Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Two Stage Job Title Identification System for Online Job Advertisements

Dr. K. Narasimhula
M.Tech, Ph.D.
Associate Professor
Department of CSE
RGMCET, Nandyal

C. Nandini
Department of CSE
RGMCET, Nandyal

B. Bhavana
Department of CSE
RGMCET, Nandyal

E. Manasa Veena
Department of CSE
RGMCET, Nandyal

Y. Indu
Department of CSE
RGMCET, Nandyal

Abstract: The booming development of online recruitment sites has led to high number of job advertisements that have unstructured and haphazardly written job titles. Differences in naming styles, use of abbreviations, and the fact that some roles are assigned in one advertisement makes proper estimation of job titles a complicated exercise. This project offers a twostep job title identification system which aims at extracting and standardizing job titles in online job postings with better accuracy. First stage to identify relevant title candidates, the raw advertisement text is analysed on the basis of linguistic patterns, positioning and contextual relevance. This move minimizes noise by isolating the areas that could to a bigger extent be job titles. The second stage consists of machine learning and semantic similarity to classify and match the extracted candidates with standardized job titles. By isolating detection and normalization, the system is able to accommodate a wide variety of both writing styles and domain specific vocabulary than methods that solely utilize a single stage. As it has been revealed in experimental assessment, the suggested approach will increase identification accuracy and strength and decrease false classification due to the presence of vague or multi-role descriptions. The system assists in a better search of jobs, better recommendation and better labour market analysis since it gives cleaner and consistent data on job titles.

Indexed Terms: Job Title Recognition, IaaS (Internet-as-a-Service) Advertisements, Text Mining, NLP, Information Extraction, Machine Learning, Recruitment Analytics, Named Entity Recognition, Semantic similarity, Unstructured Text Processing.

I. INTRODUCTION

Millions of job advertisements are published daily on online job portals, creating a rapidly growing repository of employment data that is valuable to job seekers, recruiters, and labour market analysts; however, its effective use is often constrained by ambiguous, noisy, and non-standardized job titles [1], [2]. Employers frequently adopt creative naming conventions, abbreviations, and multi-role descriptions, making it difficult to infer the core job role from advertisement text alone [3]. Existing job title extraction methods mainly rely on keyword matching or single-stage classification models [4], which perform poorly when faced with compound roles, informal language, and domain-specific terminology. Additionally, job titles are typically embedded within lengthy descriptions that include responsibilities, required skills, and organizational information, increasing the risk of misclassification or incomplete role identification [5], [6]. To address these limitations, this project proposes a two-phase job title identification framework tailored for online job advertisements, where the first phase extracts and filters candidate job title fragments using linguistic and contextual cues, and the second phase normalizes and maps these candidates to semantically equivalent job titles through machine-based job data processing and analysis, thereby improving robustness across diverse job domains, writing styles, and complex role descriptions [7].

- **The Problem**

The most frequent source of job information is online job advertisements; however, job titles are often vague, inconsistent, or misleading [1], [5]. Employers frequently use alternative naming patterns, abbreviations, combined job titles, or marketing-based expressions instead of explicit and standardized titles [2]. In many cases, the actual job role is hidden within lengthy descriptions that include responsibilities, required skills, and company information, making automated extraction difficult. This ambiguity reduces the performance of job search engines, recommendation systems, and labour market analytics that rely on accurate job titles [1], [2].

Current job title extraction systems typically rely on single-step approaches such as keyword matching or direct classification methods [4]. These methods perform poorly when job advertisements contain multiple possible titles, informal language, or domain-specific terminology [3], [4]. As a result, incorrect or incomplete identification of job titles leads to misclassification, reduced search accuracy, and unreliable analytics [7].

This project addresses the need for a robust and accurate system capable of recognizing and standardizing the primary job title from unstructured online job advertisements. The system is designed to handle noisy text, diverse writing styles, and ambiguous role descriptions while minimizing false identification through semantic learning and intelligent analysis [6], [7]. The proposed framework improves reliability, accuracy, and flexibility in automated job title identification [6].

- **Our Contribution**

The presented project introduces an innovative two-stage algorithm for job title recognition from online job advertisements, specifically designed to address challenges such as unstructured text, inconsistent nomenclature, and ambiguous job descriptions [1], [3]. Unlike traditional single-stage approaches, the proposed framework separates job title detection and job title standardization into two distinct processing phases, enabling more accurate and reliable identification of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

the primary job role [4], [7]. The first phase develops a robust candidate localization mechanism that filters relevant job title segments using contextual, positional, and linguistic cues, thereby reducing noise and improving downstream processing efficiency [2], [4]. The second phase introduces a strong classification and normalization process that maps extracted candidates to standardized job titles using machine learning and semantic similarity techniques, ensuring consistency across diverse job postings [6], [7].

Furthermore, the model is designed to operate without relying on manually defined rules, allowing it to adapt to evolving job market terminology and varied role descriptions [6]. The system demonstrates improved accuracy and robustness when handling multi-role advertisements and informal textual descriptions [3], [7]. Overall, this work provides a scalable and practical solution that enhances job search engines, recommendation systems, and labour market analytics by delivering cleaner and more reliable job title data [1], [2].

- **Content of the Paper**

The following is the structure of this paper. Section I gives the background and the motivation behind automated job title identification and proposes the challenges that are faced in unstructured online job adverts [1], [2], [5]. Part II presents review of related literature on job information extraction, text mining, and natural language processing where the limitations of the available single-stage processes are demonstrated [3], [4], [17], [7], [12]. In Section III, the problem statement is established and the general layout of the proposed job title identification system presented in two steps [3], [6] and [13]. Section IV presents the dataset, upstream processing and feature extraction that is used in detecting candidates in the first-stage [15], [11]. Section V presents the second-stage job title classification and standardization methodology, the applications of which are the machine learning and semantic similarity models [14], [8], [18]. Section VI reports on the experimental set up, performance metrics, and performance analysis of proposed system [17], [12]. Discussion of the research finding, implications, and restrictions of the approach are presented in section VII [9], [10], [16]. Lastly, the concluding part of the paper is in the 8th section where future research and system improvement directions are outlined.

II. LITERATURE REVIEW

Automated job title identification studies have evolved alongside the development of online recruitment systems. Early methods primarily relied on rule-based and keyword matching approaches to extract job titles from relatively structured advertisements [1], [5]. While effective for standardized postings, these methods struggled with informal language, abbreviations, and creatively written job titles [2].

To rise beyond these restrictions, techniques based on machine learning were subsequently proposed, which use Subsequent research incorporated textual features such as term frequency and syntactic patterns to improve job title recognition [4], [6]. Although these techniques enhanced flexibility, they required large amounts of labeled data and faced challenges when handling advertisements containing multiple roles or ambiguous descriptions [3], [4].

More recent studies have explored named entity recognition and deep learning architectures to extract job-related information from unstructured text, achieving improved accuracy [7], [8]. However, many of these approaches treat job title recognition as a single-step task, which often results in misclassification due to noisy or irrelevant content and limited emphasis on job title standardization [6].

III. METHODOLOGY

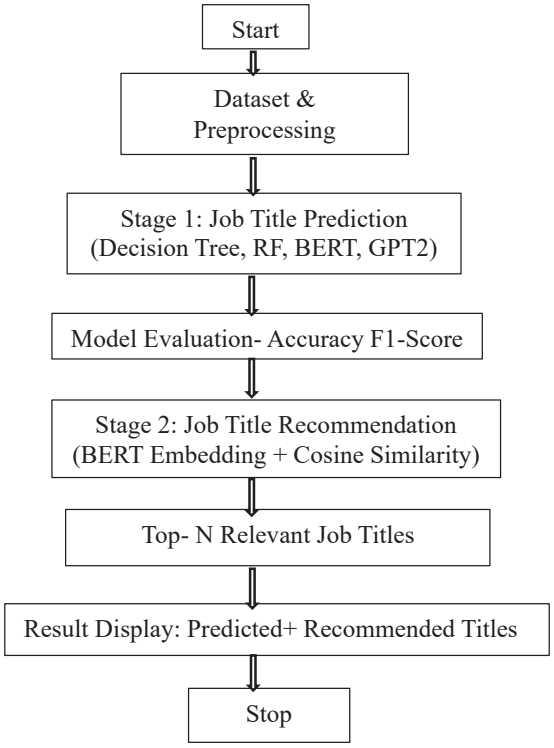

FIGURE 1: Flowchart which illustrates the process of customer churn prediction in the banking sector.

1. Dataset Collection

The dataset that will be employed in the given project will be the Kaggle "Job Descriptions Dataset," comprising of real-life job advertisements on websites like Glassdoor, Merojob.com. It is published under a CC0 Public Domain License, which is appropriated in terms of academic and research application. This data set consists of 2,277 records having three key columns; Job Title, Job description, and link field(optional). It addresses a wide variety of modern tech jobs and offers plenty of textual content, which is why it is perfect to be used with NLP based job title classification and recommendation systems.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Table 1: The information of the dataset is presented in the table

S. n	Job Title	Job Description
1	Django Developer	PYTHON/DJANGO (Developer/Lead) - Job Code(PDJ - 04)
2	Machine Learning	Data Scientist
3	Java Developer	No
4	Machine Learning	Remote, Any where
5	WordPress Developer	Experience:2-5 years
6	Java Developer	Software Developer
7	PHP Developer	Hiring for PHP-Laravel Developers
8	JavaScript Developer	JavaScript developers
9	iOS Developer	Job Code: ID190704
10	Machine Learning	Location: Bangalore
12	Flutter Developer	Zoople Technologies

2. Data Preprocessing and Cleaning

Job descriptions are being preprocessed to help optimize processing by structurally refining the quality of data and improving the effectiveness of NLP models. The first stage eliminates noises like HTML tag, special characters, duplicate elements, and irrelevant words and earmarks lowercasing and lemmatization to synchronize words. Data gaps in vital areas such as job title and job description are handled early enough to maintain completeness and reliability of the dataset to retain data on key roles to be classified and offered proper recommendations. The information is also purified using regular expression cleaning, whitespace, and URL elimination by text normalization. Even more advanced mechanisms, such as tokenization, part-of-speech tagging, and semantic filtering, are effective contextual understanding enhancers. Also, GPT-based semantic mapping matches the job titles that have been extracted and that are being matched against standardized job titles to eliminate ambiguity and increase efficiency in consistency across the listings. This two step architecture enhances robustness, sharps results of analysis, and enables better job suggestion and work market data.

3. Job Title Prediction:

After defining the features, the next step is to train Decision tree, Random Forest and BERT, GPT2

1. Decision Tree:

The supervised learning algorithm employed in this case is Decision Tree which predicts job titles given job description features.

It operates by breaking data down in a recursive fashion to gain the most information based on such measures as Gini Impurity or Entropy

.

2. BERT:

BERT (Bidirectional Encoder Representations from Transformers), a language model which was released in 2018 by Google, is built on top of transformers and learns to interpret both forward and backward context in order to comprehend the text. This learning quality is bidirectional thereby helping the BERT to acquire further semantic meaning, unlike the traditional unidirectional models. BERT is adaption able and generates rich contextual embeddings in recruitment datasets where a writing style is different and abbreviations or vocabulary specific to the field are frequently employed. These embeddings represent the semantic intent of job adverts in general and provide a good basis of characterizing, comparison of similarities and making recommendations. Consequently, BERT enhances the accuracy of job title matching and minimizes the errors in job titles related to ambiguous words or the inconsistent phrasing.

1. $H(D) = -\sum_{i=1}^{n} p_i \log_2 p_i$
   where p_i is the probability of class i in dataset D. The tree aims to choose splits that result in the greatest reduction in entropy, i.e., highest information gain.

   - Gini Impurity is calculated as:

2. $G(D) = 1 - \sum_{i=1}^{n} p_i^2$

Self-Attention Mechanism The self-attention mechanism of the transformer is used in 7BERT to figure out how words are related to each other. The difference in attention of words is determined as: Attention3. (Q, K, V) = SoftMax((QK) / [?]d[?]) V Where: A Query (Q), a Key (K) and Value (V) are matrices based on input embarkations. d L is the dimensionality of key vectors. It is the method that enables the model to put some value on significant words in a sentence that facilitate the model to comprehend complicated job definitions and multi-selections.

3. GPT2:

Generative Pre-trained Transformer (GPT) is a transformer architecture, which is a language model that is trained to comprehend and produce human-like text by discovering grammatical, semantic, and contextual patterns on large and diverse text. As opposed to rule-based methods, GPT is capable of identifying the nuanced patterns of language, and it works exceptionally well in terms of dealing with unstructured job ads that often have mixed vocabulary, abbreviations and phrasing. In systems used to identify job titles, GPT uses contextual interpretation based on analyzing the structure of a sentence and famous words, thus being able to comprehend complex or multi-role descriptions that can be difficult to classify by using a traditional model. Its capacity to find long-range dependencies is into sentence level embedding that indicates the overall role intent. Top-N recommendation policy then

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

also one way that it can find job titles that are not explicitly stated in descriptive text, making it more consistent in the labour market analysis and creates better predictions.

Further, GPT enables the normalization of semantics through facilitation of panoplies to the extracted job title candidates and normalization job role labels. The aspect will reduce ambiguity and improve consistency in numerous positions of the job advertisements. The system implementation of the incorporation of GPT in the 2-stage architecture will be more robust, versatile and accurate that will ultimately cause better job recommendation systems, cleaner data on the labour market, and more consistent recruitment analytics.

4. Model Evaluation-Accuracy and F1-Score

This is a very important process of evaluating the success of the job title identification system by evaluating the models. Accuracy is applied in the present study and is a measure of the proportion of correct titles of the job which are predicted. It only gives a clear picture of the overall performance of the model along with how reliably the system is when it comes to classifying job titles based on market analysis. Overall, presenting Top-N relevant job titles ensures higher reliability, interpretability, and practical applicability of the proposed recommendation system.

5. Job title Recommendation

Job Title Recommendation stage involves the recommendation of the most appropriate standardized job titles that depends on the semantic meaning of a job advertisement. Using contextual embedding encoding, the first stage of the system derives meaningful textual features before projecting the job description into dense vector representations. These embeddings represent the general meaning and duties of the job as opposed to a straightforward keyword matching. Measurements of similarity are then taken between the input embedding and characterizations of standardized job titles stored based on syntactic and semantic features. The system ranks and presents the N top job titles that get the highest similarity scores hence the contextual accuracy and relevance. The use of this technique through management of synonyms, overlapping duties and business terminology can also generate better recommendations, improve job-seeking experiences, reinforce recruitment analytics and complement the overall function of the two-stage approach to identifying job titles.

6. BERT Embedding of input description

The job descriptions in the proposed system get converted to the meaningful numerical values through the application of BERT embeddings that help in maintaining the context of the job descriptions by producing the dynamic word representations based on the surrounding text. In comparison with the traditional embedding algorithms, BERT has the ability to identify two-way relationships between sentences, and therefore it is effective in understanding a complex sentence of a job due to the fact that a single word can be associated with various meanings in different contexts. The processed description is turned eliminates the less relevant job titles which enhance robustness, decision support, job search filtering.

7. Cosine Similarity Calculation

To quantify semantic proximity between two vectors representations, cosine similarity and their direction are of emphasis by calculating the cosine of the angle between the two. In this system it uses a comparison of BERT-generated embeddings of input job descriptions with standardized job title embeddings to establish the relevance. Similarity score has a range of -1 to 1 with higher values showing a high level of semantic congruence. When all the candidate titles have been computed the system ranks in descending order which leaves the most ideal results. Cosine similarity is more effective than simple key word matching because it reflects more semantic associations in this case like synonyms, overlapping roles and functional language that are domain-specific. Improving accuracy, eliminating errors due to the existence of other writing types or the presence of an incomplete description, and improving the overall efficiency of the two-stage job title recognition system, the system achieves this through the introduction of this indicator into the form of the recommendations.

8. Top-N Relevant Job Titles

The procedure of selection of job title to Top-N jobs positions arranges the standardized job titles according to the cosine similarity scores between the candidate title default and job description embedding input. The order of titles in the ranking depends on decreasing similarity, and the most semantically relevant matches are offered at the top of the list. Rather than giving back one title, the system gives several top recommendations in case the position is cross-functional or even hybrid as is often the case in advertising job openings. This increases flexibility and usability through the ability to select a small list of high-confidence options between recruiters, job seekers, and analysts. Furthermore, more skill combinations can be analyzed including Java, Spring Boot, and SQL, which assist to determine relevant technology packages in the market, advance feature engineering through modeling skill dependencies, improve the accuracy of the recommendations and introduce enhanced profiling of candidates basing on in-demand skill packages.

9. Result Display

FIGURE 2 T-SNE VISUALIZATION OF JOB DESCRIPTIONS

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

t-SNE was used with n-components = 2 to plot semantic distribution of job description embeddings into a lower-dimensional space. Each job description and is represented as a point in the 2D plot with a colour code according to its job title. Clear clusters suggest that similar semantically related descriptions are put in similar clusters suggesting that the model is effective at capturing meaningful patterns of roles. Separate clusters were seen in terms of Flutter Developer, iOS Developer, and Backend Developer that demonstrate their high textual similarity. Nevertheless, overlapping sets between the occupations such as JavaScript Developer, Full Stack Developer, and Software Engineer show a common set of duties and similar skill set. The global accuracy of the Decision Tree model was 34 which shows that it is not very effective in differentiating closely related job titles using only structured features and in cases where there is a semantic overlap.

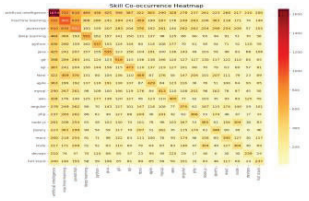

FIGURE 3 SKILL CO-OCCURRENCE HEATMAP

The heatmap highlights strong co-occurrence patterns among technical skills in job descriptions, with darker cells indicating higher frequency associations. Artificial Intelligence is most commonly linked with Machine Learning (732), Deep Learning (484), and Python (436), showing their close relationship, while JavaScript frequently appears alongside HTML (376), React (262), and Node.js (254), reflecting common frontend and full-stack development skill bundles. In classification performance, the Random Forest model significantly outperformed the Decision Tree baseline, achieving 82% accuracy and strong generalization. It delivered high precision, recall, and F1-scores across most job roles, particularly for Django Developer (0.90), Flutter Developer (0.91), Machine Learning (0.92), and iOS Developer (0.98), demonstrating effective semantic pattern recognition. Although performance for Java Developer and Full Stack Developer was moderate due to overlapping responsibilities, balanced class weighting improved fairness across categories, resulting in low class confusion and reliable overall classification.

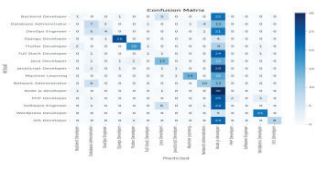

FIGURE 4 CONFUSION MATRIX FOR DT

Embeddings in the form of contextual embeddings work well in learning semantic meaning, classification and recommendation, which are captured in distributions of clusters as shown by t-SNE visualization. Though the model is fairly successful regarding separate delivery types as Django Developer (F1: 0.87) and WordPress Developer (F1: 0.81), implying that classical machine learning models can support homogeneous and well-defined sets of skills, it cannot manage combined roles like Software Engineer, JavaScript Developer, and Full Stack Developer because of the similarity of terms and duties. One example worth highlighting is Node.js Developer which has a recall of very high (0.94) and a precision of extremely low (0.11), that is due to over-prediction through the rise of ambiguous token patterns. These results support the existence of a challenge in separating job titles that are very similar and underscore the necessity of using a more sophisticated semantic modelling to enhance generalization between overlapping classes.

FIGURE 5 CONFUSION MATRIX FOR RF

The training and validation loss curves demonstrate rapid convergence and plateau validation loss which means that they generalize well and that there is little overfitting. Random Forest classifier had an accuracy of 82 percent which is very high, compared to Decision Tree which is more precise. There was high performance metrics, such as precision, recall, and F1-scores, in the majority of job categories. The model was exceptionally good in case of Django Developer, Flutter Developer, Machine Learning, and iOS Developer, whereas it was fairly good in case of Java Developer and Full Stack Developer even though both have similar terminology. In general, the weighting of the classes enhanced the equity within the categories, a fact that proves that the model of the Random Forest is efficient and dependable within the scope of job title classification.

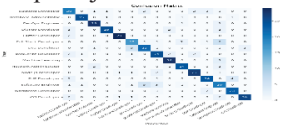

FIGURE 6 CONFUSION MATRIX FOR BERT

The BERT model achieved high results regarding job title classification, which by far would compare well to classical machine learning methods. It had consistent accuracy of precision and recall across several job categories with a total validation accuracy at approximately 80%. Strong F1-score of the position of flutter Developer, DevOps Engineer and Machine Learning is an indication of good separation between classes and strong contextual knowledge. BERT also did fairly well when it comes to conflicting titles such as JavaScript Developer vs. Full Stack Developer, and fewer misclassifications than other models, especially between similar titles, such as Software engineer and Java developer. After first-level training, the curve levelled off because of the viability of the validation. the Findings reveal that BERT is a useful tool when rich semantic context is required, thus it is highly applicable in complex and text-driven job title prediction tasks.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

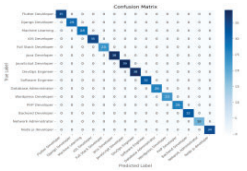

FIGURE 7 CONFUSION MATRIX FOR GPT2

The GPT-2-based model demonstrated great results in 15 types of job titles and 456 samples with a value of zero is misclassifications on 1.00 precision, recalls and F1-scores on the model. This level of accuracy indicates a high level of contextual awareness using very large scale pretraining and proper fine-tuning to the classification problem. Successful preprocessing, tokenization, Cross-Entropy Loss, and efficient training strategies of AdamW and GPU acceleration also led to its success. Although the results show that GPT-2 has a potent ability in text classification in specific domains, subseq uent research should confirm its ability in scaling and generalizability as regards larger datasets which are noisier and have more variants in the real-world.

IV. CONCLUSION & FUTURE ENHANCEMENT

This project compared the classical machine and transformer-based deep learning models in job title classification and contents-based recommendation. Decision Trees underperformed because they were not able to generalize well and they were also sensitive to noisy text whereas the Random Forest had a good performance and a good trade-off between accuracy and efficiency. BERT and GPT-2 transformer-based models demonstrated better semantic comprehension and higher separation between the closely similar job titles, but GPT-2 scores were nearly perfect, indicating potential measures of overfitting or bias in the data set, which will have to be confirmed on more versatile real-life data. Deep learning models were more effective at processing contextual meaning and class imbalance, but required greater amounts of the computational Embedding-based recommendation system mitigated the semantic gaps and cold start problems by similarity scoring, but current shortcomings such as the overfitting risks, overlapping roles, cost of deployment, limited skill ontologies, absence of behavioural personalization, and the necessity of retraining constantly suggest an improvement path in the future which seeks to improve upon hybrid solutions, multi and multi-lingual expansion, advanced balancing approaches, real-time optimization, and career path intelligence.

V. REFERENCES

[1] Z. Guan, J.-Q. Yang, Y. Yang, H. Zhu, W. Li, and H. Xiong,
“JobFormer: Skill-Aware Job Recommendation with Semantic-Enhanced Transformer,” Apr. 2024.
[2] F. Leon, M. Gavrilescu, S.-A. Floria, and A.-A. Minea, “Hierarchical Classification of Transversal Skills in Job Ads Based on Sentence Embeddings,” *Information (Switzerland)*, vol. 15, no. 3, Jan. 2024.
[3]N. Laosaengpha, T. Tativannarat, C. Piansaddhayanon, A. Rutherford, and E. Chuangsuwanich, “Learning Job Title Representation from Job Description Aggregation Network,” Jun. 2024.
[4] S. Garg, C. Sekhar, and L. Kumar, “Unlocking Potential: A Machine Learning Approach to Job Category Prediction,” *Proc. IEEE Region 10 Symposium (TENSYMP)*, 2024.
[5] M. Maitra, S. Sinha, and T. Kierszenowicz, “An Improved BERT Model for Precise Job Title Classification Using Job Descriptions,” *Proc. Int. IEEE Conf., IS*, 2024.
[6] Y. Harshavardhan and M. P. Ramesh, “Enhancing Job Title Identification With BERT and Unsupervised Learning,” vol. 12, pp. 2320–2882, 2024.
[7] A. Heakl, Y. Mohamed, N. Mohamed, A. Elsharkawy, and A. Zaky,“ResuméAtlas: Revisiting Resume Classification with Large-Scale Datasets and Large Language Models,” 2024.
[8] S. N. Charan, G. K. Suhas, L. Yathisha, and S. N. Devananda,“Job Recommendation System: Content-Based and Collaborative Filtering for Predictive Job Recommendation Systems,” *Journal of Information Systems Engineering and Management*, vol. 10, no. 2, pp. 453–459, 2025.