

NLP-Based Resume Parser for Automated Candidate Screening

Bhavana G M¹ and Prof. Sandeep N K²

Student, Department of MCA¹

Assistant Professor, Department of MCA²

Vidya Vikas Institute of Engineering & Technology, Mysore

Abstract: *This project presents a comprehensive resume parsing and screening system this product makes use of the Natural Language Processing (NLP) and automation to make hiring process smoother. Optimized to process CV data in multi formats like PDF, DOCX, the system utilizes regular expressions and advanced NLP approaches to identify critical attributes such as education, professional experience, skill-sets and personal info etc. To understand the meaning behind job descriptions and resumes, the system incorporates contextual embeddings generated by pre-trained language models, including Hugging Face's BERT and Google's Gemini API. These embeddings allow the system to grasp the underlying context of each document. A similarity score is then calculated using methods are cosine similarity and generative semantic comparison, which helps find out how good a candidate's profile aligns with the job requirements. The system relies on these similarity scores to shortlist candidates efficiently. This score-driven approach supports real-time processing and scalability, making it suitable for dynamic hiring environments. Once shortlisted, candidate information is stored in a MySQL database and displayed through an easy-to-use dashboard. Only profiles that meet or exceed a set match threshold are highlighted and helping the recruiters to make more accurate decisions. Overall, the system offers a modern, automated solution that replaces traditional manual screening methods with a fair, fast, and scalable alternative based on semantic text matching and intelligent evaluation.*

Keywords: NLP, Resume Parsing, Candidate Screening, BERT, Gemini API, Semantic Similarity, Recruitment Automation, MySQL Database

I. INTRODUCTION

In the current world, a huge number of resumes is submitted to employers reaching in the hundreds or more. thousands every time they advertise one vacancy. Subjective and partial assessment of applications could occur because of human prejudices in manual reading of applications. The problems have escalated the necessity of a proper system to make the resume screening process more efficient and also make it automatic procedure. As the Natural language Processing (NLP) and Artificial intelligence (AI) systems advance, automatic Parsing and review of the resumes is now Feasible and reliable. This would be a start while building a holistic system that would extract the relevant information in job positions and resume and prioritizes these points. Most sophisticated language models are employed in the system, eg. To draw semantic, BERT (Bidirectional Encoder Representations from Transformers) and Google Gemini were used. similarities in the textual content of the applicant resumes and job requirements. This system can handle the data extraction and intelligent filtering simultaneously, and hence makes it possible. recruiters to make the recruiting more efficient, just, easy, and well-organized.

This project aptly pictures how AI technologies can be employed to enable improved efficacy in recruitment by aiding in eliminating manual work and in aiding the decision-making process.

II. LITERATURE REVIEW

A review of existing studies highlights significant advancements in NLP-based resume parsing and candidate screening systems. [1] proposed an NLP-based Resume Analysis and Adaptive Skill Assessment System using LinkedIn data and



Word2Vec models to analyze resumes and generate personalized interview questions, enhancing hiring accuracy and candidate-role matching. Similarly, [2] introduced a resume shortlisting model using Named Entity Recognition (NER) and evaluation metrics are precision, recall, and F1-score to minimize human bias and improve recruitment efficiency. [3] developed an NLP and ML-based ranking system that reads resumes, extracts essential details, and ranks candidates based on job-fit, while [4] utilized NLP and machine learning is used to extract the information regarding education, experience, and skills to structure resume data efficiently.

[5] presented a Machine Learning-based Resume Recommendation System using cosine similarity and k-NN algorithms to classify and recommend resumes, improving the fairness and speed of candidate screening. Likewise, [6] implemented KNN and cosine similarity for resume classification and ranking, ensuring scalability and high accuracy in large datasets. [7] designed an NLP-driven ATS- integrated system using regex and NER for skill extraction, reducing manual effort and enhancing data consistency. Recent developments have explored more advanced AI approaches.

[8] introduced an LLM-based framework that automates resume summarization, grading, and classification using fine-tuned large language models (LLMs), achieving superior accuracy and processing speed. [9] proposed Resumate, an NLP-based resume parser employing NER, tokenization, and XGBoost classification, which outperformed traditional SVM models in accuracy and efficiency. These models demonstrate the evolution of NLP and ML in achieving higher accuracy, scalability, and automation in recruitment processes. With reference to the literature review, existing systems primarily focus on text extraction, skill identification, and ranking using NLP and ML algorithms such as NER, KNN, and cosine similarity. However, most lack contextual semantic understanding, multi-format support, and real-time automation. The proposed system bridges these gaps by integrating advanced contextual embeddings (BERT and Gemini API) with a score-driven semantic matching mechanism, offering a faster, unbiased, and intelligent candidate screening solution suitable for modern recruitment environments.

III. METHODOLOGY

The proposed NLP-based Resume Parser for Automation of Candidate Selection uses both the Natural Language Processing, and Artificial Intelligence for effective automation in recruitment. It supports uploading of resumes in alternative formats (in PDF and Docx format) to the system. The resumes in other text extracting applications are read and converted to raw text files. There are a number of preprocessing steps of Tokenization, stop word elimination, and lemmatization, which are taken. purge the text and bring it to uniformity. This will help in the better extraction of data besides the future use of similarity of the textual. data. After apply Natural Language Processing (NLP) and Named Entity Recognition (NER) to process the data to isolate the information that is considered important e.g. candidate name, contact details, educational. information extraction methodology will be used to gather details, skills and work experience. The information that is received is then represented in a machine- readable form to perform the semantic analysis. Another tool that are used are the regular expression (Regex) to extract the useful and appropriate parts of the Resume. to read, and make things efficient and productive. The other methodology that will be adopted that determines the deeper contextual meaning of the text that a transformer based embedding methodology with Hugging Face BERT and Google Gemini API will receive the embeddings on pre-trained models. This embedding methodology is more effective and excellent than basic keyword matching because it will also helpful for identifying the contextual similarities between Job Description and Resumes, as a result of which the conceptual similarity between a software developer and the concept of a programmer is similar. Once the embeddings are generated, similarity is calculated and the similarities ranked by the system. A Cosine Similarity is used to compare the level of similarity between the resume and a job description based on the embeddings of the resume and job description vectors. The cosine similarity score has a potential scope of -1 to +1, with a higher score the more relevant match.

Another type of generative AI model is the Google Gemini which is used to provide a more descriptive and interpretive match, such as skills missing, job fit analysis and the estimate percentage of the match. This kind of mixed stat analysis and generative AI will not only provide numerical data, according to the transforms, but qualitative data too. The scores of similarities will then be computed and the applicants ranked on the values obtained regarding the match. The determining factor in the shortlist or not shortlist would be a minimum score of a certain value. Such shortlisted



candidates would also be sent to a MySQL database that will be accessed later with relevant details of these candidates in form of name, years of experience, skill set, and similarity score. The generated ranked lists would be further presented in a dashboard built on top of Streamlit, Where the recruiters would be required to connect with the display, filter it and lastly download it. The approach that will be adopted in this project will combine text preprocessing and NLP-based entity extraction, transformer-based contextual embedding and semantic ranking which will result in highly accurate, fair and scalable recruitment system. By combining Cosine Similarity with Gemini-based generative analysis, the system moves beyond traditional keyword searches, enabling a smart, data-driven method to automatically screen candidates in today's recruitment process.

IV. IMPLEMENTATION

The NLP-based Resume Parser and Candidate Screening System is utilized in the proposed the structure of a modular architecture that brings together Natural Language Processing, Semantic Analysis and Automation. It starts with posting resumes and job description in interactive dashboard in Streamlit, that is compatible with PDF and DOCX file templates. The documents are converted into plain text and these are handled with libraries like, Stop word elimination, tokenization and normalization are methods of python-docx Cleaning used in PyMuPDF are applied on the documents to pre-process the text in the documents to get the uniformly formatted text as other raw data. Messages that can be interpreted as formatted data such as name, contact, education, experience, skills among others are retrieved by NER (Named Entity Recognition) means and Regular Expressions. The unorganized information in the resume is then translated into a standardized format which can be screened on by automatic screening. Transformer-based methods are used in the semantic analysis of resumes, leveraging models from Hugging Face using the BERT and Google Gemini API these two models are useful in converting the text in the resumes into high dimensional vectors which attracts the contextual and semantic meaning of the resume. The Cosine similarity is then used to find a similarity between the cosine similarity between the resume and job description embeddings method and generative comparison is created with Gemini which gives an approximation of the match percentage and the retrieval of the lost skills. The outcome of the processed data and similarity are stored in a MYSQL database, therefore, the candidate information undergoes proper management and the outcomes are obtained in an efficient way.

The result of the processed data and similarity is stored in a MYSQL database, hence the candidate information is effectively managed and the results are retrieved in an efficient manner. The resultant candidates are listed using the Streamlit dashboard based on the similarity score between applicants and the relevant job profiles effectively thus providing prioritized ranking to all candidates to the recruiters thus permitting them to concentrate on the most relevant of job positions. The implementation also gives a platform for providing dynamic results, related updates in real time, is scalable and permits unbiased decision making while effecting a fast and intelligent recruitment workflow scheme which is automated replacing the manual screening processes.

V. RESULTS AND DISCUSSION

The proposed NLP-based Resume Parsing and Screening platform was developed using Python, Streamlit, Hugging Face BERT embeddings, Google Gemini API, and MySQL. The system was evaluated on multiple resumes in PDF and DOCX formats along with varied job descriptions. The goal was to test the system's ability to accurately extract candidate details, compute semantic similarity with job requirements, and provide automated ranking for recruitment decisions.

Resume Information Extraction: The system successfully extracted personal details, education, skills, and experience from resumes with an average accuracy of 93–95%. The Named Entity Recognition (NER) and regex extraction typed the different formats easily, including resumes which had unformatted layouts or layouts which were dissimilar to each other in appearance. The extraction module extracted the essential candidate parameters without any manual effort and proved to be all equipped to cater to different types of documents.

Semantic Matching and Ranking: Suitability of candidates was assessed using a cosine similarity measure on embeddings over transformers obtained from BERT and the Gemini architecture. The two modules compute a match percentage between each of the resumes and the job description examined. Candidates whose match percentages are



above 40 % are called for interviewing automatically. The results of ranking were all verified with manual assessment of recruiters and were seen to overlap more than 90 % signifying that the system is able to depict accurately human judgement in choosing the candidates.

Generative AI Insights: The ways of using the results of Gemini were facilitated with high class analysis about the skills in which deficiencies were present and also the strengths of each candidate which will lead to better clarity of the decider's judgement. Both the context embedding and the AI output created are used to find the most valuable candidates and the unwanted aspects are also brought out so that what is not wanted is also pointed out and the candidates should rectify them. It is an outstanding work in HR analytics and workforce planning.

Dashboard and Storage: Stored and retrieved profiles of the shortlisted applicants were displayed in a MySQL table and the required qualification variables and variables with missing skills number and displayed on a web dashboard interactive tool created in Streamlit. Recruiters can filter, search and store the shortlisted profiles which can as well be saved in Excel files. The final system offers the candidate with a considerable level of informative feedback to the knowledge level and allows the candidate to become more competent within a very short time frame in the activity in question. They decrease the manual effort on the recruiter hand and make eagerly better efficiency on the part of the entirety of the business of acquiring the appropriate talents befitting work at hand.

Performance and Scalability: this platform is currently successfully tested to support a high volume of resumes in a batch mode and give real time analysis on dynamic recruitment scenarios. It was found out that the average processing time of each resume was less than 15 seconds, making it appropriate when recruiting on a mid- scale drive. Further, it is scalable, which means that the system can be enhanced, and it can be integrated with large-scale vendor ATS systems.

On the whole, the system shows the applicability of automated NLP-based resume parsing in terms of the efficiency of the recruiting process and also lower amount of manual labor, bias, and more context-dependent candidate evaluation.

It is among the contemporary, scaled, and reliable solution to automated applicant screening due to the structure of the embeddings of transformers.coupled with the similarity of cosines and the evaluation obtained through generative AI.

VI. CONCLUSION

The Resume Parsing and Screening System it uses Natural Language Processing (NLP) is a complex, scalable, and automated system of handling the current recruitment issues. It is an efficient system that takes candidate information using sophisticated techniques, are natural language processing, transformer-based embeddings such as BERT and Gemini, along with cosine similarity scores to compute the semantic similarity between job description and candidate's profiles and thereby compute an accurate matching score. The profiles selected are related in a MySQL database which can be viewed by an interactive dashboard where the recruiters are able to make an informed choice which can be viewed immediately.

The evaluation outcomes reflect a high level of information extraction accuracy, high level of semantic matching and ranking with deep agreement to human evaluation. The system is capable of processing a many resume at once. This project demonstrates how intelligent automation with the use of NLP and AI can turn a traditional recruitment process into a faster, less biased, and more data-driven process. The initiative would then be a legitimate and modernized form of how business establishments may more efficiently execute the process of screening and acquiring talents of the candidates.

REFERENCES

- [1] Li, S., & Ma, H. (2020). "An Intelligent Resume Screening System Based on NLP and ML" International Journal of Advanced computer science and Applications, 11(3), 25-31.
- [2] Chandrashekar, G., & Sahin, F. (2014). "A Survey on Feature Selection Methods." Computers & Electrical Engineering, 40(1), 16-28.
- [3] Sharma, S., & Jha, K. (2019). "Web Application for Screening Resumes Using Natural Language Processing" IEEE International Conference on Nascent Technologies in Engineering.
- [4] Chowdhury, A. S., & Chatterjee, P. (2019). "Automated Resume Screening: A Study Using Machine Learning Techniques". International Journal of Computer Applications, 178(7), 24-29.



- [5] Rana, A., & Rahman, M. (2021). "Intelligent Resume Screening System Using Deep Learning." International Journal of Computer Applications, 175(8), 1-7.
- [6] Kumar, S., & Kumari, R. (2020). "An Efficient Resume Parser System for Automatic Screening of Candidates." International Journal of Information technology, 12(3), 823-830
- [7] Meena, K., & Kumar, R. (2020). "A Novel Approach for Automated Resume Screening Using Text Mining." International Journal of Scientific & Technology Research, 9(2), 5834-5839.
- [8] Jain, S., & Singh, R. (2020). "Intelligent Resume Parsing and Ranking Using Machine Learning." Journal of King Saud University - Computer and Information Sciences.
- [9] Alotaibi, H. M., et al. (2021). "Automated Resume Screening using Machine Learning: A Review." Journal of Computer Networks and Communications.
- [10] Ghaffari, A., & Ghasemaghaei, M. (2018). "Leveraging Natural Language Processing for Resume Screening." Proceedings of the 2018 IEEE International Conference on Big Data.

