find out the core skills learnt by college students that can match with the job positions required by companies. At last, this paper builds a XGBoost model to forecast the employment position requirements for college students, and verifies the validity of the model using experimental results.

Keywords—big data analysis, college student, employment position, requirement analysis, forecast

1 Introduction

Online recruitment has now become the top choice for employer companies to recruit talents, and its scale is still expanding over the years compared with the conventional job fairs hold by urban public employment service agencies [1–8]. Affected by economic situations and the increasing number of college graduates, young people are facing greater employment pressure as the employer companies are posing higher requirements for talents [9–17]. The leading cause of this phenomenon is that the talent cultivation goal of colleges and universities is disconnected with the recruitment requirements of companies, and the education institutions have failed to understand the job position and skill requirements contained in the recruitment information of companies [18–24]. Based on big data analysis, we can borrow the help of natural language processing and machine learning to analyze the semi-structured or unstructured information of online recruitment text posted by employer companies and dig the requirement features of these employment positions in a fast, efficient, and intelligent manner, and this is a meaningful work with practical value.

Scholar Wang [25] argued that with the comprehensive development of higher education in China, the number of college students is on the rise, the employment pressure of college graduates is increasing, and the employment situation is not optimistic. The author introduced an employment confidence index to analyze and predict the employment confidence of college students, which is of practical significance. Then, a model named GM-BPNN (Gray model-BP neural network) of subjective employment obstacles of college students and its predictive factors were constructed based on deep learning, and the results show that the relative error of the combined model is smaller and its accuracy is higher. Yao et al. [26] pointed out that many factors can affect college students' employment, so the prediction error of employment rate is usually high. In their work, the authors designed a prediction method for the said problem based on gray system. At first, they analyzed the research progress of employment rate prediction, found out the root of the prediction error, and collected the history data of college students' employment rate; then, they employed the gray system to fit the variation features of college students' employment rate, and built a prediction model for that. Wang et al. [27] analyzed the application situation of college student employment information management system, discussed the working principle of data mining, proposed the application of data mining in the analysis of the employment of college students, and summarized the special attention of data mining in the said analysis. Scholar Cheng [28] took the data of a university graduate as subject to figure out the influence of the student's academic performance, English level, and other activities on his/her employment development through literature review and analysis, and selected data on this basis; then, according to the features of college students' employment data, a deep-seated neural network with strong learning ability and adaptability was used to predict the employment situation, so as to provide guidance for college students. Moreover, combining with the actual data of graduates, the author analyzed the practical application of the prediction model, compared it with conventional machine learning algorithm, and verified the validity of the algorithm.

The existing analysis models of employment position requirements generally perform word segmentation and clustering on website recruitment information and they found that the requirements of employer companies mainly concentrate on technology development, sales, and finance. Overall speaking, these analysis methods are quite simple, they can not handle the massive Internet data, and their dictionary information sources are not as varied, which have seriously interfered with the prediction accuracy of employment position requirements. In view of these matters, this paper attempts to study the analysis and forecast of employment position requirements for college students based on big data analysis. In the second chapter, this paper preprocesses the recruitment information of companies, extracts the keywords in recruitment text using the TF-IDF algorithm in the Python Chinese word segmentation toolkit, segments the words in recruitment text using the open source tool Word2Vec, and identifies the uncounted words in recruitment text based on the CRF model. In the third chapter, this paper compares the occurrence probability of skills learnt by college students in a certain company employment position with the occurrence probability of the skills in all employment positions on the website, so as to find out the core skills learnt by college students that can match with the job positions required by companies. In the fourth chapter, this paper builds a XGBoost model to forecast the employment position requirements for college students. At last, validity of the proposed model is verified by experimental results.

2 Preprocessing of recruitment information

Different from the traditional recruitment methods, online recruitment has a few merits such as low cost, easy acquisition, and no time or space constraint. Though online recruitment, employer companies can receive resumes submitted by college students who meet their employment position requirements, furthermore, online recruitment systems can help them screen the received resumes quickly and efficiently. As a novel channel and mode of recruitment, online recruitment has a very good application prospect, however, there are two major problems with the current online recruitment: the low information authenticity, and the difficulty in information processing.

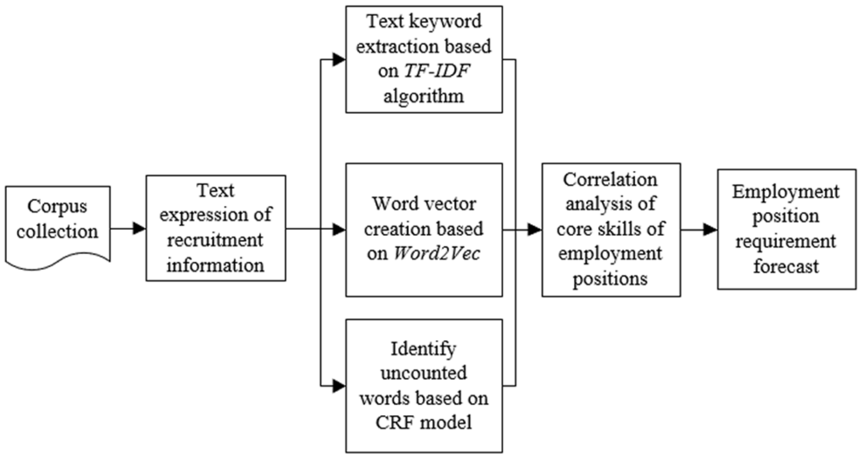


Fig. 1. Research plan of this paper

The recruitment information of companies mainly contains three parts: basic attributes of the employer company, basic information of the employment position, and the position description, specifically, the text information includes the company name, scale, industry, address, name of the position, required qualifications, salary, number of people required for the position, duty time, required education level, and required experience, etc. All these text information are the subjects to be studied in the employment position requirement analysis of college students.

Figure 1 shows the research plan of this paper. At first, the recruitment information of employer company is preprocessed, the recruitment text is processed by the *TF-IDF* algorithm in the Python Chinese word segmentation toolkit, and keywords in the text are extracted. Then, word frequency of keywords is counted using the Python script program, and keywords with a frequency higher than 1000 are extracted. In case that the recruitment text is fixed, the higher the word frequency, the greater the *TF* value of the *TF-IDF* algorithm. Assuming: $CO(q, c)$ represents the number of occurrences of a word q in the recruitment text c, SI represents the total number of words in the recruitment text, then the formula for calculating the *TF* value is:

$$TF = \frac{CO(q,c)}{SI(c)} \tag{1}$$

By taking the logarithm of the ratio of the number of recruitment text to the number of recruitment text containing a certain keyword, the *IDF* value of the *TF-IDF* algorithm could be attained, the formula for calculating the *IDF* value is:

$$IDF(q,C) = \log\left\{\frac{M - Total\ number\ of\ text}{1 + CO\left\{q\_in\_c_i \mid \{i = 1,2,3\ldots,M\}\right\}}\right\} \tag{2}$$

In the recruitment information, the higher the frequency of a word, the closer the *IDF* value is to 0, and *TF-IDF* is the product of *TF* and *IDF*.

After processed by the *TF-IDF* algorithm, this paper used the open source tool *Word2Vec* to perform word segmentation on the recruitment information, and the recruitment text after subjected to word segmentation was placed into the *Word2Vec* directory; then, the adopted *skip-gram* model was trained to get the word vectors of recruitment text under the current directory, and the *gesium* module based on *Python* was used to call the created word vectors to further attain several words that are semantically similar to the extracted keywords of recruitment text.

The *skip-gram* model was used to predict $e(q_i|q_o)$. Assuming: d represents the constant that determines the size of the context window, there are $o-d \le i \le o + d$ and $i \ne o$; $q_1, q_2, q_3, \ldots, q_o$ is a word sequence existing in the recruitment text, the objective of the adopted *skip-gram* model was to maximize the following formula, wherein O represents the size of the recruitment text set, then there is:

$$\frac{1}{O}\sum_{o=1-d\le j\le d}\sum \log\left(q_{o+j} \mid q_o\right)$$

The basic *skip-gram* model was used to define $e(q_0|q_R)$ as:

$$E(q_0|q_o) = \frac{p^{U_{q_0}^{O} U_{q_R}}}{\sum_{q=1}^{q} p^{U_q^O U_{q_R}}} \tag{3}$$

According to above formula, q_o must be in the window with q_l as the head word, that is:

$$\frac{1}{O}\sum_{o=1-d\le j\le d, j\ne 0}^{O}\sum \log e(q_{o+j}|q_o)=\frac{1}{O}\sum_{o=1-d\le j\le d, j\ne 0}\sum \log e(q_o|q_{o+j}) \tag{4}$$

The word segmentation accuracy of massive online recruitment text is mainly affected by uncounted words and ambiguous cutting, and the effect of the former is greater, uncounted words are a problem that can not be ignored with the word segmentation of recruitment text, to solve it, this paper introduced the CRF model to identify the uncounted words in the recruitment text, the model integrates the advantages of the maximum entropy model and the hidden Markov model, and it has a very good effect in identifying new words.

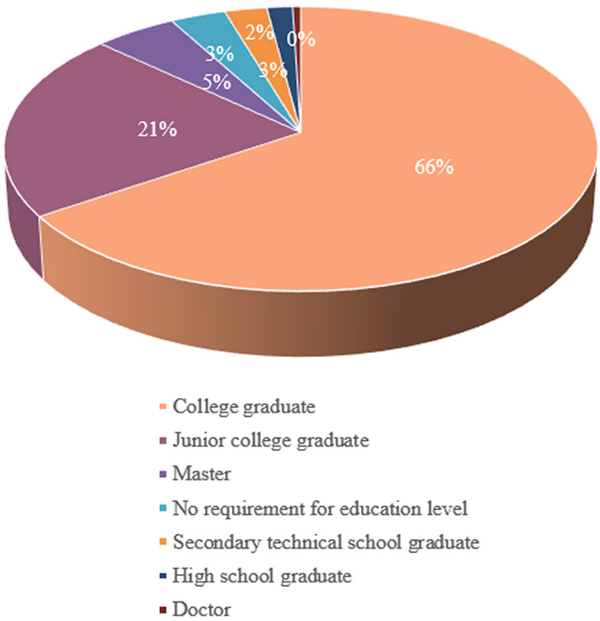


Fig. 2. Distribution of required education level of companies

Assuming: $A = \{a_1, a_2, \ldots, a_m\}$ represents the recruitment text sequence to be processed; $B = \{b_1, b_2, \ldots, b_m\}$ represents the hidden state sequence; $H = (U, P)$ is an undirected graph, U represents the nodes in the graph, P represents the connecting edges of nodes, then $B = \{B_u|u\in U\}$ represents the variable set created with nodes in H as indexes; if A is known and variables in B conform to Markov distribution, the CRF can be defined as $SC(A|B)$; assuming $C_{(a)}$ represents the the normalization factor, $G_l(b, a)$ represents the feature function, μ_l represents the weight parameter related to each feature g_l, then, based on the theory of CRF, there is:

$$SC(b_1,\ldots,b_m|a)=\frac{1}{C_{(a)}}\exp\left(\sum_{l=1}^{m}\mu_l G_l(b,l)\right) \tag{5}$$

Education level, work experience and salary level are information college students pay more attention to in the recruitment text. This paper plotted the distribution of education level, work experience, and salary level of the employment positions of companies based on the processed results of massive online recruitment text, as shown in Figures 2, 3, and 4, in terms of the lowest education level required by companies, the college graduate level takes the greatest proportion, 67.15% of the positions require a work experience of more than one year, and a salary level between 5k–15k is highly attractive to college students.

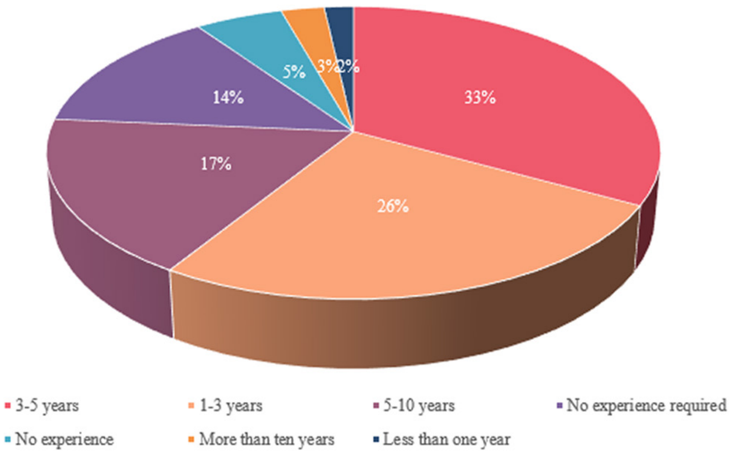


Fig. 3. Distribution of required work experience of companies

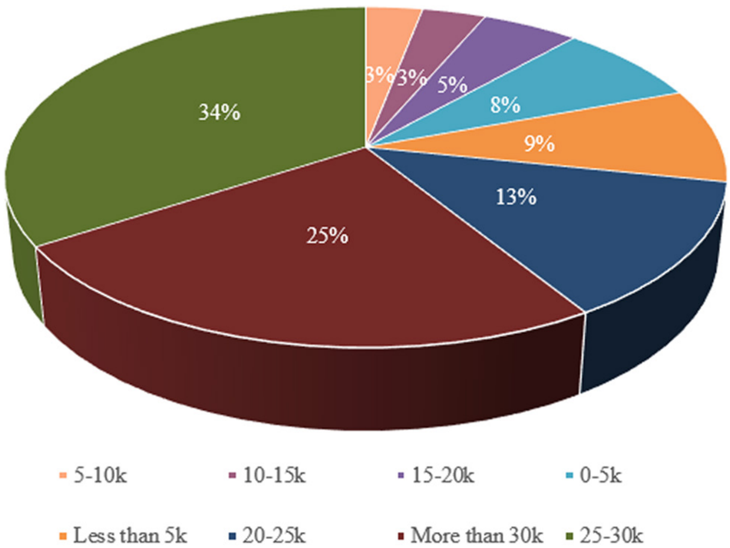


Fig. 4. Distribution of salary level of companies

3 Correlation analysis of core skills required by employment positions

The conventional requirement analysis methods of employment positions mainly count the frequency of the keywords of skills related to the positions and then analyze the differences between the features of employment positions of companies and the talent cultivation goals of schools and higher educational schools, the keyword frequency here is the *TF* and *IDF* values of the *TF-IDF* algorithm, which can indicate the number of occurrences of skills learnt by students during education in the online recruitment text, although this indicator can reflect the skills required by companies for the students, in terms of the research objective of this paper, mining the employment position text that is highly correlated with the skills required for college students is more valuable. The correlation here refers to the key indicator based on which the keywords entered by college students into the job search system during job hunting are included by the said system, that is, the inclusion of keywords in the job search system can be described by the correlation and keyword matching performance. This paper compared the occurrence probability of skills learnt by college students in a certain company employment position with the occurrence probability of the skills in all employment positions on the website, so as to attain the correlation between the skills learnt by college students and the employment positions of companies and to find out the core skills learnt by college students that can match with the employment position requirements of companies.

Assuming: S represents the correlation value, E_m represents the occurrence probability of a certain type of skill-related keywords l in a certain type of positions, E_M represents the occurrence probability of a certain type of skill-related keywords in the online recruitment text, m_l represents the frequency of a certain type of skill-related keywords l in a certain type of positions, m represents the number of online recruitment text of a certain type of positions, M_l represents the number of occurrences of a certain type of skill-related keywords l in the online recruitment text, M represents the number of online recruitment text, then the formula below gives the method for calculating the correlation degree:

$$S = \frac{E_m}{E_M} = \left(\frac{m_l}{m} \Big/ \frac{M_l}{M} \right) = \frac{m_l \times M}{m \times M_l} \tag{6}$$

If the value of S is greater than 1, it indicates that the correlation between employment positions and skills learned by college students is abnormal, and further investigation is needed. The greater the value of S, the greater the difference between the employment position requirements and the skills learned by college students, and this means that compared with other positions, this position has a higher requirement for a skill learned by college students, and companies are in need of talents with this skill.

4 Employment position forecast

According to the content of previous chapter, this paper took 5 core influencing factors in the college student employment position requirement forecast indicator system (including the number of specialized talents, the demand for specialized talents, the quality indicator of talent structure, the number of specialized talents added, and the number of specialized talents reduced) as the explanatory variables, and the number of employed college students as the explained variable to establish the *XGBoost* model. Some of the specific forecast indicators under the listed influencing factors are: the factor "number of specialized talents" can be sub-divided into the total number, the number of specialized talents in each department, the number of specialized talents in each region, the proportion of specialized talents in per 10,000 population, and the proportion of specialized talents in per 10,000 employees; the factor "demand for specialized talents" can be sub-divided into the total demand for specialized talents, the demand for specialized talents of each department, the demand for specialized talents of each region, the demand for specialized talents of per 10,000 population, and the demand for specialized talents of per 10,000 employees; the factor "quality indicator of talent structure" can be sub-divided into the the proportion of major types, the proportion of education level types, the proportion of professional title types, and the proportion of age types. Principles of the model and the construction ideas are introduced in detail as follows:

XGBoost is a *Boosting* optimization and integration algorithm based on gradient-boosted tree. The model attains the final forecast result of college student employment position requirements by summarizing the forecast scores of the leaf nodes of the random forest formed by multiple decision trees. Let $b_i^{(o)} = 0$ and assume the number of decision trees in the model is L, the l-th decision tree is represented by g_l, the eigenvector corresponding to the core influencing factor sample i is represented by a_i, and the leaf weight of indicator sample i on the l-th tree is represented by $g_l(a_i)$, then the following formula gives the expression of the forecast result:

$$\dot{b}_i^{(l)} = \sum_{l=1}^{L} g_l(a_i) \tag{7}$$

Assuming: b_i represents the real value of the i-th indicator sample, $\dot{b}$ represents the predicted value of the i-th indicator sample, n represents the total data volume of the l-th tree, L represents all trees created, then based on the conventional loss function and model complexity, the objective function of the model can be constructed as follows:

$$MO = \sum_{i=1}^{n} k(b_i, \dot{b}_i) + \sum_{l=1}^{L} \psi(g_l) \tag{8}$$

The first term of the above formula is the adjusted mean square error (MSE), namely the conventional loss function. The second term of the formula characterizes the complexity of the constructed model. The objective function of the model is solved as follows:

$$b_i^{(o)} = \sum_{l=1}^{o} g_l(a_i) = \dot{b}_i^{(o-1)} + g_o(a_i) \tag{9}$$

The following formula gives the expression of the loss function in the o-th iteration:

$$\sum_{i=1}^{n} k\left(b_i^{(o)}, \hat{b}_i^{(o-1)} + g_o(a_i)\right) \tag{10}$$

The second-order Taylor expansion of the above formula is:

$$\sum_{i=1}^{n} \left(k\left(b_i^{(o)}, \hat{b}_i^{(o)}\right) + g_o(a_i) * \frac{\partial k\left(b_i^{(o)}, \hat{b}_i^{(o-1)}\right)}{\partial \hat{b}_i^{(o-1)}} + \frac{1}{2}\left(g_o(a_i)\right)^2 * \frac{\partial^2 k\left(b_i^{(t)}, \hat{b}_i^{(o-1)}\right)}{\partial \left(\hat{b}_i^{(o-1)}\right)^2} \right) \tag{11}$$

Let $h_i = \partial k(b_i^{(o)},\hat{b}_i^{(o-1)})/\partial\hat{b}_i^{(o-1)}$, $f_i = \partial^2 k(b_i^{(o)},\hat{b}_i^{(o-1)})/\partial(\hat{b}_i^{(o-1)})^2$; $k(b_i^{(o)},\hat{b}_i^{(o-1)})$ is a known constant, then the above formula is simplified:

$$\sum_{i=1}^{n} \left(k\left(b_i^{(o)}, \hat{b}_i^{(o-1)}\right) + g_o(a_i) * h_i + \frac{1}{2}\left(g_o(a_i)\right)^2 * f_i \right) \tag{12}$$

Let $\sum_{l=1}^{o-1} \psi\ (g_l = 1 \psi\ (g_l)$, the formula below calculates the model complexity in the t-th iteration:

$$\sum_{l=1}^{o} \psi(g_l) = \sum_{l=1}^{o-1} \psi(g_l) + \psi(g_o) \tag{13}$$

The following formula gives the objective function in the o-the iteration:

$$MO^o = \sum_{i=1}^{n} \left(g_o(a_i) * h_i + \frac{1}{2}\left(g_t(a_i)\right)^2 * f_i \right) + \psi(g_o) \tag{14}$$

Assuming: O represents the the number of leaf nodes in a decision tree; j represents the index of each leaf node; θ_j represents the weight of samples on leaf nodes; α and μ are penalty coefficients, then the expression of the model complexity $\psi\ (g_o)$ is:

$$\psi(g_o) = \alpha O + \frac{1}{2} \mu \sum_{j=1}^{O} \theta_j^2 \tag{15}$$

To get the best forecast effect, the above two formulas are combined to attain the final objective function:

$$MO^o = \sum_{j=1}^{O} \left[\theta_j \sum h_i + \frac{1}{2} \theta_j^2 \left(\sum f_i + \mu \right) \right] + \alpha O \tag{16}$$

As can be seen from the above formula, MO^o is a quadratic equation with one variable about θ_j, and the value of θ_j that can minimize MO^o could be found;

let $H_j = \sum h_i$, $F_j = \sum f_i$, by calculating the extreme value of the objective function, it's known that when $\theta_j = -H_j/F_j\mu$, the value of MO^o is the smallest, that is:

$$Obj^o = -\frac{1}{2}\sum_{j=1}^{o}\frac{H_j}{F_j+\mu}+\alpha O \tag{17}$$

In this paper, 16 indicators in the established college student employment position requirement forecast indicator system that might affect the employment position requirements of college students were taken as inputs, the logarithm of the number of employed college students was taken as the output, and the training set, verification set, and test set of the model were constructed based on the data of 2001–2014, 2015–2016, and 2017–2021.

5 Experimental results and analysis

To verify the accuracy of the proposed keyword extraction algorithm, this paper compared it with other two keyword extraction algorithms: *LDA* and *Textrank*, and the comparison results are shown in Figure 5 and Table 1. Compared with *LDA* and *Textrank*, in most cases, the MAPE value of the keyword extraction error of the *TF-IDF* algorithm is 25.42% and 19.50% lower than the other two, and the average extraction accuracy has improved by 30.15% and 24.17%, respectively. Therefore, in terms of the correlation between the skills learnt by college students and the employment position requirements, the proposed algorithm can achieve a good effect in extracting keywords.

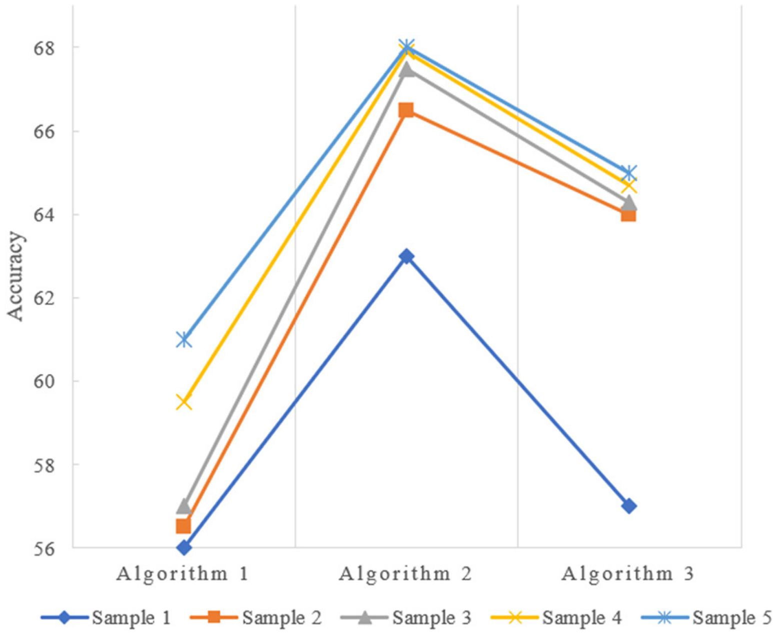


Fig. 5. Performance comparison of different keyword extraction algorithms

Table 1. Experimental results of different keyword extraction algorithms

Algorithm	*LDA*			*TF-IDF*			*Textrank*		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Sample 1	52.29	41.25	41.17	49.25	45.2	44.19	64.15	67.41	63.59
Sample 2	51.36	43.62	49.2	41.27	34.19	48.51	60.38	62.18	67.24
Sample 3	58.19	47.15	48.52	45.38	41.28	47.35	67.24	69.03	63.8
Sample 4	62.48	52.52	54.36	48.15	43.41	49.12	69.11	67.24	60.17
Sample 5	63.52	56.01	50.28	41.72	48.26	47.35	65.27	61.18	63.42

To verify the validity of the method for preprocessing recruitment information of companies proposed in this paper, ablation experiments were designed, and the preprocessing method combinations participated in the comparison included: the *TF-IDF* + CRF, *Word2Vec* + CRF, and *TF-IDF* + *Word2Vec* + CRF, the experimental results are shown in Figure 6 and Table 2. According to the data in the figure and table, completing the extraction of keywords, effective word segmentation, and identification of uncounted words are very important for extracting the features of recruitment information. Compared with the two combinations of *TF-IDF* + CRF and *Word2Vec* + CRF, the *TF-IDF* + *Word2Vec* + CRF can attain a lower *MAPE* value of recruitment information feature error in most cases, the average value has decreased by 45.21% and 39.27% respectively compared with the other two combinations, and the accuracy has improved by 35.67% and 46.87% on average.

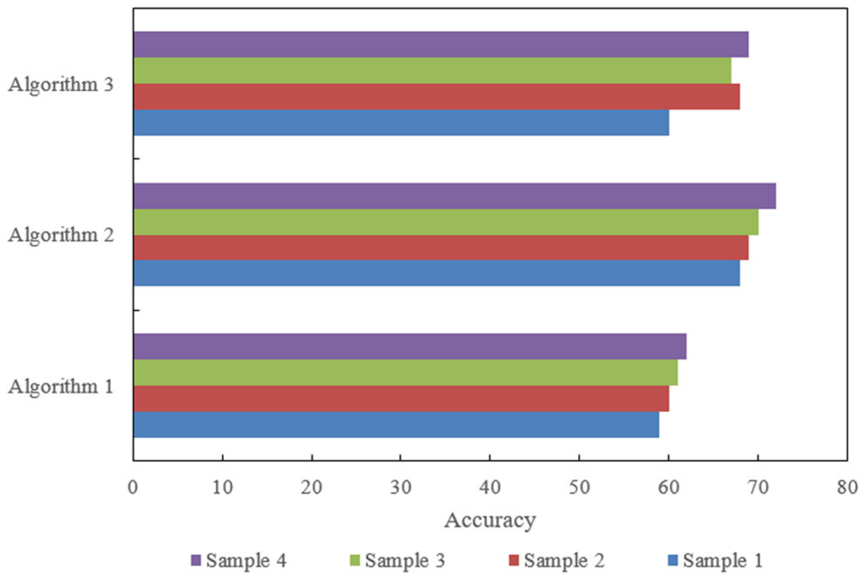


Fig. 6. Performance comparison of different text preprocessing methods

Table 2. Experimental results of different text preprocessing methods

Classification	*TF-IDF* + CRF			*Word2Vec* + CRF			*TF-IDF* + *Word2Vec* + CRF		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Sample 1	63.15	52.34	59.16	64.18	63.27	61.37	36.34	37.85	31.08
Sample 2	69.02	58.02	55.27	72.36	51.49	65.24	24.19	31.27	27.35
Sample 3	67.41	51.69	53.82	71.49	68.02	61.31	34.82	38.39	34.08
Sample 4	68.25	57.31	61.5	78.42	63.84	74.59	36.15	21.24	24.52

This paper constructed the college student employment position requirement forecast model based on the *XGBoost* model, and divided the training set and test set with a ratio of 4:1 based on the sample set of the online recruitment text of companies. To verify the validity of the forecast model, this paper designed a comparative experiment to compare its performance with the *Lasso* regression model, *LSTM* model, and *GA-LSTM* model. Figures 7 and 8 respectively give the predicted results of the *Lasso* regression model and the proposed model. It can be known from Figure 7, the *Lasso* regression model failed to well fit the overall trend of the number of employed college students, the deviation is large, on training set and test set, the *RMSE* value of core influencing factors is 1.4571, and the MAE value is 1.3426. Overall speaking, the forecast effect of the model is not satisfactory.

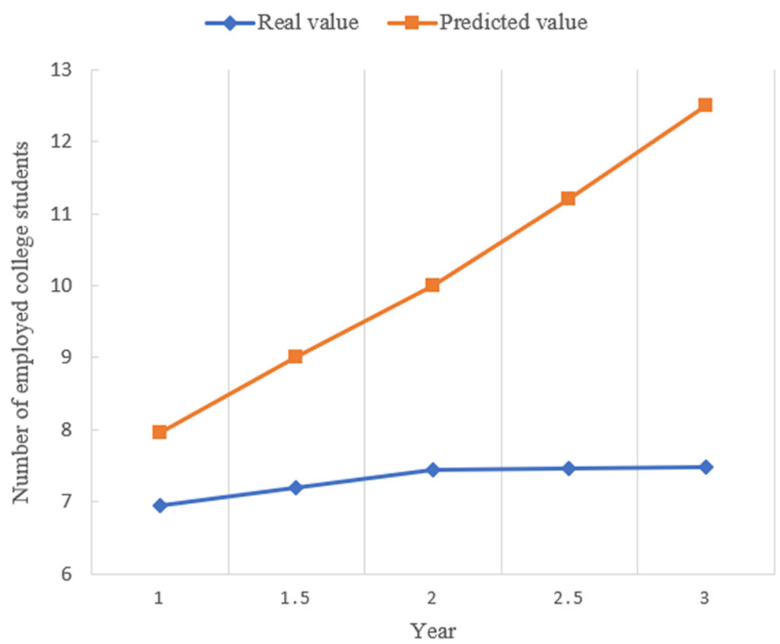


Fig. 7. Forecast results of the *Lasso* regression model

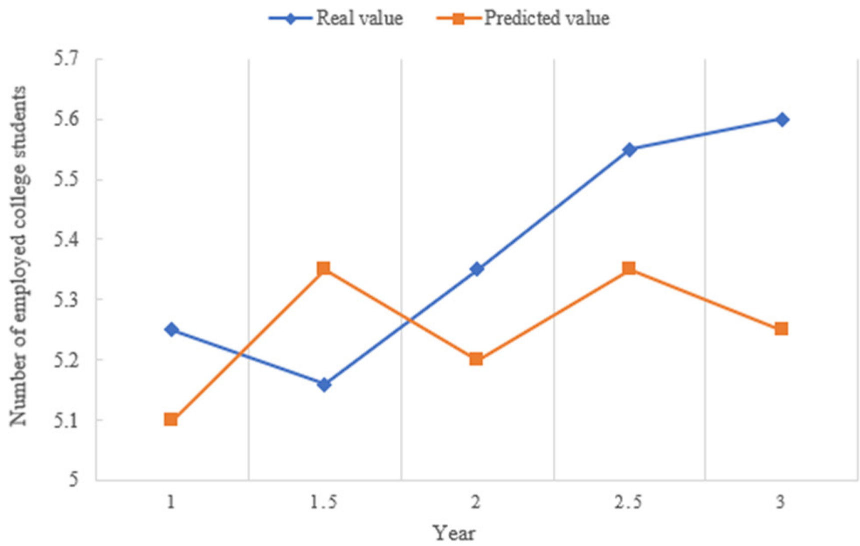


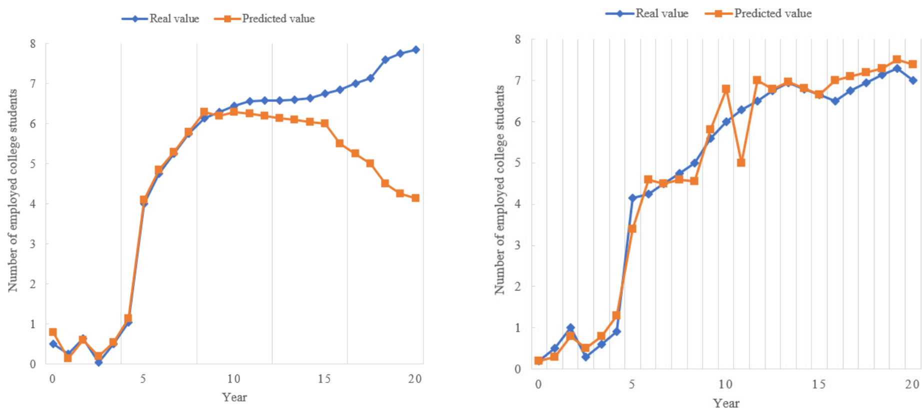
Fig. 8. Forecast results of the proposed model

Table 3. Forecast results of different models

Year	The Proposed Model	*Lasso* Regression	*LSTM*	*GA-LSTM*	Real Value
2011	1.102586	1.427416	3.417251	1.374851	1.01
2012	1.51473	1.928574	3.265982	1.623974	1.69
2013	2.517424	2.362517	7.352184	2.641528	2.53
2014	1.023598	1.302506	2.641279	1.302474	0.92
2015	1.352817	1.382547	3.251741	1.026182	1.47
2016	2.629581	3.201692	9.521486	5.032691	2.35
2017	51.30528	52.41726	42.35128	43.52174	52.6
2018	126.8475	132.581	63.52414	96.35128	124.01
2019	239.5741	162.3958	72.5147	136.5294	236.59
2020	368.5392	374.1203	114.6329	324.5018	374.16

According to Figure 8, the proposed model has fitted the overall rising trend of the number of employed college students, and the deviation between the predicted value and the real value is small. on training set and test set, the *RMSE* value of core influencing factors is 0.324, and the MAE value is 0.223. In contrast, the forecast effect of this model is more ideal than that of the *Lasso* regression model.

Table 3 lists the forecast results of the proposed model, *Lasso* regression model, *LSTM* model, and *GA-LSTM* model. Through comparison, it can be seen that the predicted value of the proposed model is closer to the real value, and its accuracy is higher than that of other models. This paper plotted the forecast results of 2011–2020 made by the *GA-LSTM* model which has a relatively good forecast effect, so as to more intuitively compare the performance of the two models, as shown in Figure 9, the forecast curve of the number of employed college students of the proposed model is basically consistent with the trend of the actual situation, while the forecast curve of the *GA-LSTM* model is obviously different from the curve of real values.



a) GA-LSTM **b) The proposed model**

Fig. 9. Forecast results of different models for 2011–2020

6 Conclusion

This paper studied the analysis and forecast of employment position requirements for college students based on big data analysis. At first, the recruitment information of companies was preprocessed, keywords in the recruitment text were extracted by the TF-IDF algorithm in the Python Chinese word segmentation toolkit, words in the recruitment text were segmented using the open source tool Word2Vec, and the uncounted words in the recruitment text were identified based on the CRF model. Then, this paper compared the occurrence probability of skills learnt by college students in a certain company employment position with the occurrence probability of the skills in all employment positions on the website, so as to find out the core skills learnt by college students that can match with the job positions required by companies. After that, this paper built a XGBoost model to forecast the employment position requirements for college students, and verified the validity of the model using experimental results. Combining with experiments, this paper compared the proposed keyword extraction algorithm with other two keyword extraction algorithms LDA and Textrank, gave the comparison results, and verified the accuracy of the proposed keyword extraction algorithm. Moreover, ablation experiments were designed, the performance of different text preprocessing methods was compared, and the validity of the method proposed in this paper for preprocessing the recruitment text of companies was verified. At last, the performance of the proposed model was experimentally compared with the Lasso regression model, LSTM model, and GA-LSTM model, and the experimental results show that, the forecast curve of the number of employed college students of the proposed model is basically consistent with the trend of the actual situation, and its performance is the best.

7 References

[1] August, T., Oliveira, N., Tan, C., Smith, N., Reinecke, K. (2018). Framing effects: Choice of slogans used to advertise online experiments can boost recruitment and lead to sample biases. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 2: 1–19. https://doi.org/10.1145/3274291

[2] Brandão, C., Morais, C., Dias, S., Silva, A.R., Mário, R. (2017). Using online recruitment: Implicit theories and candidates' profile. In World Conference on Information Systems and Technologies, 571: 293–301. https://doi.org/10.1007/978-3-319-56541-5_30

[3] Sureephong, P., Dahlan, W., Chernbumroong, S., Tongpaeng, Y. (2020). The effect of non-monetary rewards on employee performance in massive open online courses. International Journal of Emerging Technologies in Learning, 15(1): 88–102. https://doi.org/10.3991/ijet.v15i01.11470

[4] Urba, S., Chervona, O., Panchenko, V., Artemenko, L., Guk, O. (2022). Features of the application of digital technologies for human resources management of an engineering enterprise. Ingénierie des Systèmes d'Information, 27(2): 205–211. https://doi.org/10.18280/isi.270204

[5] Mahbub, S., Pardede, E., Kayes, A.S.M. (2022). Online recruitment fraud detection: A study on contextual features in Australian job industries. IEEE Access, 10: 82776–82787. https://doi.org/10.1109/ACCESS.2022.3197225

[6] Kaur, S., Sehgal, J.K., Grima, S. (2022). Virtual hiring: An effective green human resource management practice. International Journal of Sustainable Development and Planning, 17(6): 1699–1704. https://doi.org/10.18280/ijsdp.170602

[7] Hoang, P., Mahoney, T., Javed, F., McNair, M. (2018). Large-scale occupational skills normalization for online recruitment. AI Magazine, 39(1): 5–14. https://doi.org/10.1609/aimag.v39i1.2775

[8] Israfilov, N., Borisova, O., Kartashova, O., Davydova, N., Biserova, G., Gryaznukhin, A. (2020). Motivation and employee effectiveness in online learning environments: Leadership strategies of new generation and emotional intellect. International Journal of Emerging Technologies in Learning, 15(9): 258–279. https://doi.org/10.3991/ijet.v15i09.13921

[9] Liu, K. (2018). An analysis on the path of helping college students with difficult employment in micro media platform under intelligent system. In 2018 International Conference on Robots & Intelligent System (ICRIS), Changsha, China, pp. 238–241. https://doi.org/10.1109/ICRIS.2018.00068

[10] Tian, L. (2017). Simulation of college students' employment rate estimation model based on big data analysis. Boletin Tecnico/Technical Bulletin, 55(10): 437–443.

[11] Wang, J. (2017). Research of college students' employment data mining based on fuzzy decision tree (FDT). AGRO Food Industry Hi-Tech, 28(3): 2889–2893.

[12] Jiao, D. (2017). Study of the correlation between multiple intelligences theory and college students' employment motivation. AGRO Food Industry Hi-Tech, 28(3): 3277–3279.

[13] Zhu, C. (2017). Construction and research on the personalized employment recommendation system for college students based on Hadoop platform. Boletín Técnico, 55(8): 760–766.

[14] Xiang, L. (2017). Analysis of factors affecting college students' employment based on data mining algorithm. AGRO Food Industry Hi-Tech, 28(3): 1724–1728.

[15] Long, W., Jing, Q. (2017). Research on the construction of college students' employment guidance course from the perspective of multimedia. AGRO Food Industry Hi-Tech, 28(3): 2050–2054.

[16] Peng, G. (2017). Research on the employment guidance system of college students based on cluster analysis. AGRO Food Industry Hi-Tech, 28(3): 2885–2888.

[17] Xu, H., Zhang, Y. (2017). An analysis of the adjustment of industrial structure and college students' employment ability. AGRO Food Industry Hi-Tech, 28(3): 1502–1505.

[18] Guo, J., Suo, Z. (2017). Study on government management responsibility of college students' employment process based on big data analysis. Boletin Tecnico/Technical Bulletin, 55(19): 546–551.
[19] Qi, M.B. (2021). Clustering mining method of college students' employment data based on feature selection. In International Conference on Advanced Hybrid Information Processing, 105–115. https://doi.org/10.1007/978-3-030-94551-0_9
[20] Gao, J., Deng, F., Han, W. (2022). An algorithm for college students' employment tendency based on machine learning in a big data environment. In 2022 International Conference on Information System, Computing and Educational Technology (ICISCET), pp. 151–153. https://doi.org/10.1109/ICISCET56785.2022.00046
[21] Li, S. (2022). The use of machine learning model in the evaluation of college students' employment and entrepreneurship level. Wireless Communications and Mobile Computing, 2022: Article ID 8652349. https://doi.org/10.1155/2022/8652349
[22] Liu, X., Zhang, Y. (2022). Prediction of the impact of college students' behavior on employment rate. In CIBDA 2022; 3rd International Conference on Computer Information and Big Data Applications, Wuhan, China, 1–5.
[23] Li, N., Feng, L., Zeng, S., Bo, H. (202). Analysis on the regional spatial distribution of employment of college students and evaluation of employment quality. International Journal of Emerging Technologies in Learning, 17(19): 242–258. https://doi.org/10.3991/ijet.v17i19.34519
[24] Qi, M.B. (2021). An algorithm of employment resource allocation for college students based on social network mining. In International Conference on Advanced Hybrid Information Processing, 254–265. https://doi.org/10.1007/978-3-030-94551-0_21
[25] Wang, Y. (2022). Subjective employment obstacle of college students and its predictor model based on deep learning. Wireless Communications and Mobile Computing, 2022: Article ID 7955490. https://doi.org/10.1155/2022/7955490
[26] Yao, H., Dahlan Hj, A., Malek, M. (2021). Prediction of college students' employment rate based on gray system. Scientific Programming, 2021: Article ID 4182011. https://doi.org/10.1155/2021/4182011
[27] Wang, G., Cao, W., Cheng, Y. (2019). Application of data mining technology in college students' employment analysis. In International Conference on Applications and Techniques in Cyber Security and Intelligence, pp. 2087–2092. https://doi.org/10.1007/978-3-030-25128-4_276
[28] Cheng, P. (2022). Employment data screening and destination prediction of college students based on deep learning. Wireless Communications and Mobile Computing, 2022. https://doi.org/10.1155/2022/7173771

8 Authors

Yong Wei (1976–), male, born in Jiamusi, Heilongjiang Province, director and associate professor of the Admissions and Employment Division, Hebei Vocational University of Technology and Engineering, focuses on vocational education policy. Email: weiyong19760423@126.com. https://orcid.org/0000-0003-1022-4713

Yinle Zheng (1976–), male, born in Longyao, Hebei Province, associate professor of Department of Clothing Engineering, Hebei Vocational University of Technology and Engineering, focuses on education management. Email: zylkgd123@126.com. https://orcid.org/0000-0001-9908-8716

Ning Li (1981–), male, born in Xingtai, Hebei Province, lecturer of Department of Architectural Engineering, Hebei Vocational University of Technology and Engineering, focuses on teacher education. Email: Lining198107@163.com. https://orcid.org/0000-0002-7311-221X

Article submitted 2022-11-26. Resubmitted 2023-01-20. Final acceptance 2023-01-21. Final version published as submitted by the authors.