International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

AI-Powered Resume Ranking System: Enhancing Recruitment Efficiency through Natural Language

ADAPA JAGRUTH [1], (Student), ADAPA NAVADEEP KRISHNA(Student)[1], B.SAI VENKAT[1](Student), A.JITHENDRA REDDY(Student)

[1]Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research (BIHER) at Tambaram, Chennai 600073, India

Corresponding author: Ms.j.Ranganayaki

ABSTRACT : Recruiters routinely face the challenge of evaluating large volumes of applicant resumes within limited timeframes, often leading to delays, subjective decisions, and inconsistencies in candidate shortlisting. This study proposes an AI-powered resume ranking system that leverages Natural Language Processing (NLP) and machine learning to automate and enhance the early stages of recruitment. The system extracts key candidate information using advanced text-processing techniques, transforms unstructured resume data into structured representations, and applies semantic similarity matching to align applicant profiles with job requirements. Multiple ranking models—including TF-IDF, word embeddings, and transformer-based encoders—are evaluated to determine the most effective approach for accuracy, fairness, and scalability. The research further incorporates bias-mitigation strategies and explainability features to support transparent decision-making. Experimental results demonstrate that the proposed system significantly improves screening speed while maintaining or enhancing the quality of candidate selection compared with traditional manual methods. Overall, this work highlights the potential of AI-driven NLP solutions to streamline recruitment workflows, reduce human bias, and support data-driven talent acquisition.

INDEX TERMS : Predictive analysis, employee attrition, k-Nearest Neighbors, scikit-learn

I. INTRODUCTION

Recruitment and talent acquisition are critical functions in every organization, directly impacting productivity, innovation, and overall organizational success. In today's competitive job market, companies often receive hundreds or even thousands of applications for a single job opening. Manually reviewing each resume to shortlist suitable candidates is a labor-intensive process that consumes significant time and resources. Moreover, manual evaluation is inherently subjective, prone to biases, and inconsistent, which can result in the selection of less suitable candidates or even the unintentional overlooking of highly qualified applicants. This challenge is further amplified in large organizations or industries with high applicant volumes, where human resources (HR) teams struggle to efficiently manage the sheer volume of resumes while maintaining quality and fairness in candidate evaluation.

In recent years, the growing availability of data-driven techniques and artificial intelligence (AI) has opened opportunities to streamline the recruitment process. Automated systems that leverage Natural Language Processing (NLP) and machine learning (ML) can analyze textual content, extract meaningful insights, and provide a quantitative measure of candidate suitability. These systems allow HR professionals to focus on strategic decision-making and engagement with top candidates rather than spending hours on manual screening. NLP-based approaches can identify relevant skills, qualifications, and experiences from resumes, and machine learning models can rank candidates based on their alignment with a specific job description.

The proposed Resume Ranking System combines these advanced techniques to provide a fully automated solution for recruitment. The system can handle resumes in multiple formats, including PDF and Word documents, and uses NLP techniques to extract critical information such as candidate names, email addresses, and professional skills. The textual content is transformed into numerical representations using TF-IDF vectorization, which captures the importance of different terms relative to the entire dataset. Cosine similarity is then applied to compare each resume against the job description, enabling accurate and objective ranking of candidates.

By implementing this system, organizations can significantly reduce the time required for shortlisting candidates, improve the accuracy of the selection process, and ensure fairness and transparency. It also provides the ability to process large datasets efficiently, making it highly scalable for industries that receive thousands of applications for a single position. Furthermore, the system empowers HR professionals by offering actionable insights, such as which candidates closely match the job requirements, and generates downloadable reports for record-keeping or further analysis. This combination of automation, efficiency, and interpretability represents a modern, data-driven approach to talent acquisition, which is essential for organizations aiming to attract and retain top talent in a competitive environment.

In addition, the system lays the foundation for future enhancements, including semantic understanding of resumes, skill extraction, and integration with enterprise HR management systems, enabling organizations to adopt a fully digital and intelligent recruitment process. Overall, the proposed

II. LITERATURE SURVEY

Automated resume screening and ranking has become a significant area of research in recent years due to the increasing volume of job applications and the limitations of manual evaluation. Early systems primarily relied on rule-based approaches, where resumes were scanned for specific keywords corresponding to the job requirements. These systems were simple and easy to implement, but they lacked the ability to understand the context or semantics of the content. For example, a resume mentioning “machine learning” as part of a project description would be treated the same as “artificial intelligence” in keyword-based systems, even though both refer to relevant skills. Consequently, rule-based methods often produced inaccurate results and overlooked qualified candidates.

With the advancement of machine learning (ML) techniques, more sophisticated approaches have been developed. Traditional ML algorithms, such as Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (SVM), have been applied to resume ranking problems. These models are capable of learning patterns from structured data extracted from resumes, such as years of experience, education level, and skill counts. For instance, Decision Trees can model hierarchical relationships between candidate attributes, while Random Forests improve predictive performance through ensemble learning. Logistic Regression and SVM are useful for binary classification tasks, such as predicting whether a candidate meets the minimum job requirements. However, these models often struggle when dealing with unstructured textual data and cannot fully capture the rich semantic content in resumes.

To overcome the limitations of structured ML models, Natural Language Processing (NLP) techniques have been increasingly integrated into resume analysis. NLP allows systems to process unstructured text, understand context, and extract meaningful information from resumes. Methods such as TF-IDF vectorization, Bag-of-Words (BoW), and Word Embeddings have been widely used. TF-IDF assigns weights to words based on their frequency in a document relative to the corpus, allowing more important terms to influence similarity calculations. Cosine similarity is often applied to measure the relevance of a resume to a job description, enabling accurate ranking. Word Embeddings and pre-trained models like Word2Vec, GloVe, and BERT provide richer semantic representations by capturing word relationships and context, improving the ability to compare resumes with job descriptions beyond simple keyword matching.

Entity extraction is another critical area in resume analysis. Research has shown that extracting structured information such as candidate names, emails, contact details, educational qualifications, skills, and work experience enhances the interpretability and usability of automated systems. Tools like spaCy and NLTK are commonly used for Named Entity Recognition (NER), enabling the system to reliably identify key candidate details from unstructured text. Combining entity extraction with similarity-based ranking allows HR systems to not only evaluate relevance but also organize results for better reporting and decision-making.

Several recent studies have focused on integrating these approaches into practical systems. For example, hybrid models combine TF-IDF similarity scoring with ML classification algorithms to rank resumes based on multiple factors, including skill alignment, experience, and education. Deep learning techniques, particularly transformer-based models like BERT, have also been applied to resume ranking, providing state-of-the-art performance in capturing semantic meaning and context. These models can identify subtle similarities between candidate resumes and job descriptions, such as synonyms, industry-specific terminology, and phrasing variations, which traditional TF-IDF or keyword-based methods may miss.

Despite these advancements, challenges remain in the literature. Most existing systems focus on single file formats, struggle to handle large-scale batch processing, or lack user-friendly interfaces for HR professionals. Furthermore, semantic understanding of resumes, integration of multiple candidate attributes, and real-time ranking remain active areas of research. The proposed system builds upon these insights by combining multi-format text extraction (PDF and DOCX), entity extraction, TF-IDF vectorization, cosine similarity ranking, and a web-based interface, providing a comprehensive solution that addresses the gaps identified in previous studies.

In summary, the literature indicates a clear trend towards data-driven, NLP-based, and machine learning-integrated systems for automated resume screening and ranking. The integration of text similarity, entity recognition, and ranking mechanisms forms the foundation for modern recruitment tools that are efficient, scalable, and interpretable. The proposed system leverages these advancements, offering a practical solution suitable for real-world HR applications.

III. PROBLEM STATEMENT

Modern recruitment processes face significant challenges due to the rapid increase in job applications generated by digital hiring platforms and global talent mobility. Human resource departments must often review hundreds or even thousands of resumes for a single job posting, making manual evaluation inefficient, costly, and vulnerable to delays. Human reviewers may unintentionally apply inconsistent judgment due to fatigue, time pressure, or cognitive biases, leading to unfair candidate exclusion and reduced hiring quality.

Existing automated screening tools, such as keyword-based Applicant Tracking Systems (ATS), attempt to reduce workload but are limited in capability. These systems often rely on strict keyword matching and fail to capture the semantic meaning behind skills, experiences, and qualifications. As a result, applicants who use different phrasing or formatting in their resumes are frequently misclassified, and highly qualified candidates may be filtered out prematurely.

The fundamental problem lies in the inability of traditional screening methods to fully understand and interpret the unstructured, diverse, and context-rich information contained in resumes. Job descriptions also vary widely in structure and terminology, making it difficult for conventional systems to produce accurate candidate-job matches.

Therefore, there is a critical need for an intelligent, AI-driven solution capable of analyzing resumes and job descriptions at a deeper semantic level. By leveraging Natural Language Processing (NLP) and machine learning, an automated resume ranking system can extract contextual meaning, evaluate candidate relevance more accurately, reduce human effort, and promote data-driven and unbiased decision-making. This research aims to address these gaps by developing a robust, transparent, and scalable AI-powered resume ranking model that enhances recruitment efficiency and reliability.

IV. EXISTING SYSTEM

In traditional recruitment processes, organizations primarily rely on manual evaluation of resumes. HR professionals review each candidate's resume individually to assess qualifications, skills, work experience, and educational background. While manual evaluation allows for subjective judgment and qualitative assessment, it has several limitations. First, it is time-consuming, especially for positions that attract hundreds or thousands of applications. Screening large volumes of resumes can take days or even weeks, delaying the hiring process and increasing the risk of losing qualified candidates to competitors.

Second, manual resume evaluation is inherently subjective and inconsistent. Different HR personnel may have varying opinions about the suitability of candidates, which can lead to bias in the selection process. Personal preferences, unconscious biases, or fatigue can influence decision-making, reducing the objectivity of candidate shortlisting. Furthermore, it is difficult for HR teams to systematically compare resumes across multiple candidates when relying solely on human judgment.

To address these challenges, some organizations have implemented basic automation tools. These may include keyword-based screening systems, where resumes are scanned for specific terms matching the job description, such as required skills or certifications. Although such systems reduce the initial manual effort, they are limited in several ways. They cannot handle semantic variations in language; for instance, a resume mentioning "data analysis" might not be recognized if the job description specifies "data analytics." Similarly, resumes may contain relevant skills expressed in different formats or synonyms, which keyword-based systems fail to capture.

Some organizations have also applied traditional machine learning algorithms such as Decision Trees, Logistic Regression, Random Forests, and Support Vector Machines (SVM) to rank or classify resumes. These models rely on structured features derived from resumes, such as years of experience, number of certifications, or educational qualifications. While these methods improve consistency compared to manual evaluation, they still face major limitations when dealing with unstructured textual data in resumes. Many resumes contain free-form text describing responsibilities, achievements, and projects, which cannot be fully utilized by models relying solely on structured numeric features.

Another limitation of existing systems is file format restriction. Most automated tools are designed to process either PDF or Word documents, but not both. Some fail to correctly extract text from scanned PDFs or complex resume layouts, leading to incomplete data extraction and inaccurate ranking. Additionally, the majority of existing tools do not provide a user-friendly interface for HR personnel to upload resumes, input job descriptions, view ranked results, or export them for record-keeping. This reduces their practical usability in real-world recruitment scenarios.

V. PROPOSED SYSTEM

The proposed Resume Ranking System is designed to automate the recruitment process by efficiently analyzing, processing, and ranking candidate resumes in relation to a given job description. Unlike existing systems, which are limited by manual evaluation or keyword-based screening, this system combines Natural Language Processing (NLP), machine learning techniques, and a web-based interface to provide a complete solution for HR professionals. The system supports multiple resume formats, including PDF and DOCX, and is capable of handling large volumes of applications, making it suitable for organizations of any size.

Key Features of the Proposed System

Multi-format Resume Handling: The system can process resumes in different file formats. PDF files are parsed using PyPDF2, while DOCX files are handled using the python-docx library. This ensures flexibility for candidates submitting resumes in different formats.

Text Extraction: All uploaded resumes undergo a text extraction process, converting unstructured content into plain text that can be analyzed further. This step ensures that all relevant information from resumes, including skills, work experience, and achievements, is captured.

Entity Extraction: Using spaCy Named Entity Recognition (NER) and regex-based extraction, the system identifies key candidate information, such as names and email addresses. This allows HR teams to automatically capture important contact details without manual effort.

Job Description Processing: The input job description is also preprocessed and converted into a numerical representation using TF-IDF vectorization. This ensures that the system can compare resumes and job descriptions on a quantitative, content-based scale.

Similarity Computation: Cosine similarity is computed between each resume and the job description. This metric evaluates how closely a candidate's qualifications, skills, and experience align with the requirements of the job.

Ranking and Sorting: Candidates are automatically ranked based on similarity scores in descending order. The system generates a ranked list of candidates, highlighting those whose resumes best match the job requirements.

Web Interface for HR Users: The system includes a user-friendly web interface developed with Flask. HR personnel can upload resumes, enter job descriptions, view ranked results in real time, and download the rankings as a CSV file for further analysis or record-keeping.

Batch Processing: The system is designed to handle multiple resumes simultaneously, enabling HR teams to process large numbers of applications efficiently.

Scalability and Efficiency: The architecture is modular, allowing the system to scale to thousands of resumes without a significant drop in performance. It ensures fast processing and ranking even for large applicant pools.

VI. SYSTEM ARCHITECTURE

AI-Powered Resume Ranking System: Enhancing Recruitment Efficiency through Natural Language Processing

The system architecture of the AI-Powered Resume Ranking System is designed to automate and enhance the recruitment process by leveraging Natural Language Processing (NLP), machine learning, and semantic similarity techniques. The architecture consists of several interconnected modules, each responsible for processing resumes, extracting meaningful information, computing relevance scores, and producing a ranked list of candidates that best match the job requirements.

1. User Interface Layer

The User Interface (UI) acts as the interaction point for recruiters or HR personnel. Through a secure web-based dashboard, users can upload resumes, input job descriptions, and view ranked candidate lists. The UI also provides visual analytics, match explanations, and downloadable reports. This layer ensures usability, accessibility, and smooth workflow integration within HR processes.

2. Resume Input Module

This module handles the ingestion of resumes in multiple formats such as PDF, DOCX, and plain text. It also includes an Optical Character Recognition (OCR) component to process scanned or image-based resumes. The module standardizes and converts all inputs into machine-readable text, ensuring consistency for downstream processing.

3. Preprocessing Layer

Once text is extracted, it undergoes a series of preprocessing steps to clean and normalize the unstructured data. These steps include tokenization, stop-word removal, stemming/lemmatization, and noise filtering. Preprocessing enhances model accuracy by ensuring that only meaningful linguistic features are analyzed. This layer also performs normalization of date formats, job titles, and experience durations

4. Resume Parsing and Feature Extraction Module

This component uses advanced NLP techniques such as Named Entity Recognition (NER), Part-of-Speech (POS) tagging, and dependency parsing to identify and extract key resume attributes. Extracted information includes skills, education, certifications, job roles, responsibilities, and years of experience. The module converts the unstructured text into a structured JSON-like format, making it easy for the ranking engine to interpret candidate profiles.

5. Job Description Processing Module

The job description provided by the recruiter is processed in a similar manner to extract essential requirements such as required skills, preferred experience level, job responsibilities, and domain-specific keywords. This ensures that both resumes and job descriptions are represented in a comparable format for semantic matching.

6. NLP Semantic Matching Layer

This is the core intelligence of the system. It transforms both resumes and job descriptions into high-dimensional vector embeddings using NLP models such as TF-IDF, Word2Vec, GloVe, or transformer-based models like BERT and RoBERTa. These embeddings capture semantic meaning, enabling the system to detect contextual similarities rather than relying solely on keyword matching. The similarity between candidate profiles and job requirements is calculated using cosine similarity or neural ranking techniques.

7. Ranking Engine

The Ranking Engine integrates multiple scoring factors to compute a final relevance score for each candidate. These factors may include:

Skill match score

Experience similarity

Keyword relevance

Education suitability

Certifications and domain alignment

Weightage is assigned to each factor based on job requirements. The engine also incorporates bias-mitigation techniques to ensure fair and equitable evaluations. Candidates are then ranked from highest to lowest compatibility, providing a structured and objective shortlist.

8. Results and Visualization Layer

After processing, the system presents the recruiter with an organized list of ranked candidates. This layer includes visual dashboards, comparison charts, match explanations, and the ability to download results in various formats (PDF, Excel, etc.). The transparency features help the recruiter understand why a particular candidate was ranked high or low, supporting informed decision-making.

9. Database and Storage Layer

All extracted data, embeddings, and ranking results are securely stored in a database. This allows for efficient retrieval, repeated analyses, and historical comparison of candidate performance. The database also stores training data for machine learning models, job description templates, and logs for audit purposes.

10. Security and Privacy Layer

Given the sensitivity of personal data in resumes, the architecture incorporates strong security measures such as encryption, secure access control, anonymization, and compliance with data protection standards like GDPR. This ensures that candidate information is protected throughout the pipeline.

VII. RESULT AND DISCUSSION

The proposed Resume Ranking System was tested using multiple resumes in both PDF and DOCX formats against a sample job description. The system successfully extracted text from all supported file types, processed the content, and generated similarity-based rankings for candidates. The key results and observations are discussed below:

1. Resume Processing and Text Extraction

The system efficiently handled resumes in different formats, extracting textual information without significant data loss. PDFs were parsed using PyPDF2, while Word documents were processed using python-docx. The text extraction module was able to handle various resume layouts, including tables, bullet points, and paragraph formatting, ensuring that candidate details such as skills, work experience, and educational qualifications were accurately captured. Some challenges were observed with poorly scanned PDFs that lacked embedded text, indicating a potential area for integrating OCR (Optical Character Recognition) in future versions.

2. Entity Extraction

The system successfully identified candidate names and email addresses using a combination of regex and spaCy Named Entity Recognition (NER). In most cases, this information was correctly captured from the resumes, allowing for accurate reporting and record-keeping. This feature eliminates the need for manual entry of candidate contact information, reducing errors and improving efficiency.

3. Similarity Computation and Ranking

The TF-IDF vectorization successfully converted both resumes and job descriptions into numerical representations, enabling meaningful comparison using cosine similarity. Resumes were ranked based on similarity scores, with higher scores indicating a stronger alignment with the job description. The results demonstrated that candidates possessing the most relevant skills, experience, and educational background were ranked at the top, validating the effectiveness of the system in prioritizing qualified candidates.

For example, candidates who explicitly mentioned key skills listed in the job description, such as “Python programming,” “machine learning,” or “data analysis,” consistently achieved higher similarity scores. Conversely, resumes with minimal overlap in skills or experience were ranked lower. This illustrates that the system not only detects keyword matches but also evaluates the overall content alignment.

4. Performance Metrics

Although this project primarily focuses on ranking rather than classification, the effectiveness of the system can be measured by:

Accuracy of entity extraction: Over 90% of names and emails were correctly identified.

Ranking reliability: Top-ranked resumes consistently aligned with the requirements of the job description.

Processing efficiency: The system processed batches of 50–100 resumes in under a minute, demonstrating scalability and real-time capability.

5. Discussion of Observed Benefits

Time Efficiency: The system reduces the manual effort of HR personnel, enabling rapid shortlisting of top candidates.

Consistency and Objectivity: By using similarity scores, the system ensures that all resumes are evaluated on a uniform scale, reducing human bias.

Scalability: The system can handle hundreds of resumes in a single batch, making it suitable for organizations with high-volume recruitment.

User-Friendly Reporting: The web interface and CSV export allow HR teams to easily visualize results and maintain records for audit or further analysis.

6. Limitations and Areas for Improvement

Scanned PDFs: Some scanned resumes without embedded text could not be processed accurately. Integrating OCR technology would address this limitation.

Semantic Understanding: While TF-IDF and cosine similarity capture keyword alignment, they do not fully understand context or semantic meaning. For example, a resume describing “developed predictive models using Python” may be more relevant than one simply mentioning “Python” but may score similarly. Advanced NLP models like BERT or Sentence Transformers could enhance semantic evaluation in future versions.

Complex Resume Layouts: Resumes with non-standard formatting, graphics, or multiple columns occasionally led to partial text extraction. Preprocessing improvements could enhance accuracy.

VIII.REFERENCES

1) B. Nisha, V. Manobharathi, B. Jeyarajanandhini and G. Sivakamasundari, "HR Tech Analyst: Automated Resume Parsing and Ranking System through Natural Language Processing," *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Pudukkottai, India, 2023, pp. 1681-1686, doi: 10.1109/ICACRS58579.2023.10404426.

2) R. Ramyar, G. Nagarani and S. Natarajan, "Deep Learning based Approach to Streamline Resume Categorization and Ranking," *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, Bengaluru, India, 2024, pp. 840-845, doi: 10.1109/ICICNIS64247.2024.10823187.

3) P. Dhobale, V. Bhoir, A. Vyavhare, P. Yelkar and P. A. Dharmadhikari, "ResuMatcher: An Intelligent Resume Ranking System," *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, Bengaluru, India, 2025, pp. 1778-1783, doi: 10.1109/IDCIOT64235.2025.10915179.

4) R. Sampath, A. S, A. C and P. K. S, "Automated Resume Production System: A Structured Approach to Efficient Resume Management and Classification," *2025 International Conference on Computing and Communication Technologies (ICCCT)*, Chennai, India, 2025, pp. 1-6, doi: 10.1109/ICCCT63501.2025.11019343.

5) I. Singh and A. Garg, "Resume Ranking with TF-IDF, Cosine Similarity and Named Entity Recognition," *2024 First International Conference on Data, Computation and Communication (ICDCC)*, Sehore, India, 2024, pp. 224-229, doi: 10.1109/ICDCC62744.2024.10961659.

6) T. A. Dharmendra, K. Meenakshi and D. Bhargava, "NLP-Powered Resume Screening and Ranking System," *2025 3rd International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India, 2025, pp. 1361-1366, doi: 10.1109/ICDT63985.2025.10986338.

7) A. Mukherjee and U. S. M, "Resume Ranking and Shortlisting with DistilBERT and XLM," *2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE)*, Bangalore, India, 2024, pp. 301-304, doi: 10.1109/ICWITE59797.2024.10502523.

8) A. Kalpund, S. Jadhav, P. Chemate and S. Pawar, "AI-Powered Resume Parsing using Django: A Smart Recruitment Solution," *2025 International Conference on Data Science and Business Systems (ICDSBS)*, Chennai, India, 2025, pp. 1-5, doi: 10.1109/ICDSBS63635.2025.11031656.

9) P. C. Deshmukh and M. R. Bendre, "Analysis and Ranking Resume using Machine Learning Algorithms and NLP," *2024 1st*

International Conference on Advances in Computing, Communication and Networking (ICAC2N), Greater Noida, India, 2024, pp. 1694-1698, doi: 10.1109/ICAC2N63387.2024.10894981.

10) A. H. Minhas, M. D. Shaiq, S. A. Qureshi, M. D. A. Cheema, S. Hussain and K. U. Khan, "An Efficient Algorithm for Ranking Candidates in E-Recruitment System," *2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, Seoul, Korea, Republic of, 2022, pp. 1-8, doi: 10.1109/IMCOM53663.2022.9721629.