ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

Resume Screening Automation with NLP Techniques

[1] DR.T.SRAVANTI, [2] P.AKSHITHA, [3] P.SAHASRA, [4] S.KOUSHIK GOUD,[5] S.AKHIL REDDY
[1]Associate Professor, Department of CS, Sri Indu College of Engineering & Technology ,Hyderabad.
[2,3,4,5] U.G. Scholar, Department of CS, Sri Indu College of Engineering & Technology, Hyderabad.

Abstract: This project introduces an automated solution designed to simplify and enhance the resume screening process using Natural Language Processing (NLP) techniques. The system is capable of extracting essential details from resumes, including skills, work experience, and educational qualifications, to effectively match applicants with job requirements. By utilizing advanced NLP models, the system interprets and ranks resumes according to their relevance, thereby minimizing the time and manual effort required by recruiters. The proposed approach applies NLP methods to analyze unstructured resume data and convert it into a structured summary by identifying key attributes such as skills, education, and professional background. By filtering out unnecessary or irrelevant information, the system makes the screening process more efficient, allowing recruiters to evaluate candidates more quickly and accurately. After completing the text processing stage, the system implements a vectorization technique along with cosine similarity to compare resumes with job descriptions. Based on the similarity scores generated, candidates are ranked according to how well they match the job criteria. This ranking assists recruiters in identifying the most suitable applicants for a given position. The system is designed to enhance the precision of candidate selection while promoting a faster and less biased hiring process. The output is displayed through an intuitive interface, presenting a ranked list of candidates along with extracted details and matching scores, enabling recruiters to make informed decisions with ease. Overall, this tool significantly reduces the effort and time involved in the initial screening phase, thereby improving the overall efficiency of recruitment.

Keywords: *Candidate Screening, Skills Matching, Experience Analysis, Recruitment Efficiency, Hiring Process Optimization, Automated Candidate Selection, Resume Parsing.*

How to Cite: Dr. G. Sravan Kumar; K. Varshitha; K. Keerthi; N. Vikesh (2025). Resume Screening Automation with NLP Techniques. *International Journal of Innovative Science and Research Technology*, 10(5), 3294-3298. https://doi.org/10.38124/ijisrt/25may1909

I. INTRODUCTION

The application is built to handle both PDF and DOCX file formats, extracting textual data from these documents to enable further analysis. This extraction process is crucial for converting the unstructured data of a resume into a format that the application can understand and process. Once the text is extracted, it undergoes a preprocessing stage. This involves cleaning the text by removing non-alphabetic characters, converting all text to lowercase, and eliminating common English stop words. The aim of this preprocessing is to simplify the text and focus on the most relevant keywords, ensuring a more accurate comparison. A key aspect of the application is its ability to identify and extract relevant skills from the cleaned resume text. This is achieved by comparing the words in the resume to a predefined list of skills. Furthermore, the application extracts contact information, including name, email, and phone number, from the resumes. This information is essential for recruiters to easily reach out to potential candidates. The application employs a similarity calculation method to quantify how closely a resume matches the job description. This calculation involves transforming both the resume text and the job description into numerical vectors and then using cosine similarity to measure their resemblance. The modern era of recruitment is increasingly shaped by the need to efficiently process and evaluate large volumes of job applications. Traditional methods, heavily reliant on manual screening, often prove to be time-consuming, resource-intensive, and susceptible to human biases. To address these challenges, there's a growing demand for automated systems capable of streamlining the initial stages of candidate selection. The core objective is to swiftly identify the most promising candidates from a pool of applicants, allowing recruiters to focus their attention on in-depth assessments and interviews. This process allows for an objective ranking of candidates based on their suitability for the role. Finally, the application presents the results in a structured format, typically a table, displaying each candidate's contact information, relevant skills, and a percentage score indicating the match with the job description. The candidates are ranked by their match percentage, allowing recruiters to quickly identify the most promising individuals. In essence, this application automates and enhances the initial stages of resume screening, saving

time and effort for recruiters while ensuring a more data-driven and objective evaluation of candidates.

II. RELATED RESEARCH

The automated screening of resumes has become an increasingly vital area of research and development within the field of human resources and talent acquisition. The surge in digital documentation and online applications has led to an overwhelming volume of resumes for recruiters to process manually, creating a bottleneck in the hiring process. This challenge has fueled the need for efficient and effective automated systems capable of parsing, analyzing, and ranking resumes based on their relevance to specific job requirements. Research in this domain draws upon a combination of techniques from natural language processing (NLP), information retrieval, and machine learning to streamline and optimize the initial stages of candidate selection. One of the fundamental aspects of this research involves the extraction of meaningful information from resume documents. Resumes, while intended to present a candidate's qualifications, often exhibit significant variability in format, structure, and terminology. This heterogeneity poses a considerable challenge for automated systems. Consequently, a substantial body of work is dedicated to developing robust methods for extracting key data points such as contact information, work experience, educational background, and, crucially, skills. Optical Character Recognition (OCR) techniques are frequently employed to convert scanned or image-based resumes into machine-readable text, which then serves as the input for subsequent processing. However, OCR accuracy can vary, necessitating error correction and post-processing steps. Furthermore, research explores various parsing strategies, including rule-based systems, statistical models, and more recently, deep learning approaches, to accurately identify and categorize different sections and elements within a resume. The extraction of skills is a particularly critical area of focus. Skills are often a primary indicator of a candidate's suitability for a role, and accurately identifying and categorizing them is essential for effective screening. Research in this area involves the creation of comprehensive skill taxonomies and the development of algorithms to map skills mentioned in resumes to these standardized lists. It's crucial to ensure that these systems are fair, unbiased, and compliant with equal opportunity employment laws.

III. METHODOLOGY

The methodology employed in this system for automated resume screening integrates several key computational techniques to efficiently process and evaluate candidate information against job descriptions. The process begins with the ingestion and extraction of text from resume files, accommodating both PDF and DOCX formats. For PDF files, the system leverages the pdfplumber library, which facilitates the extraction of textual content page by page, ensuring that all text within the document is captured. This method is chosen for its robustness in handling various PDF layouts and text encodings, crucial for maintaining the integrity of the extracted data. For DOCX files, the docx2txt library is utilized, providing a straightforward way to retrieve the text embedded in these documents. The dual approach ensures compatibility with the two most common resume formats, enhancing the system's versatility. Once the text is extracted, it undergoes a preprocessing phase aimed at standardizing the data and removing noise that could impede accurate analysis. This preprocessing involves the use of regular expressions to eliminate non-alphabetic characters, thereby focusing the analysis on relevant textual content. The text is then converted to lowercase to ensure uniformity, treating identical words consistently regardless of their original casing. Stop words, common English words that carry little semantic weight, are removed using the Natural Language Toolkit (NLTK) library and its built-in English stop word list. This step is critical for reducing the dimensionality of the text and improving the efficiency and accuracy of subsequent analyses. Finally, the processed text is tokenized and reassembled, forming a clean, standardized text string ready for further processing. The extraction of skills from the cleaned text is a crucial step in the methodology. A predefined list of skills is loaded from an external file, providing a controlled vocabulary for skill identification. The system then identifies relevant skills by comparing the tokens in the cleaned text against this list. This intersection of terms allows for the precise extraction of skills explicitly mentioned in the resume, ensuring that only relevant qualifications are considered. This method is efficient and accurate, focusing on skills that are directly pertinent to the job description.

IV. ARCHITECTURE

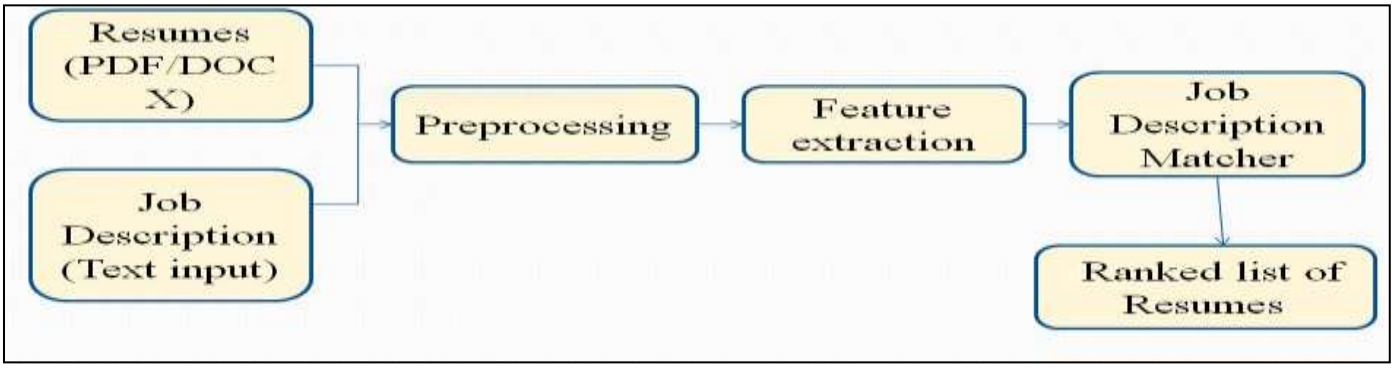

Fig 1 System Architecture

The core of this system lies in its ability to process and analyze resume data, a task that presents several architectural challenges. Firstly, the system is designed to handle multiple file formats, specifically PDF and DOCX. This necessitates an architecture that incorporates format-specific parsing modules, each optimized for its respective format. The choice of libraries becomes crucial here; libraries like pdfplumber, fitz (PyMuPDF), and docx2txt are employed to extract textual content, showcasing a modular approach to document parsing. This modularity allows for potential future expansion to accommodate other file types without requiring a complete overhaul of the system.The extracted text then undergoes a preprocessing stage, a critical architectural component for ensuring data quality and consistency. This preprocessing involves cleaning the text by removing non-alphanumeric characters, converting all text to lowercase, and eliminating stop words. These steps, while seemingly simple, are fundamental to standardizing the input for subsequent analysis, directly impacting the accuracy of skill extraction and similarity calculations. The architecture here emphasizes text manipulation and normalization, highlighting the importance of data preparation in information retrieval systems. Skill extraction is another key architectural element. The system employs a predefined list of skills against which the extracted text is compared. This "skills list" acts as a knowledge base, and the system's architecture must efficiently manage and utilize this list for accurate skill identification. The design involves tokenizing the cleaned text and then performing a set intersection to find matching skills. This approach highlights a set-based architecture for skill matching, prioritizing speed and accuracy in identifying relevant qualifications. The scalability of this skill matching component is important; as the skills list grows, the architecture should maintain efficient lookup times. Contact information extraction represents a different architectural challenge. Here, the system relies on regular expressions to identify patterns corresponding to emails, phone numbers, and names. This pattern-based extraction requires a robust regular expression engine and a well-defined set of patterns to handle variations in contact information formats. The architecture must balance flexibility in pattern matching with the need for accuracy to avoid misidentification of data. Error handling is also important in this module, as contact information may not always adhere to strict formats or may be absent altogether.

V. EVALUATION

The core functionality revolves around accurately extracting pertinent information from resumes and effectively matching it against job descriptions. Therefore, evaluation must begin with assessing the precision and recall of the information extraction modules. For instance, the ability to correctly identify skills, contact details, and other key qualifications is paramount. Erroneous extraction can lead to misrepresentation of a candidate's profile, potentially resulting in qualified individuals being overlooked. This demands a rigorous evaluation of the text processing techniques employed, including the effectiveness of regular expressions and natural language processing (NLP) methods in handling variations in resume formats and writing styles. Furthermore, the matching algorithm's performance is crucial. Cosine similarity, often used to quantify the degree of resemblance between the job description and the resumes, needs to be evaluated for its ability to prioritize candidates whose qualifications genuinely align with the job requirements.

VI. RESULT

Resume screening application designed to streamline the candidate selection process by analyzing resumes and comparing them against a given job description. This application leverages several key processes, starting with the uploading of resumes in either PDF or DOCX formats and the input of a job description. The system then extracts the textual content from these documents, regardless of their format, using specialized functions for PDFs and DOCX files. Once extracted, the text undergoes preprocessing to remove irrelevant characters, convert it to lowercase, and eliminate common English stop words, ensuring that only essential information remains for analysis. Following this cleanup, the application identifies and extracts crucial information from the resumes, specifically focusing on skills and contact details such as name, email, and phone number. The extraction of skills is facilitated by comparing the processed text against a predefined list of skills, ensuring that only relevant technical skills are captured. This list comprises a variety of skills including programming languages, cloud computing platforms, databases, and other technical proficiencies. With the cleaned text and extracted skills in hand, the application proceeds to calculate the similarity between each resume and the job description. This similarity calculation employs TF-IDF vectorization and cosine similarity to quantify how closely a resume matches the requirements outlined in the job description. The results of this analysis are then compiled into a structured format, typically a pandas DataFrame, which presents a clear overview of each candidate's information alongside their match percentage. This DataFrame is sorted by the match percentage, allowing for easy identification of the most suitable candidates. The application's output provides a concise summary of each candidate, including their name, contact information, relevant skills, and a percentage indicating how well their resume aligns with the job description. This facilitates a more efficient and data-driven approach to resume screening, enabling recruiters to quickly identify and prioritize the most promising candidates. The system effectively automates the initial screening process, saving time and effort in identifying candidates whose qualifications closely align with the requirements of a specific job. Furthermore, the use of text preprocessing and skill extraction ensures that the comparison is based on relevant information, reducing the potential for human bias and improving the accuracy of the screening process. The application's ability to handle multiple resume formats and provide a ranked list of candidates based on their match percentage makes it a valuable tool for recruiters and hiring managers seeking to streamline their recruitment efforts.

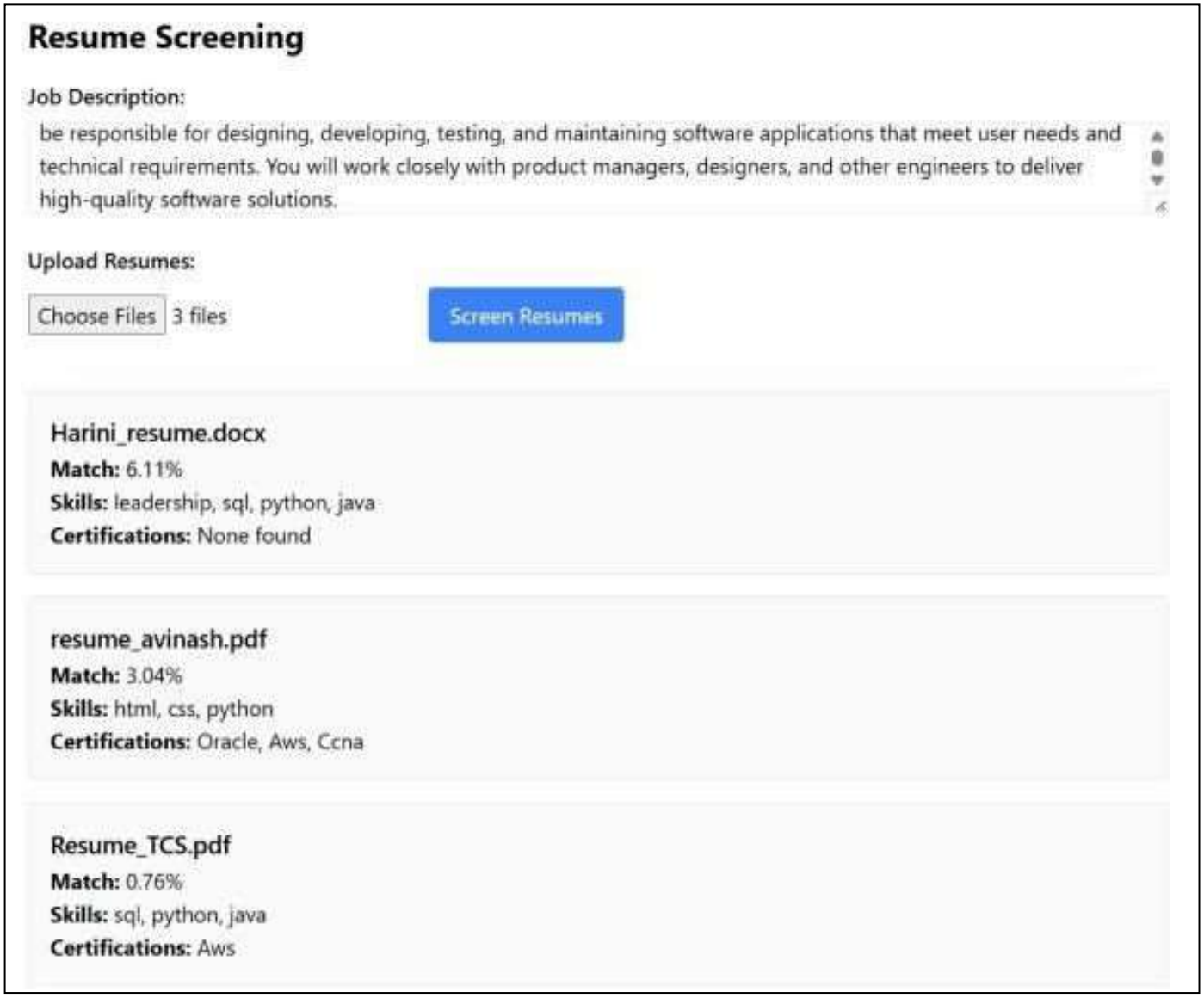

Fig 2 Resume Screening

VII. CONCLUSION

This innovative Resume Screening App streamlines the hiring process by efficiently matching uploaded resumes with a given job description. It intelligently extracts crucial information from both PDF and DOCX resume formats, including the candidate's name, email, phone number, and a comprehensive list of skills. The core functionality relies on advanced text preprocessing to clean the raw data, removing noise and standardizing the text for accurate analysis. By leveraging TF-IDF vectorization and cosine similarity, the application quantifies the relevance of each resume to the job description, providing a clear percentage match. This allows recruiters to quickly identify top candidates based on skill alignment and overall content similarity, moving beyond manual screening and significantly reducing the time and effort involved in the initial stages of recruitment. The results are presented in an organized table, showcasing the candidate's contact details, their match percentage, and their identified skills, all sorted by match percentage for immediate insights.

FUTURE SCOPE

The resume screening application has significant potential for future expansion. Enhancements could include integrating natural language processing (NLP) for more nuanced understanding of resume content, moving beyond keyword matching to semantic analysis for improved accuracy in candidate-job fit. Implementing machine learning models for predictive analytics could help identify top performers based on historical data. Expanding the range of extracted entities to include education, work experience, and projects would create more comprehensive candidate profiles. Furthermore, incorporating a user feedback loop to refine matching algorithms and providing interactive dashboards for in-depth analysis of candidate pools would significantly boost the application's utility for recruiters. Future development could focus on integrating with Applicant Tracking Systems (ATS) for seamless workflow automation, from resume upload to interview scheduling. Adding a feature for automated skill gap analysis against a job description would empower candidates to identify areas for improvement. Implementing sentiment analysis on

resume content could provide insights into a candidate's soft skills and communication style. Furthermore, enabling customizable weighting for different matching criteria (e.g., prioritizing skills over experience) would offer greater flexibility to recruiters. Finally, exploring blockchain for secure and verifiable credential management could enhance trust and efficiency in the hiring process.

REFERENCES

[1]. Malhotra, A., & Singh, Y. (2019). Automated Resume Ranking Using Natural Language Processing and Machine Learning. In Proceedings of the 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN).

[2]. Jain, D., Kumar, V., & Arora, A. (2018). An Approach towards Resume Classification and Recommendation using NLP. In Proceedings of the 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT).

[3]. Sinha, A., Bhatia, S., & Bhattacharya, S. (2020). Resume Quality Assessment Using BERT Embeddings. In Proceedings of the 2020 International Conference on Artificial Intelligence and Machine Vision (AIMV). arXiv:2006.13111

[4]. Luo, C., Zhang, H., & Wang, W. (2022). A Resume Screening System Based on BERT and Knowledge Graph. In Journal of Intelligent & FuzzySystems,42(3),2221–2232.

[5]. Tripathy, A., & Agrawal, A. (2020). Intelligent Candidate Shortlisting Using NLP and Deep Learning. In Procedia Computer Science, 167,1170–1179.

[6]. Xue, Y & Ghosh, R. (2018). Deep Learning-based Resume-Job Matching Solution. In Proceedings of the 27th International Conference on Computational Linguistics (COLING).URL: https://aclanthology.org/C18-1113/

[7]. Patil, S., & Patil, S. (2020). Resume Parsing and Matching using Natural Language Processing. In International Research Journal of Engineering and Technology (IRJET), 7(5), 2539–2542.

[8]. Lavanya, K., & Nandhini, M. (2021). Resume Screening Using NLP and Machine Learning. In International Journal of Engineering Research & Technology (IJERT), 10(4), 96–101.

[9]. Li, H., Liu, L., & Lin, C. (2017). Job Resume Matching with Learning-to-Rank and Word Embeddings. In Proceedings of the IEEE International Conference on Big Data (Big Data), 138–145.

[10]. Vasudevan, N., & Nandhini, V. (2022). Automated Recruitment System using NLP and Deep Learning. In Journal of Emerging Technologies and Innovative Research (JETIR), 9(1), 73–79.