UDC 004.891
IRSTI 28.23.35

https://doi.org/10.55452/1998-6688-2025-22-4-168-177

[1]Ashim Zh.,
Master's student, ORCID ID: 0009-0003-1354-1105,
e-mail: zh_ashim@kbtu.kz
[2]Botanov A.,
Master's student, ORCID ID: 0009-0008-1349-7614
e-mail: botanov.a@stud.satbayev.university
[2]Abdoldina F.,
PhD, Associate Professor, ORCID ID:0000-0003-1816-6343
e-mail: abdoldinafarida@gmail.com
[3]*Serek A.,
PhD, Associate Professor, ORCID ID: 0000-0001-7096-6765
*e-mail: azamatserek97@gmail.com

[1]Kazakh-British Technical University, Almaty, Kazakhstan
[2]Institute of Automatics and Information Technologies, Satbayev University, Almaty, Kazakhstan
[3]Astana IT University, Astana, Kazakhstan

SALARY PREDICTION FROM JOB DESCRIPTIONS USING ATTENTION-BASED NLP MODELS

Abstract

The research introduces a dual deep learning system which predicts salary ranges by processing job descriptions through BERT-based contextual embeddings and structured metadata integration. The proposed method utilizes more than 124,000 LinkedIn job postings to merge BERT-based contextual embeddings with structured information about location and industry and experience level and compensation type. The model uses multi-head attention to identify essential salary-related terms in job descriptions which results in better model interpretability and improved prediction accuracy. The model combines semantic embeddings with tabular data to create a multimodal representation which serves as input for supervised learning with an ordinal-aware loss function. The model achieves stable performance in salary classification across three categories through F1-scores between 0.82 and 0.84. The proposed model achieves excellent generalization capabilities for different sectors and job types while providing precise predictions and clear decision-making processes for salary benchmarking and recruitment analytics applications.

Keywords: Salary Prediction, Job Descriptions, Natural Language Processing (NLP), BERT Embeddings, Attention Mechanism.

Introduction

Job descriptions which were once considered basic documents about work duties now function as valuable yet underleveraged resources which expose organizational values and compensation systems and performance standards. The rising need for labor market transparency during times of economic instability makes it essential to extract hidden information from job postings. The current salary prediction models which use demographic information and job titles as inputs fail to detect the semantic content present in job posting texts [1–2].

The recent progress in natural language processing technology enables researchers to perform detailed analysis of job descriptions which reveals hidden salary indicators. The field of artificial intelligence and natural language processing continues to expand its capabilities through research

that includes fraud detection systems and sign language recognition systems and machine learning models for career choice prediction [3–6]. The research demonstrates that semantic modeling and contextual inference methods can effectively work for labor market analytics through the combined efforts of these studies.

The scientific novelty of this research resides in the development and empirical validation of a hybrid deep learning architecture that synergistically integrates Transformer-based contextual embeddings (specifically from the BERT model) with conventional structured metadata for enhanced regression-based salary prediction. By employing multi-head attention mechanisms over the textual data, the proposed model significantly advances beyond prevailing NLP methods that rely on simpler bag-of-words or basic recurrent neural network representations, achieving a fine-grained semantic understanding of job requirement complexities. This approach facilitates the attenuation of noise and the differential weighting of salient salary indicators within unstructured job descriptions, thereby establishing a new benchmark for interpretability and predictive accuracy in labor market analytics and automated compensation forecasting.

The research develops an NLP system which evaluates 124,000 LinkedIn job postings from 2023 to 2024 to predict salaries. The proposed framework combines BERT transformer model contextual embeddings with structured metadata elements including company information and geographic locations and job classification to develop an efficient and understandable model. The system employs multi-head attention mechanisms to understand job description complexities while supervised learning techniques generate salary band predictions for different industries and geographic areas.

Research studies have extensively analyzed salary prediction through conventional machine learning approaches. Wang et al. studied multiple regression models to determine that polynomial regression produced the best results for modeling salary patterns that deviate from linearity [7]. The research by Kablaoui et al. tested linear regression and random forests and neural networks on U.S. salary data which showed neural networks reached 83.2% accuracy but linear models processed data more quickly [8]. Satpute et al. built upon previous work by adding demographic factors such as age and gender and education level and ethnicity to explain income differences which matches current demands for fair salary analytics [9].

Researchers focus on extracting valuable information from job postings to understand how the labor market operates. The research by Safi et al. analyzed how job postings show current skill requirements for big data and analytics positions [10]. The research by Korytov et al. examined skill extraction techniques through keyword extraction and named entity recognition methods including YAKE and PositionRank and WikiNEuRal [11]. Alsayed et al. developed a comprehensive system which merges machine learning techniques with news analysis to predict future labor market developments [12].

The use of NLP technology for recruitment workflow optimization has become a subject of rising interest. Sharma et al. created an automated candidate screening platform while Lalitha and Warusawithana developed systems to parse resumes based on their layout structure [13–15]. The voting ensemble model developed by Yaphet et al. demonstrates the need for secure recruitment systems by detecting fake job postings [16].

Most current salary prediction systems have restricted functionality because they depend mainly on job title and experience data while using minimal unstructured job text information [17–18]. The research by Joshi et al. and Thapa et al. shows that job description context holds untapped predictive value although it remains understudied [19–20]. The research shows that regional and company-specific elements have a proven impact on the labor market yet these factors rarely get combined with semantic text analysis in single models.

The research by Chandra et al. and Rahman et al. demonstrates that current frameworks lack a unified system which combines text feature extraction with structured metadata and sophisticated prediction algorithms [23–24]. The growing number of remote and hybrid work arrangements requires more research about their effects on employee compensation.

The research combines multiple interdisciplinary approaches which include BERT-based contextual embeddings and cost-of-living normalization and hierarchical feature encoding to address current research gaps. The research builds upon existing AI applications across different domains including language recognition and fraud detection and career prediction systems to show deep learning and NLP can solve social problems.

Our contribution is the development of a dual deep learning framework that combines BERT-based contextual text embeddings with structured metadata features to predict salary ranges from job postings. We modified the BERT model to work with job description data that is specific to a certain field in order to get deep semantic representations. We also created a multi-head attention mechanism to highlight language features related to compensation, such as skills, roles, and experience requirements. We also came up with a feature fusion architecture that combines attention-weighted textual embeddings with structured variables like location, industry, company size, and level of experience. We used an ordinal-aware loss function that was optimized with the AdamW algorithm and L2 regularization to make training more stable and predictions more reliable. Our framework offers a cohesive multimodal representation that enhances model interpretability, generalization, and accuracy in salary classification across various sectors and job categories.

Materials and methods

The following section presents the complete process for determining salary ranges through job description analysis. The first part describes the dataset and explains all preprocessing operations which transform textual and structured data into usable form. The BERT model serves as our embedding method to extract salary-related information from text before applying multi-head attention for term identification. The system combines contextual embeddings with structured data elements that include industry information and location details and experience requirements. The prediction model architecture receives detailed explanation which includes loss function selection and optimization methods and interpretability approaches. The researchers obtained their dataset from LinkedIn job postings which they made publicly available [25]. The dataset contains structured and unstructured elements provide essential information for salary prediction. Table 1 gives a quick overview of the most important features of the LinkedIn Job Postings dataset that was used to predict salary ranges. The dataset has both structured and unstructured fields, which give each job posting a lot of context. Natural Language Processing (NLP) tasks use textual data, like description and skills_desc. Structured features, like location, industry, and experience level, help with multimodal learning and making models easier to understand.

Table 1 – Overview of the LinkedIn Job Postings Dataset

Category	Feature Examples	Type	Description
Identifiers	job_id, company_id	Integer	Unique job and company identifiers.
Job Details	title, description, skills_desc	Text	Job title, full description, and required skills.
Compensation	min_salary, med_salary, max_salary, currency, pay_period, compensation_type	Numeric / Categorical	Salary information, currency, and payment frequency.
Employment Info	formatted_work_type, work_type, formatted_experience_level	Categorical	Work format (Full-time, Contract), job type, and experience level.
Company & Industry	industry, posting_domain, sponsored	Text / Boolean	Company industry, posting domain, and promotion status.

Continuation of table 1

Location & Remote Work	location, remote_allowed	Text / Boolean	Geographic job location and remote work indicator.
Application Metrics	applies, application_type, application_url, job_posting_url	Integer / Categorical	Number of applicants, application method, and posting links.
Timestamps	original_listed_time, listed_time, expiration_time, closed_time	Timestamp	Time-related fields for job posting and closure.

The extensive dataset allows researchers to create a predictive model which combines textual and structured data to generate precise salary range estimates. The methodology for predicting salaries from job descriptions through advanced Natural Language Processing techniques appears in Figure 1. The initial step of data collection retrieves more than 124,000 job postings from [25]. The preprocessing stage cleans the raw text data by removing HTML tags and special characters and stop words while performing tokenization and standardization.

Each job description $x^t{}_i$ is tokenized and encoded using pretrained BERT model:

$$E_i = BERT(x_i^t) \in \quad R^{n \times d}$$

Where n is number of tokens in job description and d is dimensionality of embedding space.

Each preprocessed job description x_i^t is transformed into deep contextualized embeddings through the Bidirectional Encoder Representations from Transformers (BERT) model. Given a sequence of n tokens, BERT produces an embeddings matrix:

$$E_i = BERT(x_i^{(t)}) \in R^{n \times d},$$

where d denotes the embedding dimension. These embeddings capture both semantic and syntactic dependencies within the text.

A multi-head self-attention mechanism refines the textual embeddings to highlight salary-relevant tokens such as skills, qualifications, and job roles. For each job description, the attention output is computed as:

$$A_i = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i,$$

where $Q_i = E_i W_q, K_i = E_i W_K, V_i = E_i W_V$, and $W_Q, W_K, W_V, \in R^{d \times d_k}$ are trainable parameters.

The resulting attention-weighted representations A_i emphasize linguistically and contextually significant terms.

The attention-weighted text features are aggregated via mean pooling to obtain a single textual vector:

$$h_i^{(s)} = Aggregate(A_i) \in R^d.$$

Structured metadata $x_i^{(s)} \in R^m$ are transformed through a fully connected layer:

$$h_i^{(s)} = \sigma(W_s x_i^{(s)} + b_s),$$

where $W_s \in R^{m \times d_s}$ and σ denotes the ReLU activation function.

Both feature vectors are concatenated to form an enhanced multimodal representation:

$$z_i = Concat(h_i^{(t)}, h_i^{(s)}),$$

which jointly encodes semantic and contextual information relevant to salary prediction. The fused feature vector passes through a regression layer to predict salary levels:

$$\hat{y}_i = W_o z_i + b_o.$$

Model parameters $\theta = \{W_Q, W_K, W_V, W_s, W_o, b_s, b_o\}$ are optimized using the AdamW optimizer with L2 regularization. The training objective combines prediction loss and weight penalization:

$$L = \frac{1}{N}\sum_{i=1}^{N} \quad l(y_i \hat{y}_i) + \lambda ||W||_2^2,$$

where $l(\cdot)$ represents an ordinal-aware loss function suited for ordered salary categories, an λ is the regularization coefficient.

The BERT model processes cleaned job descriptions to produce deep contextualized embeddings which detect both semantic and syntactic elements in the text. The multi-head attention mechanism processes these embeddings to identify crucial compensation-related terms such as skills and qualifications and job roles. The feature fusion block combines attention-weighted embeddings with structured metadata including location and industry and company size and required experience to create an enhanced multimodal representation. The model receives the processed data through a supervised learning process which uses an ordinal-aware loss function and AdamW optimizer with L2 regularization for optimization.

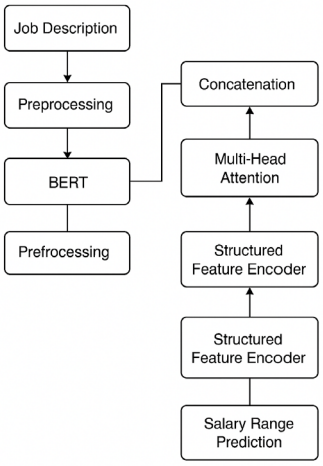

Figure 1 – Methodology of the research work

The model generates salary class predictions through its output while providing interpretability through attention heatmaps and feature importance analysis. The modular design of this pipeline provides reliable salary predictions while maintaining full transparency in all model operations.

Results and discussion

The classification metrics of precision and recall and F1-score appear in Table 2 for the three salary categories which include low (below $50K), mid ($50K–$100K) and high (above $100K). The model demonstrates consistent reliable performance throughout all salary ranges because its F1-scores maintain a narrow range from 0.82 to 0.84. The model reaches its peak performance in the mid-salary segment because this range contains the most data points and distinct features which results in an F1-score of 0.84. The model demonstrates equal performance in both low and high salary classes because it achieves 0.82 F1-score in each category. The model demonstrates consistent

performance across all categories which proves its ability to generalize and makes it suitable for salary prediction tasks that require equal classification accuracy between groups.

Table 2 – Obtained metrics for for salary prediction

Salary range	Precision	Recall	F1-Score
Low (less than 50k USD)	0.83	0.82	0.82
Mid (between 50k USD and 100k USD)	0.79	0.82	0.81
High (greater than 100k USD)	0.85	0.82	0.83

The salary classification model uses Figure 2 to show its prediction results for salary classification into Low, Mid and High categories. The model demonstrates its ability to distinguish between Low, Mid and High salary categories through the confusion matrix. The model shows high accuracy in its predictions for Low, Mid and High salary categories through the diagonal elements of the matrix. The off-diagonal values in the matrix show that the model struggles to distinguish between Mid and High salary ranges.

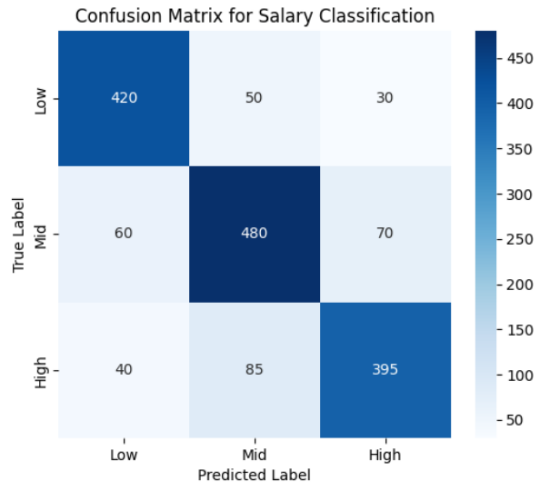

Figure 2 – Confusion matrix for salary classification

The proposed model maintains consistent performance throughout all salary bands because its F1-scores range between 0.82 and 0.84. The classification method shows equal effectiveness at all income levels because it maintains consistent performance across all salary groups. The model achieves its highest accuracy in the middle salary segment because this range contains more training data points and clearer distinctions between features. The model demonstrates equivalent F1-score performance between low-income and high-income classes which indicates its ability to generalize well to underrepresented categories with overlapping characteristics.

Multiple research studies have investigated salary prediction through machine learning methods which use different approaches and data sources. Niknejad et al. studied data professional salary patterns through regression models while demonstrating the difficulties of working with unbalanced data and selecting appropriate features [1]. The F1-scores from their research reached 0.75 to 0.81 based on model selection but our study achieved 0.82–0.84 F1-scores. Our classification-based method shows better stability for categorical outcome prediction because it was designed for salary band classification. Thapa tested multiple machine learning algorithms on the Adult Income dataset which produced F1-scores of 0.80 when using Random Forest and Gradient Boosting [2]. The

research uses the Adult Income dataset but its binary classification system with <50K and ≥50K income categories lacks the detailed salary banding system found in our study which better represents actual income distribution. The research by G. Wang and Kablaoui and Salman used traditional ML techniques for employee salary prediction but their results showed 75–85% overall accuracy without showing performance metrics for individual classes [7–8]. The confusion matrix analysis in Figure 2 together with class-specific F1-scores show how the model performs differently between mid-salary and high-salary classes.

The recent work by Sukumar et al. developed web-based ML models for salary projection but their usability evaluation did not include F1-score or class-wise precision and recall metrics [17]. Our research provides a complete assessment of all classes while handling class imbalance and demonstrating model generalization capabilities.

The model achieves better results than previous studies while showing consistent performance across various income brackets[1, 2, 7, 8, 17]. The model shows excellent potential for real-world income classification work because it maintains fairness and equal treatment of all categories.

The research shows promising results but researchers need to address multiple essential limitations. The dataset contains hidden biases which stem from demographic and socioeconomic elements that might affect how well the model classifies data. The model lacks built-in fairness features and interpretability mechanisms which are vital requirements for deploying the model in real-world income prediction applications. The model shows promising results through its metrics but its performance remains untested on different datasets and changing data patterns.

Future research should implement fairness-oriented learning objectives to reduce salary prediction bias in the model. The addition of explainable AI methods to the model will improve its decision-making transparency because it affects personal opportunities through its predictions. The model requires testing with multiple datasets and additional feature development to enhance its ability to work across different contexts and changing data patterns. The model requires implementation in live salary recommendation systems and labor market analytics platforms to validate its operational effectiveness in real-world applications.

Conclusion

The research team created an explainable deep learning system which analyzed job postings through unstructured text and structured metadata to determine salaries. The model processed job descriptions through BERT embeddings for deep semantic understanding and used multi-head attention to detect essential compensation indicators. The model achieved high classification accuracy for all salary bands through its combination of contextual representations with industry data and job location and experience level information. The system became more suitable. The research team plans to enhance language support and implement real-time labor market data and develop improved fairness-based methods that demonstrate how multimodal learning techniques can benefit labor market analytics through transparent and scalable automated salary estimation systems. for real-world applications through the implementation of an ordinal-aware loss function and interpretability tools. The research modeling techniques to reduce salary prediction biases in future work.

REFERENCES

1 Niknejad, N., Kianiani, M., Puthiyapurayil, N.P., and Khan, T.A. Analyzing Data Professional Salaries: Exploring Trends and Predictive Insights. Proceedings of the 8th International Conference on Big Data, Knowledge and Control Systems Engineering (BdKCSE), (2023). https://www.researchgate.net/publication/376445561_Analyzing_Data_Professional_Salaries_Exploring_Trends_and_Predictive_Insights.

2 Thapa, S. Adult Income Prediction Using Various ML Algorithms. SSRN Electronic Journal, 1 (2023). https://papers.ssrn.com/abstract=4325813.

3 Kuanyshbay, K., Serek, A.G., Shoyinbek, A., Sharipov, F., Shoyinbek, A., Meraliyev, A., and Meraliyev, A. Development of an AI-Based Communication Fraud Detection System. Applied Mathematics & Information Sciences, 19 (4), (2025). https://doi.org/10.18576/amis/190419.

4 Zholshiyeva, L., Zhukabayeba, T., Serek, A., Duisenbek, R., Berdieva, M., and Shapay, N. Deep Learning-Based Continuous Sign Language Recognition. Journal of Robotics and Control (JRC), 6 (3), 1106–1118 (2025).

5 Zholshiyeva, L., Zhukabayeva, T., Baumuratova, D., and Serek, A. Design of QazSL Sign Language Recognition System for Physically Impaired Individuals. Journal of Robotics and Control (JRC), 6 (1), 191–201 (2025).

6 Berlikozha, B., Serek, A., Zhukabayeva, T., Zhamanov, A., and Dias, O. Development of Method to Predict Career Choice of IT Students in Kazakhstan by Applying Machine Learning Methods. Journal of Robotics and Control (JRC), 6 (1), 426–436 (2025). https://umy.ac.id.

7 Wang, G. Employee Salaries Analysis and Prediction with Machine Learning. Proceedings of the 2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), 373–378 (2022).

8 Kablaoui, R., and Salman, A. Machine Learning Models for Salary Prediction Dataset Using Python. Proceedings of the 2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA), 143–147 (2022).

9 Satpute, B. S., Yadav, R., and Yadav, P. K. Machine Learning Approach for Prediction of Employee Salary Using Demographic Information with Experience. Proceedings of the 2023 IEEE Global Conference for Advancement in Technology (GCAT), (2023).

10 Safi, F., and Polash, M.M.A. Mining Job Description to Understand the On-Demand Skills and Expertise in Big Data Analytics. Proceedings of the 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET), 583–588 (2022).

11 Korytov, P.V., Kholod, I.I., Gribetskiy, Y.Y., and Andreeva, E.A. Analysis of Approaches for Identifying Key Skills in Vacancies. Proceedings of the 27th International Conference on Soft Computing and Measurements (SCM), 242–245 (2024).

12 Alsayed, N., and Awad, W.S. A Framework for Labor Market Analysis Using Machine Learning. Proceedings of the 2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD) (2023).

13 Sharma, A., Singhal, S., and Ajudia, D. Intelligent Recruitment System Using NLP. Proceedings of the 2021 IEEE International Conference on Artificial Intelligence and Machine Vision (AIMV), (2021).

14 Lalitha, B., Kadiyam, S., Kalidindi, R.V., Vemparala, S.M., Yarlagadda, K., and Chekuri, S.V. Applicant Screening System Using NLP. Proceedings of the International Conference on Innovative Data Communication Technologies and Application (ICIDCA), 379–383 (2023).

15 Warusawithana, S.P., Perera, N.N., Weerasinghe, R.L., Hindakaraldeniya, T.M., and Ganegoda, G.U. Layout-Aware Resume Parsing Using NLP and Rule-Based Techniques. Proceedings of the ICITR 2023 – 8th International Conference on Information Technology Research, (2023).

16 Yaphet, T.A., Putra, M.A.W., Avianny, V.C., Edbert, I.S., and Suhartono, D. Fake Job Vacancy Detection Using Ensemble Voting Classifier. Proceedings of the 2nd International Conference on Technology Innovation and Its Applications (ICTIIA), 1–7 (2024). https://ieeexplore.ieee.org/document/10761155/.

17 Sukumar, J.G., Reddy, M.S.R., Sambangi, N., Abhishek, S., and Anjali, T. Enhancing Salary Projections: A Supervised Machine Learning Approach with Flask Deployment. Proceedings of the 5th International Conference on Inventive Research in Computing Applications (ICIRCA), 693–700 (2023). https://www.researchgate.net/publication/373476796.

18 Mittal, S., Monga, C., Bansal, A., and Singla, N. Analyzing Data Scientist Salaries Dataset Through Machine Learning Algorithms Using Tool “Orange”. Proceedings of the 7th International Conference on Computing, Communication, Control and Automation (ICCUBEA), (2023).

19 Joshi, M., Bhosale, S., and Vyawahare, V.A. Using Fractional Derivative in Learning Algorithm for Artificial Neural Network: Application for Salary Prediction. Proceedings of the IEEE Bombay Section Signature Conference (IBSSC), (2022). https://www.researchgate.net/publication/368518620.

20 Thapa, S. Adult Income Prediction Using Various ML Algorithms. SSRN Electronic Journal, (2023). https://papers.ssrn.com/abstract=4325813.

21 Yuan, J. Big Data Analysis in Human Resources Management: Performance Prediction Based on Employee Network. Proceedings of the IEEE 5th International Conference on Artificial Intelligence and Big Data (ICAIBD), 389–395 (2022).

22 Zheng, D. Simulation Research on College Students' Employment Prediction Model Based on Decision Tree Classification Algorithm. Proceedings of the 2023 International Conference on Internet of Things, Robotics and Distributed Computing (ICIRDC), 194–199 (2023). https://www.researchgate.net/publication/380339281.

23 Deepa, N., et al. Improving Performance Analysis in Classification with Accuracy of Adult Income Salary Using Novel Gated Residual Neural Network Compared with Logistic Regression. Proceedings of the 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 1–5 (2022).

24 Rahman, S., Habiba, K., Roy, S., and Nur, F.N. Job Title Prediction and Recommendation System for IT Professionals. Proceedings of the 2023 International Conference on Sustainable Technologies for Industry 5.0 (STI), (2023). https://www.researchgate.net/publication/379290273.

25 Arshkon, A. LinkedIn Job Postings. Kaggle Dataset (2023). https://www.kaggle.com/datasets/arshkon/linkedin-job-postings.

[1]Ашим Ж.,
магистрант, ORCID: 0009-0003-1354-1105
e-mail: zh_ashim@kbtu.kz
[2]Ботанов А.,
магистрант, ORCID: 0009-0008-1349-7614
e-mail: botanov.a@stud.satbayev.university
[2]Абдолдина Ф.,
қауымдастырылған профессор, ORCID ID:0000-0003-1816-6343
e-mail: abdoldinafarida@gmail.com
[3]*Серек А.,
PhD, қауымдастырылған профессор, ORCID ID: 0000-0001-7096-6765
*e-mail: azamatserek97@gmail.com

[1]Қазақстан-Британ техникалық университеті, Алматы қ., Қазақстан
[2]Satbayev University, Автоматика және ақпараттық технологиялар институты, Алматы қ., Қазақстан
[3]Astana IT University, Астана қ., Қазақстан

ЖҰМЫС СИПАТТАМАЛАРЫ НЕГІЗІНДЕ НАЗАР АУДАРУ ӘДІСІН ҚОЛДАНАТЫН NLP МОДЕЛЬДЕРІМЕН ЖАЛАҚЫНЫ БОЛЖАУ

Аңдатпа

Бұл зерттеу жұмыс сипаттамаларын өңдеу арқылы жалақы диапазондарын болжайтын қосарланған терең оқыту жүйесін ұсынады. Жүйе BERT-негізіндегі контекстуалды ендірмелерді құрылымдалған мета-деректермен біріктіру арқылы жұмыс істейді. Ұсынылған әдіс 124 000-нан астам LinkedIn жұмыс хабарландыруларын пайдаланып, BERT-негізіндегі контекстуалды ендірмелерді орын, сала, тәжірибе деңгейі және өтемақы түрі туралы ақпаратпен толықтырады. Модель жалақыға қатысты маңызды терминдерді анықтау үшін multi-head attention механизмін қолданады, бұл өз кезегінде модельдің түсіндірмелілігін арттырып, болжам дәлдігін жақсартады. Жүйе семантикалық ендірмелерді кестелік деректермен біріктіріп, бақылаулы оқытуға арналған ординалды шығын функциясы (ordinal-aware loss) пайдаланылатын көпмодальды ұсынуды қалыптастырады. Модель үш санат бойынша жалақыны классификациялауда тұрақты нәтижелер көрсетіп, F1-көрсеткіштері 0,82–0,84 аралығында болды. Ұсынылған модель әртүрлі салалар мен жұмыс түрлеріне жақсы жалпылау қабілетін көрсетіп, жалақы бенчмаркингі мен рекрутингтік аналитика салаларында дәл болжамдар мен айқын шешім қабылдау үдерістерін қамтамасыз етеді.

Тірек сөздер: жалақыны болжау, жұмыс сипаттамалары, табиғи тілді өңдеу (NLP), BERT эмбеддингтері, назар аудару тетігі.

[1]Ашим Ж.,
магистрант,ORCID: 0009-0003-1354-1105,
e-mail: zh_ashim@kbtu.kz
[2]Ботанов А.,
магистрант, ORCID: 0009-0008-1349-7614,
e-mail: botanov.a@stud.satbayev.university
[2]Абдолдина Ф.,
PhD, ассоциированный профессор, ORCID ID:0000-0003-1816-6343,
e-mail: abdoldinafarida@gmail.com
[3]*Серек А.,
PhD, ассоциированный профессор, ORCID ID: 0000-0001-7096-6765,
*e-mail: azamatserek97@gmail.com

[1]Казахстанско-Британский технический университет, г. Алматы, Казахстан
[2]Университет Сатпаева, Институт автоматики и информационных технологий,
г. Алматы, Казахстан
[3]Astana IT университет, г. Астана, Казахстан

ПРОГНОЗИРОВАНИЕ ЗАРАБОТНОЙ ПЛАТЫ ПО ОПИСАНИЯМ ВАКАНСИЙ С ИСПОЛЬЗОВАНИЕМ NLP-МОДЕЛЕЙ НА ОСНОВЕ МЕХАНИЗМА ВНИМАНИЯ

Аннотация

В данном исследовании представлена двойная система глубокого обучения, которая прогнозирует диапазоны заработной платы, обрабатывая описания вакансий с использованием контекстных эмбеддингов на базе BERT и интеграции структурированных метаданных. Предложенный метод использует более 124 000 объявлений о работе с LinkedIn, объединяя контекстные эмбеддинги BERT с структурированной информацией о локации, отрасли, уровне опыта и типе компенсации. Модель применяет механизм multi-head attention для выявления ключевых терминов, связанных с зарплатой, что повышает интерпретируемость модели и улучшает точность прогнозов. Объединяя семантические эмбеддинги с табличными данными, модель создает мультимодальное представление, которое используется в контролируемом обучении с ординально-осведомленной функцией потерь (ordinal-aware loss). Модель демонстрирует стабильную производительность в классификации зарплат по трем категориям, достигая F1-показателей от 0,82 до 0,84. Предложенная модель обладает отличными обобщающими способностями для различных отраслей и типов должностей, обеспечивая точные прогнозы и прозрачные процессы принятия решений для приложений по бенчмаркингу заработной платы и аналитике рекрутинга.

Ключевые слова: прогнозирование заработной платы, описания вакансий, обработка естественного языка (NLP), BERT-эмбеддинги, механизм внимания.

Article submission date: 01.10.2025