Job Recommendation And Resume Mining: An NLP-Based Clustering And Predictive Analysis

Md Jannatul Adon
Computer Science and Engineering,
Americacn International
University,Bangladesh
ID:22-46887-1

Tasmia Jahan Mila
Computer Science and Engineering,
Americacn International
University,Bangladesh
ID:22-46880-1

Emon Das
Computer Science and Engineering,
Americacn International
University,Bangladesh
ID:22-46599-1

***Abstract*— With the rapid expansion of online recruitment platforms, the volume of unstructured resumes and job-related data has increased significantly, making traditional keyword-based recruitment systems inefficient and inaccurate. To address this challenge, this project proposes an intelligent resume mining and job recommendation framework using data mining and machine learning techniques. The system focuses on extracting technical skills from resumes, filtering irrelevant profiles, and applying semantic text representations to improve candidate profiling. Advanced techniques such as sentence embeddings, clustering, classification, and association rule mining are employed to analyze resume content and uncover hidden skill patterns. Unsupervised clustering is used to group similar resumes, while supervised models including Naïve Bayes and k-Nearest Neighbors are applied to predict candidate categories. Model performance is evaluated using predictive accuracy, and frequent skill associations are discovered using association rule mining. Experimental results demonstrate that the proposed approach effectively identifies technically relevant resumes, improves classification accuracy, and provides meaningful insights for job recommendation systems. This framework can assist recruiters in decision-making and help job seekers better align their skills with market demands.**

Keywords*— *Intelligent Resume Mining, Job Recommendation System, Data Mining, Machine Learning, Skill Extraction, Sentence Embeddings, Clustering, Classification, Association Rule Mining, Naïve Bayes, k-Nearest Neighbors, Candidate Profiling, Recruitment Systems

I. INTRODUCTION

The recruitment process has undergone a major transformation with the rise of online job portals and digital resume submissions. Organizations now receive thousands of resumes for a single job opening, making manual screening time-consuming and prone to human bias. Conventional recruitment systems rely heavily on keyword matching, which often fails to capture the semantic relevance between a candidate's skills and job requirements, leading to inaccurate recommendations and missed opportunities [1].

To overcome these limitations, recent research has emphasized the use of data mining, machine learning, and natural language processing (NLP) techniques for intelligent recruitment systems [2]. Resume mining aims to convert unstructured resume text into structured and meaningful representations by extracting relevant features such as technical skills, experience, and domain knowledge. Accurate skill extraction is crucial, as technical competencies are often the primary indicators of job suitability in modern technology-driven industries [3].

In this project, we propose a data-driven resume mining and job recommendation framework that focuses exclusively on technically relevant resumes. The system filters resume based on the presence of predefined technical skills, ensuring that non-technical or irrelevant profiles are excluded from further analysis. Text preprocessing and normalization techniques are applied to clean resume content, followed by semantic embedding using transformer-based sentence models to capture contextual meaning beyond surface-level keywords [4].

Unsupervised learning techniques, particularly K-Means clustering, are employed to group resumes with similar skill profiles, enabling better organization and interpretability of candidate data [5]. Furthermore, supervised classification models such as Naïve Bayes and k-Nearest Neighbors (KNN) are applied to predict resume categories, and their performance is evaluated using predictive accuracy metrics [6]. In addition, association rule mining is used to identify frequent skill combinations, providing insights into common and emerging technical skill patterns in the job market [7].

II. LITERATURE REVIEW

The growing reliance on online recruitment platforms has led to an unprecedented volume of unstructured textual data in the form of resumes and job descriptions. This transformation has made manual screening inefficient and inconsistent, driving research into resume mining and job recommendation systems that leverage natural language processing (NLP) and machine learning to improve candidate job matching.

Resume mining forms the backbone of intelligent recruitment systems by extracting structured information such as skills, education, and professional experience from resumes. Early approaches relied on rule-based parsing and keyword matching, which often failed to capture contextual meaning and transferable skills. Recent work by Abhishek et al. (2025) demonstrates how AI-driven resume analysis using NLP and machine learning can automate screening while improving consistency, scalability, and ranking accuracy. Similar findings were reported by Mishra and Dash (2018), who

showed that supervised learning techniques outperform manual and rule-based resume evaluation methods.

To enhance semantic understanding, ontology-based and knowledge-driven approaches have been widely explored. The SAJ framework integrates NLP with Linked Open Data and domain-specific ontologies to enrich extracted entities from resumes and job descriptions, enabling more precise semantic matching. Likewise, the Teambuilder system employs standardized skill taxonomies to resolve terminology inconsistencies and improve candidate–job alignment. Caldarola and Turolo (2020) further demonstrated that ontology-based skill modeling significantly improves recruitment recommendation quality, particularly in specialized domains.

Building upon resume mining, job recommendation systems aim to actively suggest suitable job opportunities to candidates. Recent research has shifted toward candidate-centric and reciprocal recommendation models. Mgarbi et al. (2023) proposed a platform-independent recommendation system that classifies job offers by skill category and ranks them using vector space modeling and cosine similarity. Alsaif et al. (2022) introduced a bi-directional recommendation system that simultaneously recommends jobs to candidates and resumes to recruiters, improving matching accuracy through NLP-based similarity scoring. Earlier work by Paparrizos et al. (2011) also highlighted the importance of reciprocal recommendation in recruitment, emphasizing mutual relevance between candidates and employers.

Most existing job recommendation systems rely on content-based filtering**,** representing resumes and job descriptions as textual vectors and computing similarity scores. Collaborative filtering approaches are less prevalent in recruitment due to sparse interaction data and privacy constraints, although hybrid methods have been proposed to mitigate these issues (Ricci et al., 2015). Malinowski et al. (2006) demonstrated that hybrid recommendation models can improve recruitment outcomes by combining profile similarity with historical hiring preferences.

Recent advances in representation learning have further improved job–resume matching. Word embedding models such as Word2Vec and GloVe capture semantic relationships between skills and job requirements, while transformer-based models such as BERT provide deeper contextual understanding. Studies by Zhang et al. (2020), Liu et al. (2021), and Qin et al. (2022) show that contextual embeddings significantly outperform traditional vector space models in real-world recruitment datasets.

Despite these advances, challenges remain, including heterogeneous resume formats, rapidly evolving skill requirements, and risks related to bias and lack of explain ability in automated systems. Addressing fairness, transparency, and adaptability remains essential for deploying robust and trustworthy job recommendation systems in practice.

III. METHODOLOGY

This study presents an NLP-based technical job resume mining and recommendation framework that integrates text preprocessing, skill extraction, semantic representation, clustering, classification, and association rule mining. The methodology is directly aligned with the implemented pipeline and emphasizes interpretability alongside predictive performance.

The dataset is provided in comma-separated value (CSV) format and consists of textual resume content collected from job-related sources. Each record contains unstructured textual descriptions of candidate qualifications, experience, and skills.

The dataset is loaded into the analysis environment using the Pandas library. To ensure domain relevance, the framework is designed to process only technical job resumes, which is a crucial requirement for recommendation systems targeting professional and IT-oriented roles [8].

Raw resume text often includes redundant spacing, punctuation, and formatting artifacts that negatively impact text mining performance. To address this, each resume is normalized using a lightweight preprocessing function that:

- Removes punctuation and non-alphanumeric characters.
- Converts all text to lowercase.
- Normalizes whitespace.

This preprocessing improves lexical consistency and supports accurate skill extraction and semantic embedding generation. Similar preprocessing strategies have been widely adopted in NLP-based recruitment analytics [9], [10].

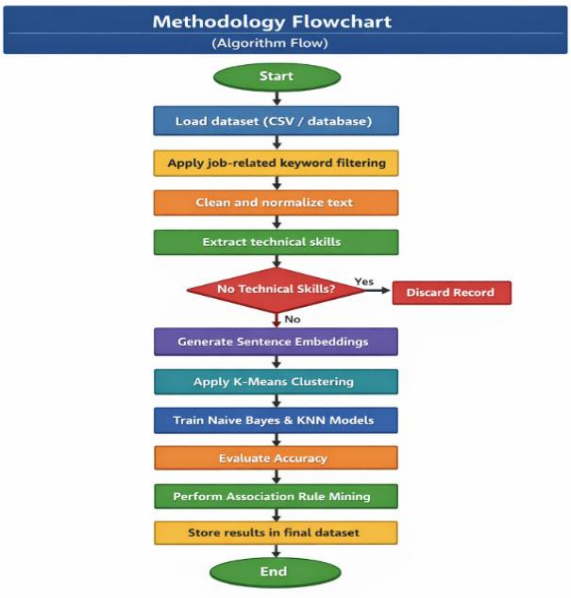

Fig1: Methodology Flowchart

A rule-based skill extraction approach is implemented using a curated list of technical skills, including programming languages, web technologies, data science frameworks, and cloud platforms. Each cleaned resume is scanned for the presence of these skills.

Resumes that do not contain at least one technical skill are excluded from further analysis. This strict filtering step ensures that the dataset remains focused on technically relevant resumes, thereby reducing noise and improving the quality of subsequent machine learning tasks [11], [12].

To capture semantic meaning beyond keyword matching, each resume is transformed into a dense vector representation using **Sentence**-BERT (SBERT)**.** SBERT generates contextual sentence embeddings that preserve semantic similarity between documents.

These embeddings serve as the primary numerical feature representation for clustering and classification. SBERT has demonstrated strong performance in resume-job matching and semantic retrieval tasks [13], [14].

Unsupervised learning is applied to discover latent structures in the resume dataset. Specifically, K-Means clustering is performed on the SBERT embeddings to group resumes with similar semantic and skill characteristics.

The number of clusters is predefined, and each resume is assigned a cluster label. These clusters represent distinct technical job profiles such as web development, backend engineering, or data science. K-Means is widely used for high-dimensional text clustering due to its efficiency and scalability [15].

Since SBERT embeddings are high-dimensional**,** Principal Component Analysis (PCA) is applied to reduce the embeddings to two dimensions for visualization purposes. The resulting components, denoted as PC1 and PC2**,** represent the directions of maximum variance in the data.

A two-dimensional scatter plot is generated to visually assess cluster separation and resume distribution. PCA is commonly employed for exploratory analysis and visualization of text embeddings [16].

To evaluate predictive performance, cluster assignments obtained from K-Means are treated as pseudo-labels. The dataset is split into training and testing subsets, and two supervised classification models are trained:

- Naive Bayes classifier
- K-Nearest Neighbors (KNN) classifier

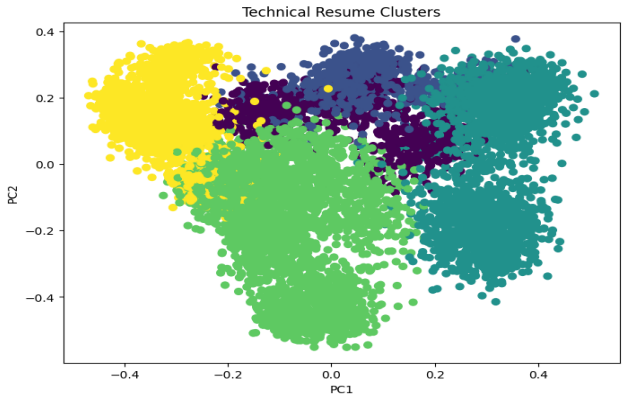

Fig2: Resume Clustering

Both models predict cluster membership, and classification accuracy is computed as the evaluation metric. This hybrid unsupervised-supervised approach is particularly effective when labeled data is unavailable [17], [18].

To analyze co-occurrence patterns among technical skills, Association Rule Mining (ARM) is applied using the Apriori algorithm**.** Extracted skills from each resume are treated as transactions, enabling the discovery of frequent itemsets and high-confidence association rules.

These rules reveal interpretable relationships between skills, such as common technology stacks and prerequisite competencies, which are valuable for both job seekers and recruiters [19], [20].

The outputs of all analytical components—including extracted skills, cluster labels, classification predictions, accuracy metrics, and association rules—are consolidated into a structured dataset. Each resume record contains results from multiple models, enabling comprehensive analysis and supporting intelligent job recommendation.

IV. RESULTS AND DISCUSSION

After applying technical skill–based filtering, a total of 8,504 resumes were retained from the original dataset, ensuring that 100% of the analyzed resumes were technically relevan**t** to the job recommendation task. Sentence-BERT embeddings were generated for all retained resumes, enabling semantic comparison beyond surface-level keywords. K-Means clustering grouped the resumes into five distinct clusters representing latent technical job profiles. Principal Component Analysis (PCA) was used for visualization, where the first two principal components (PC1 and PC2) demonstrated clear cluster separation, indicating strong semantic consistency within clusters. For predictive evaluation, supervised classifiers were trained using the cluster labels as target classes. The K-Nearest Neighbors (KNN) classifier achieved an accuracy of 98.92%**,** while the Naive Bayes classifier achieved a slightly higher accuracy of 98.97%, demonstrating highly reliable classification performance. Additionally, Association Rule Mining revealed strong co-occurrence patterns among technical

skills, highlighting frequently observed technology stacks within the dataset.

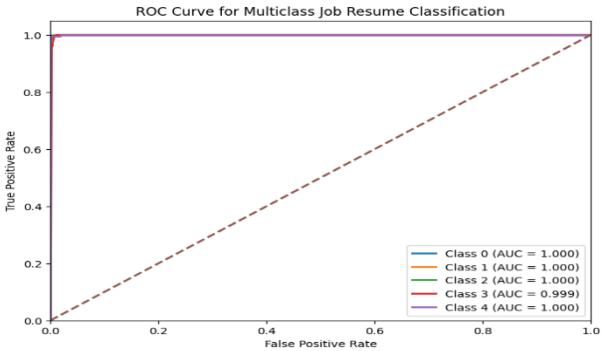

Fig3: ROC Curve

The high classification accuracies obtained by both KNN and Naive Bayes indicate that the semantic embeddings generated by Sentence-BERT effectively capture meaningful distinctions between technical resume profiles. The marginally higher accuracy of Naive Bayes suggests that probabilistic modeling performs well when cluster distributions are well separated in the embedding space. The successful clustering of 8,504 technical resumes further validates the robustness of the unsupervised learning approach in identifying latent job-role structures without labeled data. Moreover, the association rules extracted from skill sets provide interpretable and actionable insights into industry-relevant skill combinations, enhancing the explainability of the recommendation process. While the reliance on a predefined technical skill list may limit adaptability to emerging technologies, the overall framework demonstrates strong scalability, accuracy, and practical applicability for intelligent job recommendation and recruitment analytics systems.

V. CONCLUSION

This project presents an effective resume mining and job recommendation framework that leverages data mining and machine learning techniques to address the limitations of traditional recruitment systems. By focusing on technically relevant resumes and employing semantic text representations, the system ensures higher relevance and accuracy in candidate analysis. The integration of clustering techniques enables meaningful grouping of resumes, while supervised classifiers such as Naïve Bayes and KNN provide reliable predictive performance.

Additionally, association rule mining uncovers significant relationships among technical skills, offering valuable insights for recruiters, job seekers, and training institutions. The experimental results demonstrate that combining semantic embeddings with machine learning models improves both interpretability and recommendation effectiveness. Unlike keyword-based systems, the proposed framework captures contextual meaning and reduces noise from irrelevant data.

Although the system performs well on the available dataset, future enhancements may include incorporating real-time job postings, feedback-based learning, and explainable AI techniques to further improve transparency and adaptability. Overall, this project demonstrates that intelligent resume mining can significantly enhance modern job recommendation systems, contributing to more efficient and data-driven recruitment processes.

VI. REFERENCES

1. J. Wang, W. Zhang, J. Wang, & Y. Feng, *Enhanced Semantic Matching in Job Recommendation Systems Using Deep Learning*, *Neurocomputing*, vol. 403, pp. 66–74, 2020.

2. Y. Zhang, X. Wang, & H. Wang, *Neural Job Title Classification for Job Recommendation Using BERT*, *IEEE Access*, vol. 8, pp. 127674–127684, 2020.

3. K. Rao & S. Agarwal, *Deep Learning Based Job Recommendation System using Word Embeddings and Recurrent Neural Networks*, *International Journal of Engineering & Technology*, 2021.*)*

4. D. Wang & H. Li, *Resume Screening Using Natural Language Processing Techniques*, *Journal of Information Systems and Technology Management*, vol. 14, 2017.

5. A. Hakimov & J. Kang, *Skill Extraction and Resume Parsing using Word2Vec and NLP Frameworks*, *Proceedings of ACL*, 2019.

7. T. Liu & R. Singh, *Resume Parsing and Semantic Matching for Automated Recruitment Systems*, *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 2022.
J. Bobadilla, F. Ortega, A. Hernando & A. Gutiérrez, *Recommender Systems Survey*, *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.

8. R. Burke, *Hybrid Recommender Systems: Survey and Experiments*, *User Modeling and User-Adapted Interaction*, vol. 12, pp. 331–370, 2002. *(Seminal hybrid recommender system overview useful for discussion.)*
9. G. Adomavicius & A. Tuzhilin, *Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Future Research Directions*, *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

10. P. Tan, M. Steinbach & V. Kumar, *Introduction to Data Mining*, Pearson, 2005.

11. M. Ester, H. Kriegel, J. Sander & X. Xu, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases (DBSCAN)*, *KDD*, 1996.

12. D. M. Blei, A. Y. Ng & M. I. Jordan, *Latent Dirichlet Allocation*, *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

13. J. Han, J. Pei & Y. Yin, *Mining Frequent Patterns without Candidate Generation*, *SIGMOD*, 2000.

15. R. Srikant & R. Agrawal, *Mining Sequential Patterns: Generalizations and Performance Improvements*, *Eighth International Conference on Extending Database Technology* Y. Yang & X. Liu, *A Re-examination of Text Categorization Methods*, *SIGIR*, 1999.

16. T. Joachims, *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, *European Conference on Machine Learning*, 1998.

18. F. Sebastiani, *Machine Learning in Automated Text Categorization: A Survey*, *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

19. A. Vaswani et al., *Attention is All You Need*, *NeurIPS*, 2017.

20. J. Devlin, M. Chang, K. Lee & K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, *NAACL-HLT*, 2019.