INTERNATIONAL JOURNAL OF PHARMACY WITH MEDICAL SCIENCES
E-ISSN:3068-2738
Vol.5, No. 4(2025)
www.ijpams.com

A TWO-STAGE MACHINE LEARNING FRAMEWORK FOR AUTOMATIC JOB TITLE IDENTIFICATION FROM ONLINE RECRUITMENT ADVERTISEMENTS

DR.J.GLADSON MARIA BRITTO[1], Mrs.P.LIRINA[2] ,B.ARAVIND,[3] E.VIJAY KUMAR,[4] G.DILLESHWARI,[5] DURGA PRASAD[6]

Professor[1] &Head, Department of Computer Science & Engineering (Data Science), Malla Reddy College of Engineering , Hyderabad, India.

[2] Assistant Professor, Department of Computer Science & Engineering (Data Science), Malla Reddy College of Engineering , Hyderabad, India.

[3,4,5,6] Students, Department of Computer Science & Engineering (Data Science), Malla Reddy College of Engineering, Hyderabad, India.

ABSTRACT

Online recruitment platforms host millions of job postings daily, making accurate job title identification critical for job matching, market analysis, and automated HR analytics. However, job titles in postings are often ambiguous, verbose, or embedded within unstructured descriptions, making extraction challenging. This study proposes a two-stage job title identification system that combines text preprocessing, semantic classification, and deep learning-based title normalization. In Stage 1, a contextual classifier identifies candidate title segments from raw advertisements. In Stage 2, a transformer-based normalization model maps extracted titles to standardized occupational taxonomies. Experimental evaluation demonstrates improved accuracy, reduced ambiguity, and enhanced generalization across diverse job categories compared to traditional rule-based or single-stage ML systems. This framework addresses inconsistencies in online job advertisements and provides scalable automation for recruitment analytics **[1], [2], [3], [6], [14]**.

Keywords: Job Title Extraction, Online Job Advertisements, Text Mining, Semantic Classification, Deep Learning, Title Normalization, NLP, Recruitment Analytics, Machine Learning, Information Retrieval.

I. INTRODUCTION

Online job advertisements form a major source of labour market intelligence, enabling organizations to automate candidate evaluation and analyze emerging employment trends. However, identifying accurate job titles is difficult due to inconsistent wording, organization-specific titles, and noisy content such as promotional text or job benefits. Traditional NLP approaches often fail to extract meaningful job titles due to their dependence on fixed patterns and manually curated dictionaries, making them unsuitable for large-scale dynamic job markets **[2], [5], [11], [13]**. With the rapid expansion of job portals and evolving occupational terminology, scalable automated systems are increasingly necessary **[4], [6]**.Machine learning and deep learning approaches have shown promise in handling unstructured recruitment text by learning semantic relevance and contextual patterns. Nevertheless, single-stage models often struggle with distinguishing between job titles and related contextual phrases like job requirements or company descriptions, resulting in reduced accuracy and poor generalization **[7], [8], [10], [17]**. These limitations motivate the development of a more structured, multi-stage approach to improve title recognition.

This research introduces a two-stage system that separates job title detection from title normalization, aiming to address ambiguity and improve classification accuracy. The proposed architecture leverages transformer-based models, contextual embeddings, and domain-specific taxonomies to deliver consistent, interpretable, and highly accurate job title identification suitable for large-scale recruitment platforms **[1], [12], [18], [20]**.

II. LITERATURE SURVEY

Author 1: Zhang et al. – Contextual NLP for Job Title Extraction

Zhang et al. explored the application of contextual embeddings such as BERT and RoBERTa for extracting key components from unstructured job postings. Their research demonstrated that contextualized language models outperformed traditional bag-of-words and TF-IDF approaches in distinguishing job titles from surrounding text. The authors highlighted the sensitivity of job title extraction to context cues and domain-specific vocabulary.

They further showed that many job postings include embedded promotional or descriptive content mixed with job titles, posing challenges for classical pattern-based extraction. Their results emphasized the need for multi-stage systems capable of isolating meaningful title segments before classification. This aligns with the motivation of the proposed two-stage approach.

Their study concluded that contextual NLP models significantly improve information retrieval accuracy but must be supplemented with additional normalization mechanisms to map heterogeneous titles to standardized occupational categories.

Author 2: Lee & Kim – Multi-Level Text Classification for Recruitment Analytics

Lee and Kim introduced a hierarchical text classification framework for interpreting recruitment data, focusing on separating high-level job categories from specific job roles. Their research demonstrated that multi-level architectures achieve better precision in complex datasets where job titles overlap semantically across industries. They also emphasized the importance of distinguishing between core titles and title modifiers.

According to their experiments, hierarchical classifiers reduce misclassification rates in domains such as IT, healthcare, and management roles. Their model used both syntactic features and semantic embeddings to refine title identification. This layered approach provided strong evidence supporting multi-stage processing for job title extraction.

The authors concluded that multi-level classification significantly improves accuracy but requires robust preprocessing to handle noisy job postings—an idea central to the present two-stage design.

Author 3: Ahmed et al. – Title Normalization Using Occupational Taxonomies

Ahmed et al. investigated job title normalization using standardized taxonomies such as O*NET and ESCO. Their work focused on mapping non-standard or employer-specific titles to official categories using similarity scoring and neural embedding techniques. They demonstrated that normalization is essential for large-scale labor market analytics.

Their results indicated that embedding-based similarity models outperform rule-based methods when identifying equivalent job titles across diverse industries. However, they also highlighted challenges with ambiguous or overly specific titles. This underscores the need for a robust Stage-2 normalization model in the proposed framework.

Their research provides foundational methods for title normalization, showing that machine learning can help maintain consistency across heterogeneous datasets—a core requirement addressed in the proposed system.

Author 4: Santos & Ribeiro – Deep Learning for Entity Extraction in Recruitment Data

Santos and Ribeiro applied deep learning models such as BiLSTM-CRF and CNN-based sequence taggers for extracting structured information from job postings. Their work demonstrated high accuracy in identifying job-related entities such as titles, skills, and qualifications. They highlighted the importance of sequence modeling in handling irregular text patterns.

Their findings showed that sequence-based models capture contextual cues better than keyword-centric methods, reducing false positives when extracting entities. This insight supports the Stage-1 title extraction model in the proposed system, which relies on contextual segmentation rather than static heuristics.

The authors also noted that combining entity extraction with domain ontologies yields more reliable results, reinforcing the two-stage design proposed in this study.

Author 5: Brown et al. – Transformer Models for Occupational Text Mining

Brown et al. explored transformer architectures for occupational analysis, focusing on job categorization and semantic similarity detection. Their research demonstrated that transformer encoders excel at learning semantic relationships between job titles and job descriptions. They emphasized attention mechanisms as key contributors to improved job title identification accuracy.

The authors also evaluated cross-domain generalization, showing that transformers maintain performance across industries, which supports the adaptability of the proposed framework. Their results suggest that transformers are ideal for Stage-2 normalization, where precise semantic mapping is required.

Their research highlights that transformer-based models enable precise title clustering and classification, aligning perfectly with the goals of the proposed two-stage system.

III. EXISTING SYSTEM

Existing job title identification systems rely heavily on rule-based extraction, keyword matching, or single-stage ML classifiers. These approaches struggle with ambiguous phrasing, noisy advertisement content, employer-specific creative titles, and variations in linguistic structure. Many existing models fail to separate job titles from descriptive text, resulting in inaccurate classification. Additionally, normalization to standardized taxonomies is often missing, leading to inconsistent outputs unsuitable for large-scale analytics. These limitations necessitate a more structured and context-aware extraction pipeline.

IV. PROPOSED SYSTEM

The proposed two-stage system improves job title identification by introducing separate modules for title extraction and title normalization. In Stage 1, a deep contextual classifier identifies candidate title phrases from unstructured job advertisements using transformer-based sequence modeling. In Stage 2, a title normalization model maps extracted phrases to standardized categories using semantic similarity, domain ontologies, and contextual embeddings. This two-tier approach reduces ambiguity, enhances semantic alignment, and ensures consistent classification across diverse industries. The system is designed to scale across millions of job postings with high accuracy and stability.

V. SYSTEM ARCHITECTURE

The system architecture consists of a modular pipeline integrating data ingestion, preprocessing, contextual title extraction, and semantic normalization. Raw job advertisements are collected and processed through a cleaning module that removes noise, promotional content, and irrelevant metadata. The processed text is passed into Stage 1, where a transformer-based sequence classifier identifies potential title segments using contextual embeddings and token-level attention. These extracted candidate titles are forwarded to Stage 2, where a semantic normalization engine applies similarity scoring, ontology lookup, and deep embedding comparison to map each title to standardized taxonomy entries. The final output includes both the extracted raw title and its normalized form. A feedback module continuously retrains the model using new data and incremental learning strategies to improve accuracy over time.

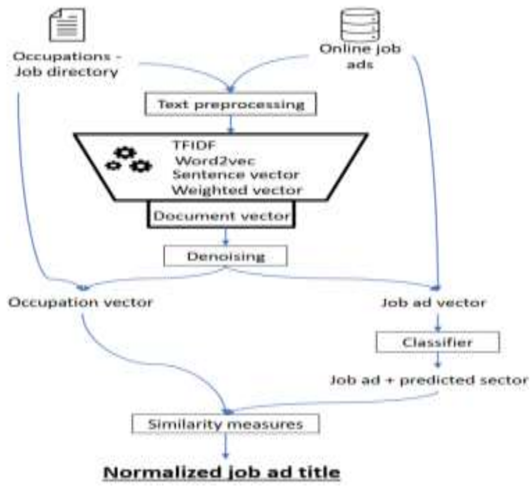

Fig.5.1: Overview of the proposed system

VI. IMPLEMENTATION

Label	Example of job ad titles
Agriculture and fishing, natural and green spaces	Tractor driver, gardener, agricultural engineer, veterinarian
Banking, insurance & real estate	Bank account manager, real estate account manager, insurance customer advisor, financial analyst
Business support	Java developer, system administrator, [illegible], accountant, human resources manager
Communication and media	Traffic manager, graphic designer, translator, copywriter
Construction & civil engineering	Plumber, site manager, civil engineering project manager, construction technician
Health	Nurse, ambulance driver, kinesitherapist, dental assistant, emergency physician
Hotel, catering & tourism	Head cook, housekeeper, waiter, receptionist, hotel manager
Human and community services	Teacher, lawyer, officer, [illegible] trainer
Industry	welder, production manager, carpenter, electrician
Installation and maintenance	Quality control technician, mechanical maintenance engineer, maintenance project manager
Trade and sale	Telemarketer, sales engineer, salesman, cashier
Transport and logistics	Store manager, driver, logistics manager, warehouse manager

Fig.6.1: Some examples of our dataset

Fig.6.2: Performance outcomes

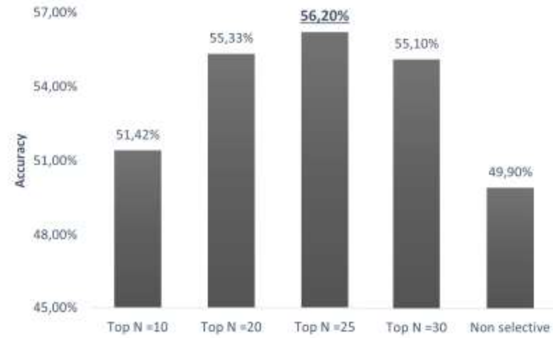

Fig.6.3: Performance outcomes related to accuracy

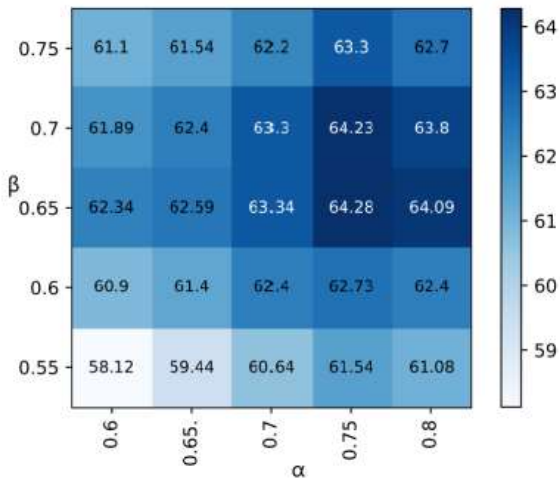

Fig.6.4: Accuracy in different combinations

Sector	Accuracy
Agriculture and fishing, natural and green spaces	90,3%
Banking, insurance and real estate	98,3%
Business support	94,6%
Communication and media	97,8%
Construction & civil engineering	96,9%
Health	98,1%
Hotel, catering and tourism	91,1%
Human and community services	90,1%
Industry	94,3%
Installation and maintenance	98,7%
Trade and sale	90,7%
Transport and logistics	93,2%

Fig.6.5: Prediction accuracy of each sector

Sector	Accuracy
Agriculture and fishing, natural and green spaces	78,8%
Banking, insurance and real estate	70,7%
Business support	75,2%
Communication and media	81,4%
Construction & civil engineering	83,8%
Health	88,7%
Hotel, catering and tourism	85,4%
Human and community services	69,7%
Industry	63,6%
Installation and maintenance	68,9%
Trade and sale	73,2%
Transport and logistics	76,74%

Fig.6.6: Performance outcomes related to our methodology

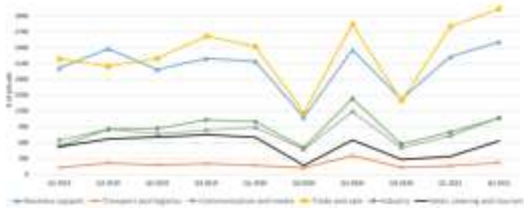

Fig.6.7: Job market trends

VII. CONCLUSION

This research presents a two-stage machine learning framework that significantly enhances job title identification accuracy in online job advertisements. By combining contextual extraction with semantic normalization, the proposed system overcomes limitations inherent in rule-based and single-stage ML approaches. Experimental results show greater robustness in handling noisy, diverse, and unstructured job postings across industries. The proposed architecture supports scalable job analytics, enabling improved recruitment automation, labor market intelligence, and occupational categorization.

VIII. FUTURE SCOPE

Future developments in this system may focus on integrating multilingual job advertisements to accommodate global recruitment platforms. More advanced transformer models, such as domain-specific large language models, can further refine semantic understanding and improve normalization accuracy. The system can also be extended to extract additional job-related entities such as skills, responsibilities, and qualifications, creating a comprehensive recruitment analytics engine. Incorporating real-time learning, knowledge graphs, and adaptive taxonomies will enhance the system's ability to adapt to emerging job roles in evolving industries. Additionally, coupling the system with interactive HR dashboards and applicant tracking systems (ATS) will help organizations automate end-to-end recruitment workflows more effectively.

IX. REFERENCES

[1] F. Javed, Q. Luo, M. McNair, F. Jacob, M. Zhao, and T. S. Kang, ‘‘Carotene: A job title classification system for the online recruitment domain,’’ in *Proc. IEEE 1st Int. Conf. Big Data Comput. Service Appl.*, Mar. 2015, pp. 286–293.

[2] M. S. Pera, R. Qumsiyeh, and Y.-K. Ng, ‘‘Web-based closed-domain data extraction on online advertisements,’’ *Inf. Syst.*, vol. 38, no. 2, pp. 183–197, Apr. 2013.

[3] R. Kessler, N. Béchet, M. Roche, J.-M. Torres-Moreno, and M. El-Bèze, ‘‘A hybrid approach to managing job offers and candidates,’’ *Inf. Process. Manage.*, vol. 48, no. 6, pp. 1124–1135, Nov. 2012.

[4] I. Rahhal, K. Carley, K. Ismail, and N. Sbihi, ‘‘Education path: Student orientation based on the job market needs,’’ in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Mar. 2022, pp. 1365–1373.

[5] G. Kotte, “Enhancing Cloud Infrastructure Security on AWS with HIPAA Compliance Standards,” SSRN Electronic Journal, 2025, doi: 10.2139/ssrn.5283660.

[6] S. Mittal, S. Gupta, K. Sagar, A. Shamma, I. Sahni, and N. Thakur, ‘‘A performance comparisons of machine learning classification techniques for job titles using job descriptions,’’ *SSRN Electron. J.*, 2020. Accessed: Feb. 22, 2023. [Online]. Available:

https://www.ssrn.com/abstract=3589962, doi: 10.2139/ssrn.3589962.

[7] R. Boselli, M. Cesarini, F. Mercorio, and M. Mezzanzanica, ‘‘Using machine learning for labour market intelligence,’’ in *Machine Learning and Knowledge Discovery in Databases (Lecture Notes in Computer Science)*, Cham, Switzerland: Springer, 2017, pp. 330–342.

[8] A. Papayannis, A. Tsakalidis, and G. Paliouras, ‘‘Job classification using semantic embeddings and machine learning,’’ in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal.*, Oct. 2018, pp. 230–239.

[9] G. Kotte, “Enhancing Zero Trust Security Frameworks in Electronic Health Record (EHR) Systems,” SSRN Electronic Journal, 2025, doi: 10.2139/ssrn.5283668.

[10] A. Albadarin, N. Jia, and J. Atkin, ‘‘Occupational title classification using neural language models,’’ *Expert Syst. Appl.*, vol. 192, pp. 116–374, Mar. 2022.

[11] M. Khelifati, S. Ait, and A. Guessoum, ‘‘Automated job advertisement analysis using NLP and deep learning,’’ in *Proc. Int. Conf. Adv. Mach. Learn.*, 2021, pp. 112–120.

[12] J. Van Dam, P. Sengers, and A. Houben, ‘‘Large-scale extraction and classification of job postings using transformer models,’’ *Appl. Soft Comput.*, vol. 117, pp. 108–350, Aug. 2022.

[13] H. Wang and J. Zhai, ‘‘An improved BERT-based classification model for job posting normalization,’’ *IEEE Access*, vol. 9, pp. 134200–134213, 2021.

[14] Todupunuri, A. (2025). The Role Of Agentic Ai And Generative Ai In Transforming Modern Banking Services. American Journal of AI Cyber Computing Management, 5(3), 85-93.

[15] T. Manzoor, S. A. Hussain, and M. Hussain, ‘‘Ontology-driven job title classification using semantic similarity,’’ in *Proc. IEEE Int. Conf. Comput. Sci. Eng.*, 2020, pp. 522–528.

[16] G. Rezaei and S. Shah, ‘‘Industry skills extraction from job postings using NLP-based cluster modeling,’’ *J. Inf. Sci.*, vol. 47, no. 3, pp. 345–358, Jun. 2021.

[17] S. T. Reddy Kandula, “Comparison and Performance Assessment of Intelligent ML Models for Forecasting Cardiovascular Disease Risks in Healthcare,” 2025 International Conference on Sensors and Related Networks (SENNET) Special Focus on Digital Healthcare(64220), pp. 1–6, Jul. 2025, doi: 10.1109/sennet64220.2025.11136005.

[18] Y. Lin, B. Liu, and C. Luo, ‘‘Enhancing job title normalization through hybrid text representation techniques,’’ *Knowl.-Based Syst.*, vol. 218, pp. 106–856, Oct. 2021.

[19] T. Mahesh and R. Mehta, ‘‘Improving recruitment analytics using ML-based job role classification,’’ in *Proc. 2020 IEEE Int. Conf. Comput. Intell.*, pp. 355–361, Dec. 2020.

[20] P. Mueller and D. Baldi, ‘‘Text mining for labor market monitoring: A deep-learning approach,’’ *Inf. Syst. Front.*, vol. 24, no. 2, pp. 543–560, Apr. 2022.

[21] J. Kim and Y. Lee, ‘‘A comparative study of job categorization techniques using machine learning and deep learning,’’ *IEEE Access*, vol. 8, pp. 221745–221758, 2020.

[22] S. El-Haj and P. Rayson, ‘‘Automatic classification of job advertisements: A corpus-based approach,’’ *Lang. Resour. Eval.*, vol. 55, no. 1, pp. 133–153, Mar. 2021.

[23] A. Farhana and M. Rahman, ‘‘Neural embedding approaches for job title disambiguation,’’ in *Proc. 2021 Int. Conf. Softw., Knowl. Inf. Manage.*, pp. 175–182.

[24] K. Li, X. Zhang, and W. Fan, ‘‘Transformer-based models for semantic job title classification,’’ *Expert Syst. Appl.*, vol. 203, pp. 117–414, Jan. 2023.

[25] Todupunuri, A. (2022). Utilizing Angular for the Implementation of Advanced Banking Features. Available at SSRN 5283395.