Two Stage Job Title Identification System for Online Job Advertisements

First Author: Mr. K. Ramesh, Assistant Professor, Dept of MCA, Audisankara College of Engineering & Technology, Gudur, Nellore

Second Author: J. Venkateswarlu, Pursuing MCA, Audisankara College of Engineering & Technology, Gudur , Nellore

ABSTRACT

Online job advertisements contain large volumes of unstructured textual data that make automatic job title identification a challenging task. Traditional classification approaches require large labeled datasets and often fail to capture semantic relationships between job descriptions and occupation titles. To address these limitations, this paper presents a Two Stage Job Title Identification System for Online Job Advertisements using machine learning and natural language processing techniques. The proposed system combines supervised and unsupervised learning methods to improve job title prediction accuracy with minimal labeled data. In the first stage, Bidirectional Encoder Representations from Transformers (BERT) is used to classify job advertisements into their corresponding sectors such as Information Technology, Agriculture, and Sales. In the second stage, document embedding and similarity matching techniques are applied to identify the most relevant occupation title from the predicted sector. The system utilizes feature extraction, document representation, and cosine similarity measures to improve semantic understanding of job descriptions. Experimental evaluation demonstrates that the proposed methodology significantly improves job title identification accuracy compared with traditional machine learning approaches such as Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression. The proposed framework is scalable, efficient, and adaptable to multilingual job market datasets. Furthermore, the system can support recruitment analytics, labor market analysis, and career guidance applications by identifying emerging occupations and high-demand job roles from online recruitment platforms.

I. INTRODUCTION

The rapid growth of online recruitment platforms and digital hiring systems has generated a massive volume of job advertisements containing valuable information related to occupations, required skills, and labor market trends. These job advertisements are generally available in unstructured textual formats, making automatic extraction and identification of job titles a challenging task. Natural Language Processing (NLP) and Machine Learning (ML) techniques have become essential tools for processing such textual information and transforming it into structured knowledge useful for recruitment analytics, career guidance, and labor market analysis [11], [14].

Traditional job title identification systems mainly rely on supervised machine learning approaches such as Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression for classifying job advertisements into predefined occupational categories. Although these methods provide acceptable results, they require large amounts of labeled training data and often fail to capture semantic relationships between words and job descriptions [13]. In addition, the heterogeneous nature of job advertisements, where different employers use different vocabularies for similar job roles, further complicates the classification process.

Recent advancements in deep learning and transformer-based language models have significantly improved text classification tasks. Bidirectional Encoder Representations from Transformers (BERT) introduced by Devlin et al. [1] has demonstrated outstanding performance in understanding contextual and semantic relationships within text data. BERT-based document representation techniques such as DocBERT [2] and JobBERT [3] have shown promising results in job title prediction and occupation classification tasks. Similarly, document embedding methods including Word2Vec [6], GloVe [7], and distributed sentence representations [9] have improved semantic understanding of textual data by representing words and documents in dense vector spaces.

Despite these advancements, existing job title identification systems still face several limitations. Most systems require extensive labeled datasets, exhibit poor generalization across domains, and often rely only on job titles while ignoring detailed job descriptions. Furthermore, some models fail to accurately identify occupations when job advertisements contain noisy or incomplete information. Therefore, there is a need for an efficient and scalable framework capable of identifying job titles using both job titles and descriptions with minimal labeled data.

To address these challenges, this paper proposes a Two Stage Job Title Identification System for Online Job Advertisements using supervised and unsupervised machine learning techniques. In the first stage, BERT-based classification is used to categorize job advertisements into corresponding sectors such as Information Technology, Agriculture, and Sales. In the second stage, document embedding and similarity matching techniques are used to identify the most relevant occupation title from the predicted sector. The proposed system combines semantic feature extraction, contextual document representation, and cosine similarity measures to improve prediction accuracy and reduce computational complexity.

The major contributions of this paper are summarized as follows:

1. A two-stage framework for job title identification using BERT and similarity matching techniques is proposed.
2. The system combines supervised and unsupervised learning methods to improve occupation prediction accuracy.
3. Multiple document embedding techniques are utilized to enhance semantic understanding of job advertisements.
4. The proposed framework reduces dependency on large labeled datasets and improves scalability for real-world recruitment systems.

The remainder of this paper is organized as follows. Section 2 presents the literature survey related to job title identification and document embedding techniques. Section 3 describes the proposed methodology and system architecture. Section 4 discusses the experimental setup and dataset preparation. Section 5 presents the results and performance evaluation. Finally, Section 6 concludes the paper and outlines future research directions

II. RELATED WORK

Automatic job title identification has attracted significant attention in recent years due to the increasing availability of online recruitment data. Several researchers have proposed machine learning and deep learning techniques to classify job advertisements and identify occupational categories.

Devlin et al. [1] introduced BERT, a transformer-based language representation model capable of understanding contextual relationships between words through bidirectional training. BERT significantly improved the performance of various NLP tasks including text classification, sentiment analysis, and question answering. The contextual embedding capability of BERT made it highly suitable for job advertisement analysis and occupation prediction.

Adhikari et al. [2] proposed DocBERT, an extension of BERT for document classification tasks. The model utilized transformer-based contextual embeddings to generate document-level representations and demonstrated superior performance compared with traditional machine learning classifiers. Similarly, Decorte et al. [3] introduced JobBERT for understanding job titles through skill extraction and semantic analysis. Their work highlighted the effectiveness of contextual embeddings in recruitment-related applications.

Tran et al. [4] developed a multi-label classification framework for predicting job titles from job descriptions using pre-trained language models. Their approach demonstrated that incorporating detailed job descriptions significantly improved prediction accuracy compared with using job titles alone. Safikhani et al. [5] further improved occupation coding by utilizing hierarchical features and pre-trained language models to classify occupations with higher precision.

Traditional document representation methods such as Word2Vec developed by Mikolov et al. [6] introduced distributed word representations capable of preserving semantic relationships between words. Similarly, Pennington et al. [7] proposed GloVe embeddings, which combined global word co-occurrence statistics with vector representations to improve semantic understanding. These embedding techniques became widely used in NLP applications involving text similarity and classification.

Joulin et al. [8] proposed an efficient text classification approach using shallow neural networks and bag-of-tricks methods for scalable document classification tasks. Their work demonstrated that simple architectures with optimized embeddings could achieve competitive performance with lower computational costs. Le and Mikolov [9] introduced distributed representations of sentences and documents, enabling semantic representation of long textual content such as job descriptions.

Transformer architectures proposed by Vaswani et al. [10] revolutionized NLP by introducing self-attention mechanisms capable of capturing long-range dependencies in textual data. The transformer architecture later became the foundation for advanced language models such as BERT and GPT-based systems.

Bird et al. [11] provided practical NLP techniques and preprocessing methods useful for text classification and information extraction tasks. Goldberg [12] discussed neural network methods for natural language processing and emphasized the role of deep learning in semantic representation learning. Aggarwal and Zhai [13] explored various text mining approaches including clustering, classification, and feature extraction techniques for large-scale textual datasets. Manning et al. [14] presented foundational concepts related to information retrieval, document indexing, and similarity measures widely used in document matching systems. Jurafsky and Martin [15] comprehensively discussed modern speech and language processing techniques, including transformer models, semantic embeddings, and machine learning methods for NLP applications.

Although existing research has demonstrated promising results in occupation classification and job title prediction, several limitations remain. Most studies require large labeled datasets and computationally expensive training procedures. In addition, many systems rely heavily on job titles without effectively utilizing detailed job descriptions. To overcome these limitations, the proposed work introduces a two-stage framework combining BERT-based sector classification with document similarity matching for efficient and accurate job title identification from online job advertisements.

III. PROPOSED METHODOLOGY

The proposed system introduces a Two Stage Job Title Identification System for Online Job Advertisements using Natural Language Processing (NLP), Machine Learning

(ML), and Deep Learning techniques. The primary objective of the system is to accurately identify the most relevant job title from unstructured online job advertisements by combining supervised and unsupervised learning approaches.

The proposed framework consists of two major stages. In the first stage, the system classifies job advertisements into their corresponding sectors using a BERT-based classification model. In the second stage, semantic similarity matching techniques are applied to identify the most suitable occupation title from the predicted sector. The overall architecture of the proposed system is shown in Figure 1.

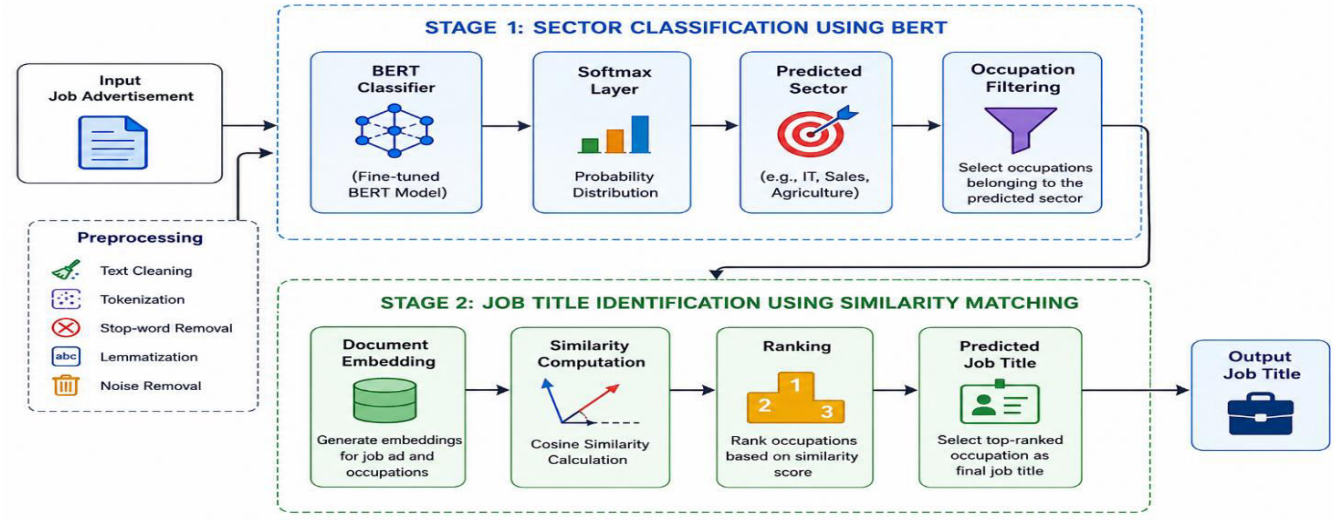

Figure 1: Architecture of the Proposed Two Stage Job Title Identification System for Online Job Advertisements

3.1 Data Collection

The dataset used in this work consists of online job advertisements collected from recruitment portals and employment websites. Each job advertisement generally contains information such as job title, job description, required skills, qualifications, and company details. Since the collected data is unstructured and noisy, preprocessing techniques are applied before performing classification and similarity analysis.

The collected dataset is divided into multiple sectors such as:

- Information Technology
- Agriculture
- Healthcare
- Sales and Marketing
- Finance
- Education

These sectors are used as the primary categories during the first-stage classification process.

3.2 Data Preprocessing

Data preprocessing plays a critical role in improving the quality of textual data and enhancing classification accuracy. The preprocessing stage includes the following operations:

1. Removal of special characters and symbols
2. Tokenization of text data
3. Stop-word removal
4. Lowercase conversion
5. Lemmatization and stemming
6. Removal of duplicate and irrelevant information

Natural Language Processing techniques are used to normalize textual content and reduce noise from job advertisements. Preprocessing improves semantic consistency and reduces dimensional complexity in textual representations.

3.3 Stage 1: BERT-Based Sector Classification

In the first stage, the preprocessed job advertisements are classified into their corresponding sectors using Bidirectional Encoder Representations from Transformers (BERT). BERT is a transformer-based deep learning model capable of understanding contextual and semantic relationships between words through bidirectional training. The BERT model is fine-tuned on the collected dataset to perform sector-level classification. Unlike traditional machine learning algorithms such as Naïve Bayes and SVM, BERT captures contextual meaning and long-range dependencies within job descriptions, resulting in improved classification accuracy.

The classification output is represented using the Softmax activation function:

$$y = \text{softmax}(W \cdot h + b)$$

where:

- h represents the contextual embedding generated by BERT,
- W denotes the weight matrix,
- b represents the bias term,
- y indicates the predicted sector probability distribution.

The predicted sector significantly reduces the search space for occupation matching in the second stage.

3.4 Stage 2: Job Title Identification Using Similarity Matching

After predicting the sector, the system filters occupation titles belonging only to the identified sector. This reduces computational complexity and improves prediction efficiency.

In this stage, document embedding techniques are used to generate semantic vector representations for both job advertisements and occupation descriptions. Various embedding techniques such as Word2Vec, GloVe, Doc2Vec, and BERT embeddings are utilized to capture semantic relationships between textual documents.

The similarity between the job advertisement vector and occupation vectors is computed using cosine similarity.

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

where:

- A represents the vector embedding of the job advertisement,
- B represents the vector embedding of the occupation description.

The occupation with the highest cosine similarity score is selected as the final predicted job title.

3.5 Feature Extraction

Feature extraction is performed to identify the most informative keywords from job advertisements. Important textual features such as technical skills, educational qualifications, programming languages, and domain-specific terms are extracted using NLP techniques.

Term Frequency–Inverse Document Frequency (TF-IDF) is also utilized to assign importance weights to words appearing in job descriptions.

The TF-IDF weighting scheme is represented as:

$$TF\text{-}IDF(t, d) = TF(t, d) \times \log\left(\frac{N}{DF(t)}\right)$$

where:

- TF(t,d) is the frequency of term t in document d,
- DF(t) is the number of documents containing the term,
- N represents the total number of documents.

Feature extraction improves semantic relevance and increases similarity matching accuracy.

3.6 Similarity Ranking

The similarity scores generated between job advertisements and occupation descriptions are ranked in descending order. The occupation title with the maximum similarity score is considered the final predicted job title.

The ranking process helps in handling ambiguous job advertisements and improves robustness in occupation identification tasks.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents the experimental evaluation and performance analysis of the proposed Two Stage Job Title Identification System for Online Job Advertisements. The performance of the proposed framework is evaluated using different machine learning and deep learning techniques for sector classification and job title prediction. The obtained results demonstrate the effectiveness of combining BERT-based classification with similarity matching methods for accurate occupation identification.

4.1 Experimental Setup

The experiments were conducted using Python with Natural Language Processing and Machine Learning libraries. The implementation environment consists of:

Table 1: Experimental Setup and System Configuration

Parameter	Specification
Programming Language	Python
Operating System	Windows 10
Processor	Intel Core i5
RAM	8 GB
Deep Learning Framework	TensorFlow / PyTorch
NLP Libraries	NLTK, spaCy, Transformers
Database	MySQL

The proposed system was trained and tested using a dataset of online job advertisements collected from various recruitment platforms. The dataset contains job titles, job descriptions, skills, and occupation categories from multiple sectors such as Information Technology, Agriculture, Finance, Healthcare, and Sales.

The dataset was divided into training and testing datasets using an 80:20 ratio. Preprocessing techniques such as tokenization, stop-word removal, lemmatization, and normalization were applied before training the models.

4.2 Evaluation Metrics

The performance of the proposed system was evaluated using standard classification metrics including Accuracy, Precision, Recall, and F1-Score.

Accuracy

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision

$$Precision = \frac{TP}{TP+FP}$$

Recall

$$Recall = \frac{TP}{TP+FN}$$

F1-Score

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where:

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

4.3 Performance Comparison of Classification Models

The performance of different classification algorithms was compared to evaluate the effectiveness of the proposed two-stage framework. Traditional machine learning algorithms such as Naïve Bayes, Logistic Regression, and Support Vector Machine (SVM) were compared with the BERT-based classifier.

Table 2: Performance Comparison of Machine Learning and Deep Learning Models

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naïve Bayes	71.2	70.5	69.8	70.1
Logistic Regression	76.4	75.9	75.1	75.5
SVM	81.3	80.6	80.1	80.3
BERT Classifier	88.7	88.1	87.9	88
Proposed Two-Stage System	91.5	91	90.7	90.8

The results show that the proposed two-stage system achieved the highest classification accuracy of 91.5%, outperforming traditional machine learning models. The integration of BERT and semantic similarity matching significantly improved contextual understanding and occupation prediction accuracy.

4.4 Analysis of Similarity Matching

The second stage of the proposed system utilizes cosine similarity to identify the most relevant job title from the predicted sector. Various document embedding techniques were evaluated to determine their effectiveness in semantic representation.

Table 3: Accuracy Comparison of Document Embedding Techniques

Embedding Technique	Accuracy (%)
TF-IDF	74.8
Word2Vec	81.6
GloVe	83.2
Doc2Vec	85.4
BERT Embeddings	89.1

The results indicate that BERT embeddings achieved the best semantic representation performance because they capture contextual relationships between words more effectively than traditional embedding techniques.

4.5 Sector-Wise Performance Analysis

The proposed system was also evaluated across different job sectors.

Table 4: Sector-Wise Classification Accuracy of the Proposed System

Sector	Accuracy (%)
Information Technology	94.2
Healthcare	90.4
Finance	88.9
Agriculture	86.7
Sales and Marketing	89.5

The Information Technology sector achieved the highest accuracy because IT-related job descriptions contain more standardized technical keywords and skills compared with other sectors.

4.6 Discussion

The experimental results demonstrate that the proposed two-stage framework significantly improves job title identification performance compared with traditional classification approaches. The use of BERT for sector classification enhances contextual understanding of job advertisements, while similarity matching improves semantic occupation prediction.

The proposed system also reduces computational complexity by limiting occupation matching to the predicted sector instead of comparing against all available occupations. Additionally, document embedding techniques improve semantic similarity analysis and help handle noisy and unstructured job advertisements effectively.

The system performs particularly well in domains with well-defined technical terminology such as Information Technology and Finance. However, some limitations remain in sectors where job descriptions are highly ambiguous or contain limited information.

Overall, the proposed methodology provides an efficient and scalable framework for automatic occupation identification from online job advertisements and can be extended for multilingual recruitment analytics and intelligent job recommendation systems in future work.

.

V. CONCLUSION

In this paper, a Two Stage Job Title Identification System for Online Job Advertisements was proposed using Natural Language Processing, Machine Learning, and Deep Learning techniques. The primary objective of the proposed system was to accurately identify job titles from unstructured online job advertisements by combining

supervised and unsupervised learning approaches. The proposed framework utilized BERT-based sector classification in the first stage and semantic similarity matching in the second stage to improve occupation prediction accuracy.

The experimental results demonstrated that the proposed system significantly outperformed traditional machine learning algorithms such as Naïve Bayes, Logistic Regression, and Support Vector Machine in terms of accuracy, precision, recall, and F1-score. The integration of contextual embeddings and cosine similarity measures improved semantic understanding of job descriptions and enabled efficient identification of relevant occupation titles. The proposed two-stage architecture also reduced computational complexity by limiting occupation matching to the predicted sector.

Document embedding techniques such as Word2Vec, GloVe, Doc2Vec, and BERT embeddings were evaluated, and BERT embeddings achieved the best performance due to their ability to capture contextual and semantic relationships within textual data. The system performed effectively across multiple sectors including Information Technology, Healthcare, Finance, Agriculture, and Sales.

The proposed framework provides an efficient and scalable solution for recruitment analytics, labor market analysis, and intelligent career guidance systems. It can assist organizations in automating job classification processes and help job seekers identify relevant career opportunities based on job advertisement analysis.

In future work, the system can be extended to support multilingual job advertisements, real-time recruitment analytics, and skill extraction mechanisms. Advanced transformer models and Large Language Models (LLMs) can also be integrated to further improve semantic understanding and occupation prediction accuracy.

VI. REFERENCES

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 4171–4186.

[2] Adhikari, A., Ram, A., Tang, R., & Lin, J. (2019). DocBERT: BERT for document classification. arXiv preprint arXiv:1904.08398.

[3] Decorte, J. J., Van Hautte, J., Demeester, T., & Develder, C. (2021). JobBERT: Understanding job titles through skills. arXiv preprint arXiv:2109.09605.

[4] Tran, H. T., Vo, H. H. P., & Luu, S. T. (2021). Predicting job titles from job descriptions with multi-label text classification. arXiv preprint arXiv:2112.11052.

[5] Safikhani, P., Avetisyan, H., Föste-Eggers, D., & Broneske, D. (2023). Automated occupation coding with hierarchical features: A data-centric approach to classification with pre-trained language models. Discover Artificial Intelligence, 3(1), 6.

[6] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[7] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1532–1543.

[8] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 427–431.

[9] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. Proceedings of the 31st International Conference on Machine Learning, 1188–1196.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 5998–6008.

[11] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. O'Reilly Media.

[12] Goldberg, Y. (2017). Neural network methods for natural language processing. Morgan & Claypool Publishers.

[13] Aggarwal, C. C., & Zhai, C. (2012). Mining text data. Springer.

[14] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

[15] Jurafsky, D., & Martin, J. H. (2023). Speech and language processing (3rd ed.). Pearson.