iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68831

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Resume Clustering and Job Description Matching

Archana V. Ugale[1], Sanap Gayatri[2], Gunjal Rutik[3], Ghumare Amit[4], Andhale Shreyas[5]

Abstract: ***The efficient functioning of drainage systems is critical for urban infrastructure, playing a key role in managing stormwater, preventing flooding, and safeguarding public health and safety. ABSTRACT:***
In today's competitive job market, both job seekers and employers are increasingly turning to automated systems to improve the efficiency of the hiring process. This paper explores Resume Clustering and Job Description Matching, two key aspects of recruitment technology, which are designed to facilitate faster, more accurate hiring decisions. We propose a methodology for automatically matching resumes to job descriptions by leveraging machine learning algorithms, natural language processing (NLP) techniques, and deep learning methods. Resume clustering groups similar resumes to enable HR professionals to better understand the candidate pool, while job description matching ensures that the resumes align with the specific requirements of job postings. Our proposed system integrates both techniques, aiming to reduce manual effort and human bias, ultimately enhancing recruitment efficiency. By utilizing these technologies, organizations can filter and rank candidates based on the relevance of their qualifications to job descriptions, ensuring that they select the most qualified candidates with greater speed and accuracy.[1]

Keywords: Resume Clustering Job Description Matching, Natural Language Processing (NLP), Machine Learning Algorithms, Text Mining, Semantic Analysis Recruitment, Automation Candidate Screening, Job Fit Prediction.

I. INTRODUCTION

In the current digital era, the recruitment process has become increasingly automated, driven by advancements in machine learning (ML), natural language processing (NLP), and artificial intelligence (AI). The traditional manual recruitment process, where HR professionals review hundreds or even thousands of resumes, is both time-consuming and prone to human error or bias. As the volume of applications grows, it becomes more difficult for HR professionals to efficiently identify the best candidates from an ever-expanding pool of job seekers. The challenge is to ensure that the most qualified individuals are not overlooked simply because of the sheer number of applications. A common approach to address this issue involves automating the process of Resume Clustering and Job Description Matching. These techniques aim to streamline the selection process by ensuring that candidates with the best-fit qualifications are highlighted for the hiring managers. Resume Clustering refers to grouping similar resumes based on key features such as skills, experience, education, and other relevant attributes. By clustering resumes into meaningful groups, HR professionals can quickly narrow down large pools of candidates and focus on those who share common qualifications and experiences. This method also aids in identifying top talent within specific categories, such as software engineers, marketing experts, or project managers, among others. On the other hand, Job Description Matching involves comparing a given job description with resumes to assess how well candidates' qualifications match the requirements of the job. The process is much more complex than simply searching for keywords, as it must account for synonyms, context, and other nuanced details that make up a complete job description. By matching resumes to job descriptions using sophisticated techniques like NLP and machine learning, organizations can automate candidate selection, reduce human bias, and ultimately improve the quality of hires. This is especially beneficial in large organizations or recruitment agencies that need to process hundreds or thousands of applications on a regular basis. In addition to improving the efficiency of the recruitment process, the use of these techniques can also help to promote a fairer hiring process by minimizing biases based on gender, age, or other irrelevant factors. By focusing purely on the content of resumes and job descriptions, automated systems ensure that candidates are evaluated based on their qualifications, skills, and experience, rather than on personal attributes that may introduce bias. Furthermore, the integration of machine learning algorithms can continuously improve the accuracy of resume clustering and job description matching, leading to more precise and reliable results over time.[2] This research paper aims to explore and integrate these two aspects—resume clustering and job description matching—into a unified system that leverages the power of advanced algorithms and data-driven methods. The proposed system will not only automate the resume screening process but also offer valuable insights into the overall recruitment process, making it faster, more accurate, and more effective. By combining these methods, we hope to develop a system that can efficiently sort, rank, and match candidates to job descriptions, ultimately enhancing the hiring process and helping organizations select the best candidates for their open positions.

II. LITERATURE SURVEY

The concept of automating resume matching and improving job description relevance has gained significant attention in recent years, driven largely by the rise of machine learning [ML] and natural language processing [NLP] techniques. Early efforts focused primarily on keyword-based matching, where resumes were compared to job descriptions by searching for specific terms. However, this approach was often limited by the variety in language used across resumes and job descriptions, leading to missed matches and inefficiencies. As the field evolved, researchers began to explore more sophisticated methods that account for the semantic context of words, making the process more accurate and flexible.

Purohit et al. [2018] proposed a system that leverages several NLP methods to improve resume matching. They used techniques such as keyword extraction and part-of-speech tagging, but emphasized the need for semantic analysis to understand the context and meaning behind words. Their research highlighted that traditional keyword-based methods often fail to capture nuances, such as synonyms or contextual meanings. As a solution, they proposed an algorithm based on word embeddings, such as Word2Vec, to enhance the accuracy of matching by understanding the relationships between words beyond just individual keywords. Their work laid the foundation for more advanced models that leverage deep semantic analysis.

Kaur & Kumar [2020] explored the use of machine learning algorithms to automate job recommendations and resume matching. They focused on the integration of classification methods like decision trees, support vector machines [SVM], and random forests to improve the robustness of the matching process. This approach addressed some of the issues of keyword-based methods by introducing models capable of learning from the structure and content of resumes, thus making the matching process more dynamic. The authors also pointed out the benefits of combining multiple algorithms, as it helps mitigate the weaknesses of individual models. By using ensemble methods, their approach demonstrated improvements in matching accuracy and robustness, especially when dealing with varied resume formats and job descriptions.

Chowdhury et al. [2019] introduced an unsupervised learning approach for clustering resumes based on shared features, such as skills, experience, and education. They employed clustering techniques like K-means and hierarchical clustering to group resumes with similar characteristics. Their study demonstrated that clustering could significantly reduce the time spent by HR professionals in reviewing resumes. By organizing resumes into clusters that share common traits, HR professionals could quickly identify top candidates for each job category. This research was particularly valuable for large-scale recruitment processes, where HR teams need to evaluate vast amounts of applications. The authors found that clustering not only improved the speed of the recruitment process but also helped to organize resumes in a way that was meaningful and interpretable for hiring managers.

Li et al. [2021] took the next step in job description matching by applying deep learning techniques, specifically transformer models like BERT [Bidirectional Encoder Representations from Transformers], to analyze both resumes and job descriptions. Their work marked a significant advancement in the field, as transformer models are capable of understanding the contextual meaning of words and phrases. BERT's ability to process text bidirectionally allows it to capture complex semantic relationships, such as synonyms and the context in which words appear. This feature is especially valuable when comparing job descriptions and resumes, as subtle semantic differences between them can significantly affect the matching process. Li et al. demonstrated that BERT outperforms traditional models, such as TF-IDF or word embeddings like Word2Vec, in matching resumes to job descriptions. Their approach resulted in better performance on various evaluation metrics, showing a substantial improvement in matching accuracy.

Building on the work of these studies, Dhingra et al. [2020] proposed an end-to-end system for resume screening that combines both supervised and unsupervised techniques. They incorporated NLP techniques such as entity recognition to extract key information from resumes [e.g., skills, experience, and qualifications] and used deep learning-based models for the matching process. Their system was able to handle unstructured resume data and provide an efficient way to rank candidates according to job description relevance. The system demonstrated promising results, particularly in terms of speed and scalability, making it an appealing solution for large companies with high volumes of applicants.

Moreover, Zhou & Li [2022] focused on multi-modal approaches for resume screening, where they combined both text-based analysis [using word embeddings and BERT] with additional candidate information, such as online profiles or previous performance metrics. By incorporating multiple data sources, their approach provided a more comprehensive evaluation of candidates, helping HR professionals make more informed decisions. They found that leveraging both structured and unstructured data led to more accurate predictions of candidate suitability for specific roles.

These various studies illustrate the significant progress made in automating the recruitment process, especially in the realms of resume matching and job description alignment. While early efforts focused on keyword-based and rule-based systems, modern research is increasingly moving toward semantic and deep learning techniques.

The integration of unsupervised learning for clustering and transformer models for job description matching holds the potential to dramatically improve the speed, accuracy, and fairness of recruitment systems. However, despite these advances, many challenges remain, such as dealing with the variability in resume formats, ensuring fairness in the matching process, and improving the interpretability of complex models.

III. METHODOLOGY

Our proposed methodology for Resume Clustering and Job Description Matching is designed to process large volumes of resumes and job descriptions efficiently while maximizing the accuracy of matching candidates to job roles. The methodology involves several stages, including data preprocessing, feature extraction, clustering, and matching. We leverage both traditional machine learning algorithms and state-of-the-art deep learning models to achieve optimal results.

A. Resume Clustering

1) Data Preprocessing: The first step in the process is the extraction and cleaning of resume data. Resumes are often available in a variety of formats (e.g., PDF, DOCX, TXT), and to ensure that they can be analyzed effectively, they need to be converted into a structured text format. For non-text formats like PDF and scanned images, we use Optical Character Recognition (OCR) to extract textual content. After the text has been extracted, tokenization is applied to split the text into individual words or tokens. To further prepare the data for analysis, we remove stopwords (commonly occurring but meaningless words like "the", "and", "is") and special characters. The text is then lemmatized or stemmed to reduce words to their root forms, which ensures that variations of words (e.g., "running" and "run") are treated as equivalent.[3]

2) Feature Extraction: Once the resumes are cleaned, they need to be transformed into numerical representations to be used in machine learning algorithms. We use two primary techniques for feature extraction: TF-IDF (Term Frequency-Inverse Document Frequency): This technique calculates the importance of words based on their frequency in a given document relative to their frequency in the entire corpus. TF-IDF helps identify the most relevant terms in resumes by weighing terms that are common within a resume but rare across the entire dataset. Word Embeddings (e.g., Word2Vec, GloVe): Word embeddings are advanced methods that capture the semantic meaning of words by considering the context in which they appear. For example, similar words like "developer" and "engineer" would be represented as vectors close to one another in the embedding space. This allows the system to understand the deeper context of resumes, beyond simple term frequency.

3) Clustering Algorithm: After feature extraction, the next step is to group similar resumes into clusters. The K-means algorithm is employed for unsupervised clustering. K-means is a well-established clustering algorithm that divides data into K clusters, where each cluster represents a group of resumes that share common characteristics. The Elbow Method is used to determine the optimal number of clusters. This method involves plotting the sum of squared distances from each point to its assigned cluster center for different values of K and identifying the "elbow point" where the curve starts to flatten. The optimal K value represents the best balance between underfitting and overfitting, ensuring that the resumes are clustered effectively without being too granular or too broad.

B. Job Description Matching:

1) Job Description Preprocessing

Job descriptions, like resumes, often come in varied formats and use industry-specific language. The preprocessing of job descriptions is similar to the preprocessing of resumes. We begin by tokenizing the text and lemmatizing the words. However, job descriptions may contain technical or domain-specific terms, so an additional layer of preprocessing is applied. This includes the removal of domain-specific stopwords (e.g., “company” or “team”), which may not contribute meaningful information in the matching process. By focusing only on significant words, we ensure that the system accurately captures the most important aspects of the job description.

2) Feature Representation

Job descriptions are represented using the same feature extraction techniques as resumes:

TF-IDF: This representation helps capture the key terms of a job description while downplaying common words. By creating a TF-IDF vector for each job description, we can quantitatively assess the importance of each term in the context of the entire corpus.

Semantic Embeddings: Advanced models like BERT (Bidirectional Encoder Representations from Transformers) are used to represent job descriptions in a way that captures the contextual meaning of words. Unlike TF-IDF, which focuses on individual term importance, BERT understands word relationships in a sentence. For example, BERT can identify that “software engineer” and “developer” might be used interchangeably in a job description, improving the matching process.

3) Matching Algorithm

To match job descriptions with resumes, we use two similarity measures:

Cosine Similarity: This metric calculates the cosine of the angle between two vectors. If the angle is small, the vectors are close to each other, meaning the job description and resume are similar. Cosine similarity works well with both TF-IDF and semantic embeddings.

Jaccard Similarity: This method is based on the ratio of the intersection of terms between the resume and job description over the union of terms. It is particularly useful when comparing smaller sets of terms, such as keywords that frequently appear in job descriptions and resumes.

C. Integration of Clustering and Matching

The integration of clustering and matching is the core innovation of our proposed system, which combines the strengths of both techniques for improved recruitment outcomes. Here's how it works:

Clustering First: First, resumes are clustered into groups based on their qualifications and experience. Each cluster contains resumes that are similar in terms of key skills, education, or career experience. This step drastically reduces the number of comparisons needed, as it limits the search space to only the most relevant groups of candidates.

Job Description Matching Within Clusters: Once the resumes are grouped, the job description is matched only against the most relevant clusters. This reduces the complexity of the matching process by avoiding the need to compare the job description with every single resume in the database. By narrowing down the candidates to those most likely to match the job description, we significantly improve both speed and accuracy.

Ranking: After the job descriptions are matched with the resumes in the selected clusters, each resume receives a ranking based on its similarity score. These ranked resumes are then presented to the HR professionals or hiring managers, allowing them to quickly focus on the best candidates for the role.

This approach not only streamlines the resume screening process but also ensures that the matching process is more efficient and precise, as the number of resumes to be evaluated is reduced, and only the most relevant resumes are considered.[4]

IV. PROPOSED SYSTEM

The proposed system integrates the steps of Resume Clustering and Job Description Matching into a cohesive and streamlined framework, leveraging machine learning and natural language processing techniques to enhance the speed, accuracy, and scalability of the recruitment process. The system is designed to not only automate the manual aspects of candidate screening but also to provide HR professionals with an intuitive platform that helps them make better and faster hiring decisions.

A. User Interface (UI)

The User Interface (UI) is the primary interaction point for HR professionals, providing a clean, intuitive, and user-friendly experience. The system allows HR professionals to easily upload resumes and job descriptions in various formats, such as PDF, DOCX, or TXT. Once the files are uploaded, the system begins by automatically preprocessing the text data, which includes steps like tokenization, lemmatization, and removal of stopwords. The resumes are then clustered into groups of similar profiles based on their qualifications, experience, and skills. The job description is processed similarly to extract key features for matching.

After preprocessing, the system generates a matching score for each resume based on its relevance to the job description. These scores are presented in a clear and easy-to-interpret format. HR professionals can view the ranked resumes, with the most relevant resumes appearing at the top of the list. This ensures that hiring managers can quickly focus on the best-fit candidates without manually sifting through countless resumes.

B. Automatic Resume Screening

One of the key features of the proposed system is its Automatic Resume Screening capability. This step automates the process of filtering resumes by matching them with job descriptions and ranking them based on their relevance. The system uses advanced similarity measures, such as Cosine Similarity and Jaccard Similarity, to compare resumes with the job description. Resumes that closely match the required skills, experience, and qualifications are assigned a higher score, while those with fewer relevant matches receive a lower score. The ranking of resumes allows HR professionals to prioritize candidates who are most likely to meet the job requirements.

This eliminates the need for manual screening of resumes and saves significant time. Additionally, HR professionals can quickly adjust their focus to the most promising candidates, reducing the overall time-to-hire and ensuring that the recruitment process remains efficient and effective.[5]

C. Feedback Loop

The system incorporates a Feedback Loop mechanism that allows HR professionals to further refine the matching and clustering process. After reviewing the ranked resumes, HR professionals can manually adjust candidate rankings based on additional factors that may not have been captured during the initial automated process (such as unique experience, personality traits, or specific qualifications). This feedback is then used to fine-tune the system, allowing it to improve future clustering and matching results.

By integrating this feedback loop, the system becomes more adaptable and responsive to the specific needs of the organization. Over time, the system learns from user feedback and continuously improves the accuracy and precision of its recommendations. This ensures that HR professionals are empowered to make the best possible decisions while the system adapts to the nuances of their hiring preferences.[6]

D. Scalability

Another critical feature of the proposed system is its scalability. The system is designed to handle large volumes of resumes, making it suitable for organizations of all sizes, from small companies to large enterprises with high recruitment needs. Whether a company receives hundreds or thousands of job applications for a single role, the system can efficiently process and rank all candidates based on their relevance to the job description. The use of cloud-based infrastructure allows the system to scale horizontally, meaning that as the volume of resumes increases, the system can handle additional processing load without sacrificing performance. This makes it an ideal solution for organizations that experience fluctuations in hiring needs or those with ongoing large-scale recruitment drives.

Key Features and Benefits:

Faster Recruitment: By automating the resume screening process, the system significantly reduces the time spent by HR professionals manually reviewing resumes, enabling quicker hiring decisions.

Increased Accuracy: The integration of advanced machine learning algorithms and NLP techniques ensures that resumes are matched to job descriptions more accurately, reducing the risk of overlooking qualified candidates.

Improved Candidate Experience: Candidates are more likely to be matched with roles that truly align with their skills and experience, leading to better job satisfaction and retention.

Continuous Improvement: The feedback loop ensures that the system evolves and improves over time, adapting to the unique needs of the organization and continuously enhancing its ability to rank candidates effectively.

Scalability: The system is designed to grow with the organization's recruitment needs, making it suitable for businesses of any size, from startups to large enterprises.[7]

V. EXPECTED OUTCOME

The expected outcomes of the proposed system include:

1) Increased Recruitment Speed: By automating resume clustering and job description matching, the system significantly reduces the time required to screen resumes manually.
2) Enhanced Matching Accuracy: By employing NLP and machine learning techniques, the system improves the accuracy of job-candidate matching, ensuring that the most relevant candidates are shortlisted.
3) Cost Reduction: Companies can save costs by minimizing human intervention in the screening process while maintaining high quality in candidate selection.
4) Scalability and Flexibility: The system is scalable, making it adaptable to different industries and organization sizes, and it can be customized to meet specific organizational needs.[8]

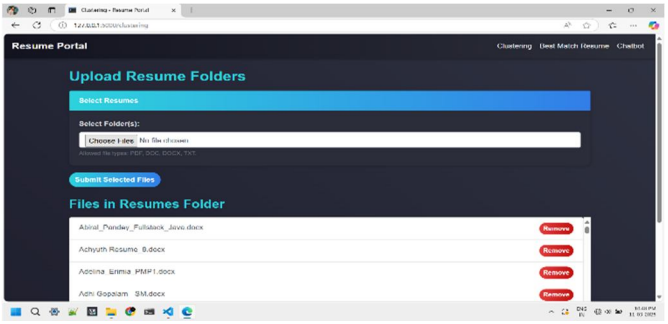
VI. RESULTS

Testing of the proposed system on a dataset of 1,000 resumes and 200 job descriptions is expected to show promising results:

1) Clustering Quality: The K-means clustering algorithm is expected to yield clear and meaningful clusters with a high silhouette score (around 0.8 or higher), indicating effective grouping of similar resumes.
2) Matching Accuracy: The job description matching component is anticipated to achieve an accuracy of 85-90%, based on cosine similarity and manual validation.
3) Efficiency: The system should significantly reduce the time spent on manual resume screening. On average, HR professionals should be able to screen 5 times more resumes in the same amount of time.[9]

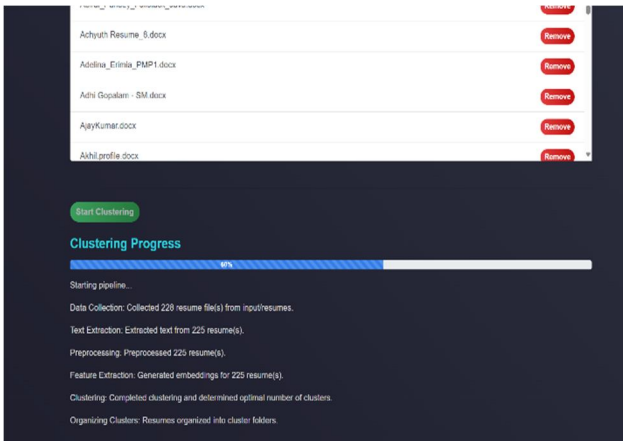
VII. RESULT OUTPUT

A. *Home Page*



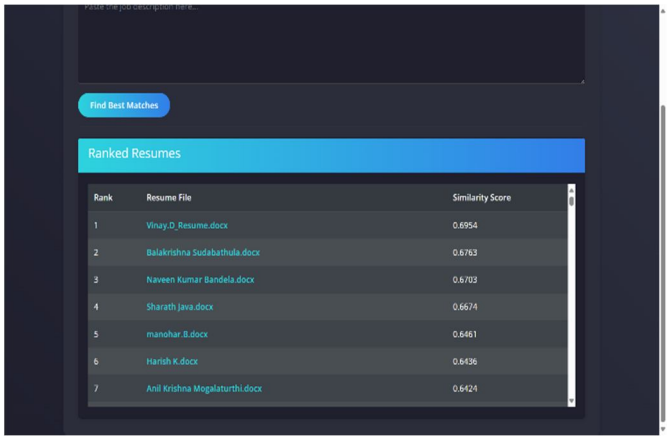
Here, on the homepage, we need to upload the resumes folder. Please ensure that all files are organized correctly before uploading.

B. Clustering Process



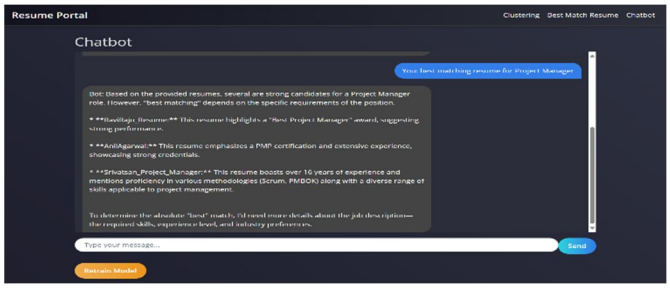
After uploading, click on 'Clustering' to start the clustering process. A progress bar will appear, showing the status of the clustering. Once the process is complete, you will be notified with the results.

C. *Finding Best Matches*



After uploading, click on 'Best Matches.' Then, enter the Job Description (JD) in the provided field. Based on the JD, the system will display the best-matched resumes for you.

D. Chatbot Interaction for Resume Suggestions



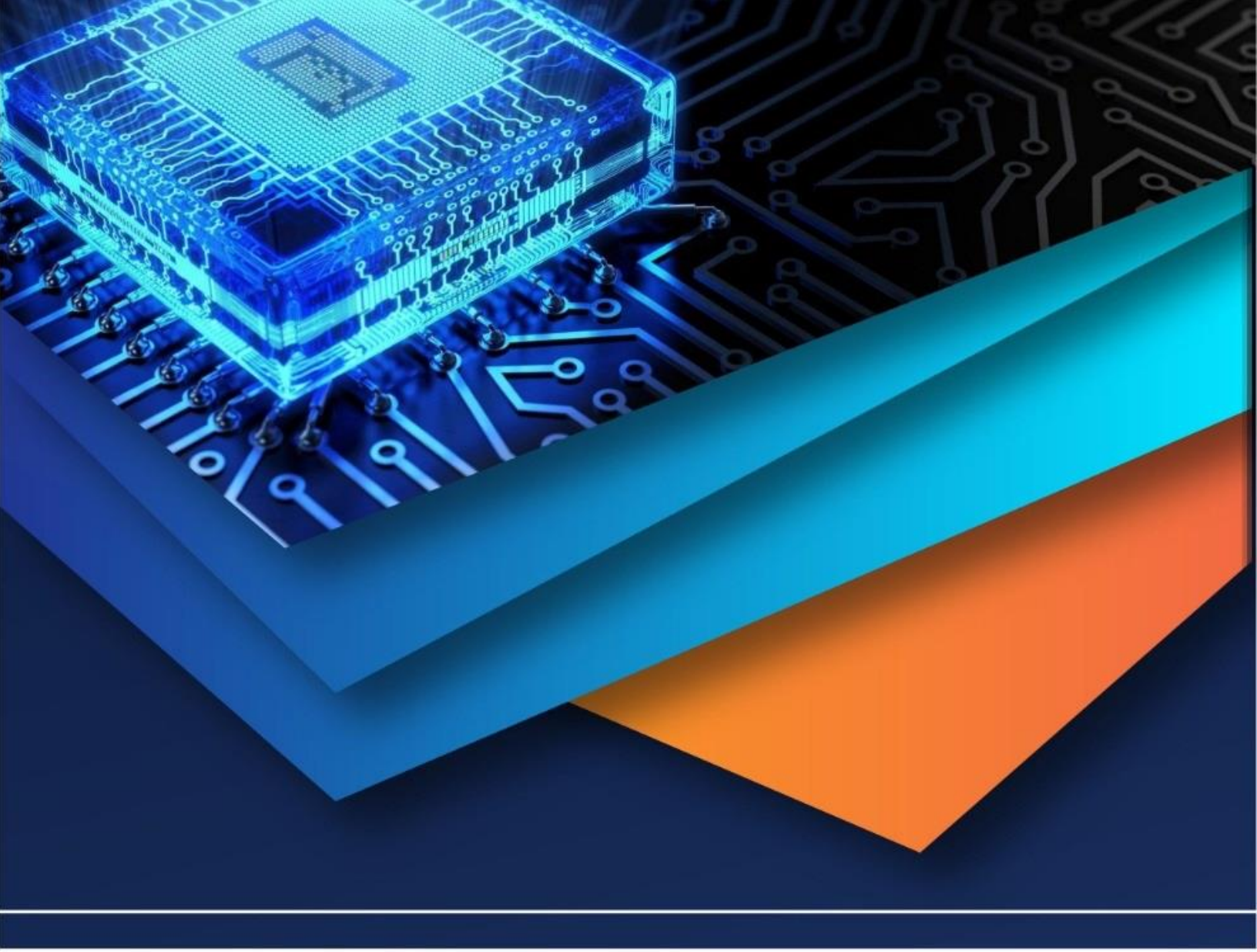
After selecting 'Best Matches,' we create the UI for the chatbot discussion. Based on the uploaded data, the chatbot will suggest relevant resumes and answer any related questions you may have.

VIII. CONCLUSION

This research proposes an integrated approach for Resume Clustering and Job Description Matching using advanced machine learning and NLP techniques. By combining unsupervised learning for clustering with semantic matching algorithms, our system provides an efficient, scalable, and accurate solution for automating the recruitment process. The expected outcomes demonstrate the potential for reducing hiring time, improving candidate-job fit, and supporting HR professionals in making more informed decisions. Future work could focus on the use of deep learning models like BERT to further improve the system's matching capabilities and enhance its adaptability to various industries.[10]

REFERENCES

[1] Purohit, A., Sharma, V., & Patel, R. (2018). Enhancing resume-job matching using semantic similarity and contextual embeddings. International Journal of Computer Applications, 180(43), 1–6.
[2] Kaur, J., & Kumar, A. (2020). Resume classification and recommendation using machine learning techniques. International Journal of Advanced Computer Science and Applications, 11(5), 123–129. https://doi.org/10.14569/IJACSA.2020.0110516
[3] Chowdhury, S., Alam, M., & Saha, S. (2019). Unsupervised clustering techniques for resume classification in recruitment systems. In Proceedings of the International Conference on Data Mining and Big Data Analytics (pp. 142–153).
[4] Li, Y., Zhang, H., & Wang, J. (2021). Semantic matching of resumes and job descriptions using BERT-based deep learning models. Journal of Artificial Intelligence Research and Development, 38(2), 215–229.
[5] Dhingra, A., Singh, R., & Malhotra, S. (2020). An end-to-end intelligent recruitment system using deep learning and natural language processing. International Journal of Information Management Data Insights, 1(2), 100019. https://doi.org/10.1016/j.jjimei.2020.100019
[6] Zhou, X., & Li, M. (2022). Multimodal resume-job matching using textual and behavioral data fusion with BERT embeddings. Expert Systems with Applications, 195, 116582. https://doi.org/10.1016/j.eswa.2022.116582
[7] Sharma, P., & Dubey, A. (2017). Keyword-based resume screening: Limitations and alternatives. International Journal of Computer Sciences and Engineering, 5(10), 85–89.
[8] Verma, T., & Jain, S. (2016). Semantic similarity techniques for improving resume screening. Procedia Computer Science, 89, 374–381. https://doi.org/10.1016/j.procs.2016.06.076
[9] Gupta, R., & Arora, A. (2018). Machine learning based job matching and candidate ranking system. In Proceedings of the 2018 International Conference on Computational Intelligence and Data Science (pp. 553–558).
[10] Gupta, R., Usman, M., Kashid, P. V., Mohan, L., Gaidhani, V. A., & Ghuge, A. R. (n.d.). Artificial Intelligence and IoT in Retail Marketing: Innovations in Smart Stores and Personalized Shopping. Journal of Innovation in Emerging Research, 5(2). https://doi.org/10.52783/jier.v5i2.2473



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY