iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81432

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 14 Issue IV Apr 2026- Available at www.ijraset.com

AI-Powered Resume Screening System Using Natural Language Processing (NLP) and Machine Learning (ML)

Navya Sri Bitra[1], K.Sudeepa Kumari[2], Sri Manikanta Ramisetty[3], Ravi Shanker Jonna[4]
Department of Computer Science and Engineering Acharya Nagarjuna University Guntur, Andhra Pradesh,

Abstract**: The rapid increase in digital job applications has created substantial challenges for recruiters in efficiently identifying suitable candidates from large volumes of resumes. Recruiters frequently process large applicant pools under strict hiring deadlines, making manual resume evaluation slower and less consistent in large-scale recruitment. This work presents an AI-Powered Resume Screening System that automates candidate evaluation by integrating Natural Language Processing (NLP), Machine Learning (ML), and intelligent assessment modules within a unified recruitment framework. The system performs resume parsing, text preprocessing, skill extraction, and feature representation using TF-IDF and semantic similarity techniques to compare candidate profiles with job descriptions. Logistic Regression serves as the primary classification model for predicting candidate relevance and generating ranking scores. To enhance recruitment effectiveness, the framework also incorporates GPT-based interview question generation, chatbot assistance, gamified cognitive skill assessment, and real-time analytics dashboards. MongoDB is utilized for scalable storage of resumes, user activity, screening history, and performance metrics. Experimental analysis indicates that the proposed framework improves screening speed, classification accuracy, and recruiter decision-making while reducing repetitive manual effort. By combining automated screening, intelligent interaction, and scalable architecture, the system provides a practical, adaptive, and comprehensive solution for modern recruitment environments.

Keywords: Automated Resume Screening, Natural Language Processing, Machine Learning Classification, Candidate Ranking, Semantic Analysis, Intelligent Recruitment System, TF-IDF Feature Extraction, Logistic Regression, GPT-Based Interview Module, Recruitment Analytics.

I. INTRODUCTION

The widespread adoption of digital recruitment platforms has significantly increased the number of applications received for individual job openings, creating new challenges in candidate screening and selection. Organizations across technical and non-technical sectors often process hundreds of resumes for a single position, making manual evaluation increasingly inefficient in modern hiring environments. Traditional screening practices require recruiters to spend substantial time reviewing resumes individually, which can slow hiring cycles and introduce inconsistencies due to repetitive workload, subjective interpretation, or limited contextual analysis. In addition, many conventional Applicant Tracking Systems (ATS) primarily rely on keyword-based matching, which may fail to recognize semantically relevant qualifications when resume terminology differs from job descriptions [5].

To address these limitations, this research proposes an intelligent recruitment framework that integrates Natural Language Processing (NLP) [9], Machine Learning (ML) [7], semantic similarity analysis [10], and intelligent automation to improve recruitment effectiveness. The system is designed to process resumes through structured stages including text extraction, preprocessing, tokenization, stop-word removal, and TF-IDF-based feature representation [9]. These processed features are then evaluated using Logistic Regression [7] to classify candidate relevance according to job-specific requirements. Furthermore, cosine similarity [10] is incorporated to measure contextual alignment between resume content and job descriptions, enabling the system to move beyond direct keyword dependency and improve semantic candidate matching.

The proposed framework extends beyond resume shortlisting by integrating advanced recruitment intelligence modules. A GPT-based interview generation component dynamically produces technical, skill-based, and project-oriented interview questions according to candidate profiles [3], [4], while a gamified cognitive assessment module evaluates logical reasoning and decision-making capability. In addition, real-time dashboards provide recruiters with analytical insights into candidate rankings, system performance, and activity trends [14].

©IJRASET: All Rights are Reserved

MongoDB is utilized as a scalable NoSQL database for storing resumes, screening outcomes, interview history, and user activity logs [8]. By combining automated screening, contextual candidate analysis, adaptive interview support, and scalable data management, the proposed system provides a unified and intelligent recruitment solution capable of improving hiring speed, screening precision, and operational efficiency.

A. Need for Automated Resume Screening

The rapid growth of online employment opportunities has made conventional resume screening methods increasingly insufficient for large-scale recruitment operations. Recruiters are frequently expected to evaluate large applicant pools within limited timeframes, creating practical challenges in maintaining speed, fairness, and consistency. Manual screening can become repetitive and resource-intensive, particularly when candidate resumes vary in formatting, vocabulary, and skill representation. In such scenarios, qualified applicants may be overlooked when their resumes do not exactly match predefined keywords despite possessing suitable competencies.

Automated resume screening systems provide a practical solution by applying NLP [9] and ML [7] techniques to evaluate resumes more systematically and objectively. These systems can process large volumes of applications efficiently, identify important candidate attributes, rank profiles according to relevance, and support data-driven hiring decisions [5]. Through semantic analysis [10] and intelligent classification, automated frameworks improve scalability while reducing recruiter burden and enhancing candidate selection quality.

B. Identified Research Gaps

Although AI-based recruitment systems have introduced automation into hiring workflows, several important research limitations remain. Many existing systems continue to depend heavily on keyword-matching strategies [5], which restrict their ability to evaluate contextual relevance accurately. Some machine learningbased screening solutions improve classification [7] but do not integrate semantic similarity mechanisms [10] for deeper resume-job understanding. In addition, most current systems focus primarily on filtering and ranking while lacking advanced capabilities such as AI-driven interview generation [3], [4], candidate interaction support, cognitive skill evaluation, and integrated performance monitoring [14].

Another major gap is the absence of unified recruitment architectures that combine resume parsing, semantic screening, candidate ranking, interview intelligence, gamified assessment, dashboard analytics, and scalable historical tracking within a single framework. Many existing platforms operate as isolated solutions, forcing recruiters to depend on multiple disconnected systems.

This research addresses these gaps by proposing a comprehensive AI-Powered Recruitment Ecosystem that combines NLP-based semantic analysis, Logistic Regression-based classification, GPT-assisted interview generation, gamified evaluation, real-time analytics, and MongoDB-supported scalable data management into one integrated platform. Such an approach enhances recruitment intelligence, improves system scalability, and provides a more practical solution for modern hiring challenges.

II. LITERATURE REVIEW

The use of Artificial Intelligence (AI) in recruitment systems has expanded significantly as organizations seek faster and more reliable methods for candidate screening [5]. Initial digital recruitment solutions primarily focused on automating resume filtering through keyword-based approaches and predefined selection rules. These systems reduced manual workload to some extent, but their effectiveness was limited [5] because candidate evaluation depended largely on exact keyword presence rather than contextual understanding. As a result, resumes containing relevant skills expressed in alternative forms were often misclassified or excluded, reducing the overall efficiency of candidate selection.

To improve recruitment accuracy, researchers introduced Machine Learning (ML) techniques for automated resume classification and candidate prediction. Algorithms such as Logistic Regression, Naïve Bayes, Support Vector Machines (SVM), and Random Forest have been widely applied [7] to classify resumes based on extracted candidate features including education, technical skills, certifications, and work experience. These models demonstrated improved screening performance compared to traditional filtering systems because they learned predictive relationships from training datasets. However, many existing ML-based recruitment systems primarily emphasize classification efficiency while offering limited semantic interpretation of resume content, which restricts their effectiveness when job descriptions and candidate profiles use varied terminology.

Natural Language Processing (NLP) has played a critical role in addressing these limitations by enabling more advanced textual understanding in recruitment applications.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Techniques such as tokenization, stop-word removal, stemming, Named Entity Recognition (NER), and TF-IDF feature extraction [9], [13] have improved the ability to process unstructured resume data systematically. In addition, semantic similarity techniques such as cosine similarity [10] have been utilized to strengthen resumejob description alignment by evaluating contextual relevance rather than relying solely on keyword overlap. These developments have significantly enhanced candidate ranking quality and improved the identification of suitable applicants [10].

Recent advancements in AI have further extended recruitment systems beyond resume screening by incorporating intelligent interaction models. GPT-based technologies and large language models have introduced capabilities [3], [4] such as automated interview question generation, chatbot-assisted candidate support, and adaptive communication frameworks. These technologies improve recruiter efficiency by supporting dynamic candidate engagement [3], [4] and personalized technical assessment. Despite these advancements, many currently available solutions function as isolated tools that focus on specific recruitment stages rather than providing an integrated end-to-end recruitment architecture.

Scalability and data management have also become increasingly important in recruitment research. Modern systems frequently utilize NoSQL databases such as MongoDB [8] to manage resumes, recruiter activity, candidate history, and evaluation records efficiently [8]. Although scalable storage improves operational flexibility, many systems still lack unified integration of semantic resume analysis, intelligent interviews, gamified assessment, and recruiter analytics within a centralized framework.

Based on existing literature, it is evident that AI, ML, and

NLP technologies have substantially improved recruitment automation [5], [7], [9]; however, challenges remain in semantic precision, workflow integration, recruiter decision support, and system scalability. This research addresses these limitations by proposing a unified automated hiring platform that integrates NLPbased preprocessing, TF-IDF feature extraction, Logistic Regression classification, semantic similarity analysis, GPT-assisted interview generation, gamified cognitive assessment, real-time analytics dashboards, and MongoDB-based data management. By combining these components within a single framework, the proposed system aims to provide a more comprehensive, scalable, and intelligent recruitment solution.

III. PROBLEM STATEMENT

The increasing dependence on online recruitment systems has significantly raised the number of job applications submitted for available positions, making candidate screening more complex for recruiters and hiring teams. In many organizations, a single job posting can attract hundreds of resumes, creating substantial pressure on recruitment departments to identify qualified applicants quickly and accurately. Manual resume screening under such conditions is often slow, repetitive, and inconsistent, particularly when recruiters must evaluate diverse candidate profiles within limited hiring deadlines. This process may also introduce human bias, subjective decision-making, and the possibility of overlooking capable candidates [5].

Conventional recruitment systems and many Applicant Tracking Systems (ATS) primarily rely on keyword-based filtering mechanisms [5] to shortlist candidates. Although these methods provide basic automation, they often fail to evaluate semantic relationships between candidate qualifications and job requirements effectively [10]. As a result, applicants with relevant competencies may be excluded simply because their resumes do not contain exact predefined keywords. Such limitations reduce screening precision and restrict the system's ability to identify contextually suitable candidates across varied resume formats and vocabulary styles.

Several AI-based recruitment models have introduced Machine Learning (ML) [7] and Natural Language Processing (NLP) [9] techniques to improve automation; however, many existing solutions remain limited in scope. Some systems focus only on resume classification without incorporating semantic similarity for deeper candidate-job alignment, while others provide filtering capabilities without advanced modules such as interview intelligence [3], [4], candidate interaction, skill-based cognitive assessment, or recruiter analytics [11], [14]. In many cases, recruiters must still depend on multiple disconnected tools to complete the hiring workflow, which reduces efficiency and increases system fragmentation.

Another critical challenge is scalability. Modern recruitment environments require systems capable of handling large-scale resume databases, user interactions, ranking histories, and real-time analytical insights simultaneously. Traditional frameworks may struggle to provide integrated storage, adaptive assessment, and longterm tracking within a unified architecture [8].

Therefore, there is a strong need for a comprehensive AIdriven recruitment framework that can automate resume screening while improving contextual accuracy, operational scalability, and decision-making quality. Such a system should combine NLP-based preprocessing [9], semantic similarity analysis [10], ML-driven candidate classification [7], GPT-assisted interview generation [3], [4], gamified cognitive evaluation, real-time recruiter dashboards, and scalable database management [8] into a single integrated platform.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Developing such a unified solution can significantly reduce recruiter workload, improve candidate selection precision, and provide a more intelligent and efficient approach to modern digital recruitment.

IV. RELATED WORK

A. Traditional Resume Screening Methods

Early recruitment systems primarily depended on manual resume evaluation and keyword-based Applicant Tracking Systems (ATS) to shortlist candidates [5]. In these approaches, recruiters or automated platforms screened resumes by matching predefined job-related terms with candidate documents. Although such systems reduced some level of manual workload, their effectiveness was limited because they relied heavily on direct keyword occurrence rather than contextual relevance. As a result, candidates with appropriate qualifications were sometimes overlooked when their resumes used alternate terminology or non-standard skill descriptions. These limitations highlighted the need for more adaptive and intelligent screening solutions.

B. Machine Learning-Based Recruitment Systems

To improve recruitment efficiency, Machine Learning (ML) techniques were introduced for candidate classification and prediction. Algorithms such as Logistic Regression, Naïve Bayes, Support Vector Machines (SVM) [7], Decision Trees, and Random Forest have been widely explored for resume screening applications [7]. These systems classify candidates by learning patterns from labeled recruitment datasets that include technical skills, education, certifications, and experience-related attributes. ML-based systems improved automation by reducing dependency on manual filtering and increasing predictive consistency. However, many such systems focused primarily on classification performance and often lacked deeper semantic understanding of resume content, which restricted their adaptability when candidate profiles varied significantly in vocabulary or structure.

C. NLP-Based Resume Analysis

Natural Language Processing (NLP) significantly advanced resume screening by enabling systems to interpret unstructured textual resumes more effectively. Techniques such as tokenization, stop-word removal, stemming, Named Entity Recognition (NER), and TF-IDF feature extraction [9] have been widely applied to extract meaningful information from resumes. These methods improved the ability to identify candidate skills, qualifications, and technical competencies systematically. In addition, semantic similarity techniques such as cosine similarity [10] enhanced candidate-job matching by evaluating contextual alignment between resumes and job descriptions. NLP-based systems therefore improved screening precision beyond traditional keyword filtering by supporting more context-aware candidate evaluation.

D. AI-Driven Interview and Intelligent Recruitment Systems

Recent advancements in Artificial Intelligence have expanded recruitment technologies beyond screening into intelligent candidate interaction and adaptive assessment. GPT-based systems and advanced language models [3], [4] have introduced capabilities such as automated interview question generation, chatbot-assisted communication, and personalized technical assessments. These systems improve recruiter productivity by automating interview preparation and supporting candidate engagement. Despite these innovations, many existing solutions remain focused on isolated functionalities such as interview generation or resume parsing rather than providing a fully integrated recruitment ecosystem.

E. Limitations of Existing Systems

Although substantial progress has been made in recruitment automation, several limitations remain in current systems. Many existing platforms still emphasize isolated functionalities, including standalone resume filtering, classification, or interview assistance, without combining all recruitment stages into one framework.

Semantic precision may also remain limited in systems that rely primarily on classification without contextual ranking. In addition, several solutions lack gamified cognitive evaluation, real-time recruiter dashboards [14], scalable historical tracking, and unified database integration. These gaps reduce operational efficiency and often require recruiters to depend on multiple disconnected tools.

The proposed system addresses these limitations by integrating NLP-based preprocessing, TF-IDF feature extraction, Logistic Regression classification, semantic similarity analysis, GPT-assisted interview generation, gamified cognitive skill evaluation, real-time analytics dashboards, and MongoDB-supported scalable data management within a single architecture. This unified approach aims to improve recruitment precision, recruiter efficiency, and system scalability while providing a more comprehensive solution for modern hiring environments.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

V. PROPOSED WORK

A. System Architecture

The proposed intelligent recruitment architecture is designed as a comprehensive recruitment framework that integrates resume analysis, candidate ranking, intelligent interview generation, cognitive assessment, and recruiter analytics within a unified architecture. The system combines Natural Language Processing (NLP), Machine Learning (ML), semantic similarity analysis, GPT-assisted intelligence, and scalable data management to automate and optimize modern hiring processes.

The architectural workflow begins with resume and job description collection through the system interface. Candidates upload resumes in standard formats such as PDF, DOC, or DOCX, while recruiters provide job requirements and role-specific descriptions. This information is processed through structured modules that extract, analyze, and evaluate candidate suitability. The architecture supports complete recruitment automation by linking resume parsing, preprocessing, feature engineering, classification, interview generation, and final recruiter decision-making into one integrated ecosystem.

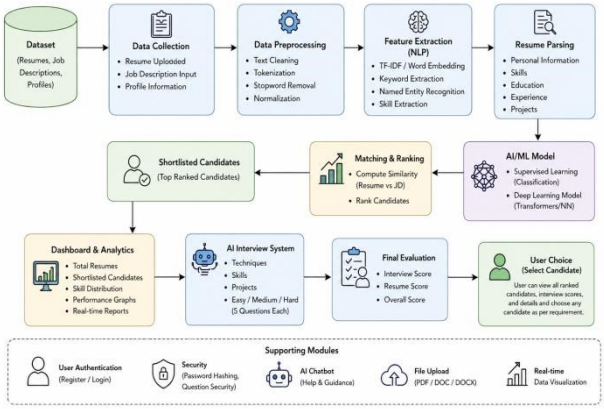


Fig. 1. System Architecture of the Proposed AI-Powered Resume Screening System

B. Data Collection and Preprocessing

The system collects candidate resumes, profile details, and job descriptions through an interactive recruitment platform. Uploaded resumes are processed using a resume parser that extracts important information such as educational background, technical skills, certifications, projects, and work experience. This extracted textual data serves as the foundational input for candidate evaluation.

To improve data quality, preprocessing operations are applied to the extracted content. These operations include text cleaning to remove irrelevant symbols and formatting noise, normalization to standardize textual data, tokenization to divide text into meaningful units, and stopword removal [9], [13] to eliminate analytically insignificant terms. Additional techniques such as stemming or lemmatization may also be applied to improve textual consistency. These preprocessing steps ensure that candidate resumes are converted into structured analytical content suitable for feature extraction and predictive modeling.

C. Feature Extraction

Feature extraction is performed using TF-IDF [9] (Term Frequency-Inverse Document Frequency), which transforms processed textual resumes into weighted numerical vectors representing the importance of candidate skills, qualifications, and domain-specific terms. TF-IDF allows the system to quantify textual resume content for Machine Learning analysis.

In addition to TF-IDF, the framework incorporates keyword extraction, Named Entity Recognition (NER) [13], and skill extraction to identify critical candidate features such as programming languages, certifications, educational credentials, and technical expertise. Semantic similarity analysis using cosine similarity [10] compares candidate resumes with job descriptions contextually, enabling deeper candidate-job alignment beyond direct keyword overlap. These combined methods improve the system's ability to evaluate candidate suitability accurately.

D. Candidate Classification and Ranking

The proposed system utilizes Logistic Regression [7] as the primary Machine Learning algorithm for candidate classification. Based on extracted features, the model predicts whether a candidate profile aligns with jobspecific requirements. Logistic Regression is selected because of its computational efficiency, interpretability, and effectiveness in high-dimensional text classification tasks.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

After classification, semantic similarity scores and predictive outputs are combined within the Matching and Ranking Module to rank candidates according to contextual relevance and predicted suitability. This ranking mechanism enables recruiters to prioritize applicants systematically while reducing repetitive manual effort. Candidates with the highest scores are shortlisted for advanced evaluation.

E. AI-Based Interview and Cognitive Assessment

To enhance recruitment intelligence beyond resume screening, the system incorporates a GPT-based interview generation module [3], [4] that dynamically produces technical, skill-based, and project-oriented interview questions according to candidate profiles. This adaptive interview system improves technical evaluation quality while reducing recruiter effort in interview preparation.

Additionally, a gamified cognitive assessment module is integrated to evaluate candidate reasoning ability, logical thinking, and decision-making performance. Through interactive assessments, the system collects supplementary performance metrics that contribute to more holistic candidate evaluation. These modules extend the system from basic resume filtering to multi-dimensional recruitment intelligence.

F. Dashboard, Data Storage, and Analytics

MongoDB [8] is implemented as the primary NoSQL database for storing resumes, candidate profiles, ranking histories, interview responses, cognitive assessment outcomes, and recruiter activity logs. Its scalable architecture supports efficient storage and retrieval of both structured and unstructured recruitment data.

A real-time dashboard and analytics module [14] provides recruiters with visual insights including total resumes processed, shortlisted candidates, candidate ranking statistics, skill distributions, interview performance, and system efficiency metrics. This analytical layer supports data-driven decision-making and improves recruiter transparency throughout the hiring process.

Overall, the proposed framework combines NLP preprocessing, TF-IDF feature engineering, Logistic Regression classification, semantic similarity analysis, GPT-assisted interview intelligence, gamified cognitive evaluation, MongoDB-supported storage, and real-time analytics into a complete recruitment ecosystem. This integrated architecture improves candidate screening precision, operational scalability, recruiter productivity, and overall hiring effectiveness.

VI. IMPLEMENTATION

A. Development Environment

The implementation of the proposed AI-Powered Resume Screening System is developed using an integrated technological environment that supports Natural Language Processing (NLP), Machine Learning (ML), web interaction, and scalable data management. Python is used as the primary programming language because of its flexibility and strong ecosystem for AI-based development. Machine Learning functionalities are implemented using Scikit-learn [7], while NLP tasks such as tokenization, preprocessing, and skill extraction are supported through libraries such as NLTK or spaCy [9], [13]. Resume parsing for PDF, DOC, and DOCX files is handled using document processing tools capable of extracting structured textual information from multiple file formats. For user interaction, the system interface is developed using web technologies such as HTML, CSS, and JavaScript, enabling accessible communication between candidates and recruiters. MongoDB is used as the backend NoSQL database [8] to manage resumes, user profiles, ranking history, interview outputs, and system activity records.

B. Data Acquisition and Resume Parsing

The implementation process begins with candidate data acquisition through the resume upload module. Applicants submit resumes through the platform along with profile details and job-specific information. Uploaded files are processed by the resume parser, which extracts meaningful textual content from resumes, including personal information, education, certifications, technical skills, projects, and work experience.

This extracted information is then transformed into structured candidate profiles. By standardizing resumes from multiple document formats into a unified textual representation, the system ensures consistent processing regardless of resume design or formatting style. This stage serves as the foundation for subsequent analytical operations.

C. Data Preprocessing

After resume parsing, the extracted text undergoes preprocessing to improve data quality and analytical reliability. Text cleaning removes unnecessary symbols, hyperlinks, duplicate formatting patterns, and irrelevant characters.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Tokenization divides textual content into meaningful lexical units, while stop-word removal eliminates commonly occurring but analytically insignificant words. Normalization is applied to standardize textual content, ensuring consistency across resumes. Where necessary, stemming or lemmatization [9], [13] techniques are used to reduce vocabulary variation and improve text uniformity.

These preprocessing operations convert raw resume data into a refined and structured format that is more suitable for feature engineering and predictive analysis.

D. Feature Extraction and Semantic Analysis

Feature extraction is implemented using TF-IDF [9] (Term Frequency-Inverse Document Frequency), which converts preprocessed textual data into weighted numerical vectors. This representation identifies the importance of candidate-specific skills, technical expertise, and domain-relevant qualifications while reducing the influence of common terms. TF-IDF enables efficient computational analysis of resume content for classification tasks.

To strengthen contextual candidate-job matching, semantic similarity analysis is incorporated using cosine similarity [10]. This technique compares candidate resumes with job descriptions by measuring contextual alignment between vectorized representations. Unlike traditional keyword-only methods, semantic similarity improves screening precision by recognizing relevant profiles even when candidate terminology differs from recruiter-defined language.

E. Model Training and Candidate Classification

Candidate suitability prediction is implemented using Logistic Regression [7] as the primary supervised Machine Learning model. The algorithm is trained on labeled recruitment datasets in which resumes are categorized according to relevance for specific job roles. Training and testing datasets are separated to evaluate predictive reliability objectively.

After training, the model analyzes new candidate resumes using extracted features and predicts relevance scores. These predictive outputs are combined with semantic similarity measurements to generate ranked candidate lists. Candidates are then prioritized according to classification probability and contextual suitability. Performance is evaluated using Accuracy, Precision, Recall, and F1-score to ensure effective and balanced screening quality.

F. AI Interview and Cognitive Assessment Integration

To extend implementation beyond resume classification, the system integrates a GPT-based interview generation [3], [4] module that dynamically creates personalized interview questions based on candidate skills, project experience, and technical background. This module supports adaptive candidate evaluation by generating technical, practical, and project-oriented questions across multiple difficulty levels.

In addition, a gamified cognitive assessment module is implemented to measure candidate reasoning ability, analytical thinking, and decision-making performance through interactive tasks. Results from interview and cognitive evaluations are incorporated into the final candidate assessment process, providing a more comprehensive recruitment analysis.

G. Database Management and Dashboard Implementation

MongoDB is implemented as the core database infrastructure [8] for storing resumes, parsed candidate profiles, ranking histories, interview records, cognitive assessment outcomes, and recruiter interactions. Its NoSQL architecture provides scalability, flexibility, and efficient handling of large recruitment datasets containing both structured and unstructured data.

A real-time dashboard and analytics interface [14] is also implemented to provide recruiters with centralized monitoring capabilities. This dashboard displays metrics such as total resumes processed, shortlisted candidates, skill distributions, interview scores, system performance trends, and final evaluation reports. These analytical capabilities improve recruiter decision-making by enabling transparent and data-driven hiring workflows.

Overall, the implementation integrates resume acquisition, NLP preprocessing, TF-IDF feature extraction, semantic similarity analysis, Logistic Regression classification, GPT-assisted interview intelligence, cognitive assessment, MongoDB data management, and dashboard analytics into a unified and scalable recruitment system. This implementation strategy enhances automation, improves candidate selection precision, and provides a practical solution for modern digital hiring environments.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

VII. ALGORITHMS USED

A. Logistic Regression

Logistic Regression is implemented as the primary supervised Machine Learning algorithm [7] for candidate classification within the proposed AI-Powered Resume Screening System. The objective of this algorithm is to predict the probability that a candidate resume matches job-specific requirements based on extracted textual and semantic features. Logistic Regression is particularly effective for classification tasks involving highdimensional textual data because it provides computational efficiency, interpretability, and reliable predictive performance.

In the proposed framework, TF-IDF-generated feature vectors are used as input variables, and Logistic Regression predicts candidate relevance using the sigmoid function:

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}} \quad (1)$$

where (P(Y=1|X)) represents the probability that a candidate is suitable for a given role, (X_1...X_n) represent extracted resume features, and (\beta_0...\beta_n) are model coefficients learned during training.

This model is selected because it performs efficiently in sparse text-based feature environments and supports scalable recruitment classification.

B. TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF [9] is used to convert unstructured textual resumes into weighted numerical vectors by measuring the importance of terms relative to their occurrence within individual resumes and across the overall dataset. This technique strengthens feature quality by emphasizing candidate-specific skills and reducing the influence of commonly repeated terms.

The TF-IDF formula is expressed as:

$$TF\text{-}IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right) \quad (2)$$

where (TF(t,d)) is the frequency of term (t) in document (d), (N) is the total number of documents, and (DF(t)) is the number of documents containing term (t).

Within the system, TF-IDF improves candidate representation by assigning greater weight to meaningful technical skills, certifications, and domain-relevant qualifications.

C. Semantic Similarity Using Cosine Similarity

Cosine similarity [10] is implemented to measure contextual alignment between candidate resumes and job descriptions. Unlike traditional keyword-only filtering, cosine similarity evaluates the directional similarity between vectorized textual representations, allowing the system to identify relevant candidates even when different terminology is used.

The cosine similarity formula is:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \quad (3)$$

where (A) represents the resume feature vector and (B) represents the job description vector.

This technique improves candidate-job matching precision by reducing false negatives caused by vocabulary variation.

D. Named Entity Recognition (NER)

Named Entity Recognition (NER) [13] is applied within the NLP pipeline to extract structured entities such as educational qualifications, institution names, certifications, technical tools, programming languages, and professional roles from resumes. NER strengthens the system's ability to organize candidate information into meaningful categories.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Although NER does not rely on a single universal formula like classification models, its role in the proposed system is to improve structured feature extraction, thereby supporting more accurate classification and candidate ranking.

E. GPT-Based Interview Question Generation

The GPT-based interview module [3], [4] is integrated to dynamically generate personalized interview questions according to candidate skills, projects, and technical expertise. This module analyzes extracted candidate features and produces adaptive technical questions that improve interview quality and recruiter efficiency.

This component extends the recruitment framework from resume screening to intelligent candidate evaluation by automating technical assessment preparation.

F. MongoDB for Scalable Data Management

MongoDB functions as the primary NoSQL data management system [8] within the framework. It stores resumes, extracted features, candidate rankings, interview outputs, gamified cognitive assessment results, and recruiter activity logs. Its flexible architecture supports large-scale recruitment environments by efficiently handling structured and unstructured candidate data. *G. Performance Evaluation Metrics*

To validate algorithm effectiveness, the proposed system uses Accuracy, Precision, Recall, and F1-score [7] as performance evaluation metrics.

Accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Precision is defined as:

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Recall is calculated as:

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

F1-score is expressed as:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

where (TP), (TN), (FP), and (FN) represent True Positive, True Negative, False Positive, and False Negative values respectively.

Overall, the proposed system combines Logistic Regression, TF-IDF, cosine similarity, NER, GPT-based interview intelligence, and MongoDB-supported data management into a unified recruitment framework. The integration of these algorithms improves automation depth, contextual precision, recruiter support, and overall hiring effectiveness.

VIII. COMPARING METHODS

To determine the most effective Machine Learning approach for automated resume screening, multiple classification algorithms were comparatively evaluated [7] based on predictive performance, computational efficiency, interpretability, scalability, and suitability for text-based recruitment tasks. The selected methods include Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Random Forest, and Bagging. Each algorithm was studied to identify the most practical model for the proposed AI-Powered Resume Screening System.

A. Logistic Regression

Type: Supervised Learning (Classification)

Description:
Logistic Regression is employed as the primary classification algorithm [7] in the proposed system to predict candidate suitability based on extracted resume features such as technical skills, educational qualifications, certifications, and professional experience. It estimates the probability of candidate relevance using structured numerical representations derived from textual data. Logistic Regression performs effectively in sparse and high-dimensional text classification environments, making it highly suitable for scalable resume screening frameworks that require efficiency, interpretability, and deployment simplicity.
Advantages:
1) Simple and computationally efficient implementation
2) Performs effectively in sparse and highdimensional textual environments
3) Fast model training and prediction
4) Produces interpretable outputs for recruiter analysis
5) Highly suitable for binary relevance classification

Limitations:
1) Assumes linear relationships between input features and output
2) Limited ability to model highly complex nonlinear patterns

B. Naïve Bayes
Type: Probabilistic Classifier
Description:
Naïve Bayes is a probabilistic Machine Learning algorithm [7] based on Bayes' Theorem that predicts candidate suitability by assuming conditional independence among extracted features. It is widely used in text classification because of its speed, simplicity, and computational efficiency. Within resume screening, Naïve Bayes can classify candidate profiles efficiently; however, its simplifying assumptions may reduce predictive precision when candidate attributes are strongly interdependent.
Advantages:
1) Fast and lightweight computational performance
2) Effective for basic text classification tasks
3) Suitable for smaller datasets
4) Easy to implement and deploy
Limitations:
1) Assumes feature independence
2) Lower contextual precision for complex recruitment datasets
3) Less effective with correlated resume attributes

C. Support Vector Machine (SVM)
Type: Supervised Learning (Classification)
Description:
Support Vector Machine (SVM) is a classification technique [7] that identifies an optimal hyperplane to separate candidate categories effectively. It performs strongly in high-dimensional feature spaces and often provides high predictive capability for structured text classification tasks. SVM is particularly effective when classification boundaries are clearly distinguishable; however, computational complexity and parameter optimization requirements may reduce implementation efficiency in large-scale recruitment platforms.
Advantages:
1) Strong predictive capability
2) Effective in high-dimensional feature spaces
3) Handles complex classification boundaries
4) Suitable for advanced classification tasks
Limitations:
1) Higher computational complexity
2) Parameter optimization can be challenging

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

3) Lower interpretability compared to simpler models

D. Random Forest

Type: Ensemble Learning

Description:

Random Forest is an ensemble classification algorithm [7] that combines multiple decision trees to improve predictive accuracy and reduce overfitting. It can model nonlinear feature relationships effectively and often produces robust classification outcomes across diverse datasets. In resume screening, Random Forest may improve predictive consistency but introduces additional computational requirements and model complexity.

Advantages:

1) High predictive performance
2) Handles nonlinear feature relationships effectively
3) Reduces overfitting
4) Strong robustness across varied datasets

Limitations:

1) Increased computational requirements
2) More complex implementation
3) Reduced interpretability compared to simpler classification models

E. Bagging

Type: Ensemble Learning

Description:

Bagging (Bootstrap Aggregating) improves predictive stability [7] by training multiple models on randomly sampled subsets of data and aggregating their outputs. This technique reduces variance and enhances classification consistency. In recruitment applications, Bagging can strengthen predictive reliability, although increased computational overhead may affect deployment simplicity.

Advantages:

1) Improves model stability
2) Reduces variance
3) Enhances prediction reliability
4) Useful for strengthening weaker classifiers

Limitations:

1) Higher computational overhead
2) Reduced interpretability
3) Performance depends on selected base learners

F. Comparative Performance Analysis

TABLE I
COMPARATIVE PERFORMANCE OF MACHINE LEARNING METHODS FOR RESUME SCREENING

Algorithm	Accuracy	Precision	Recall	F1-Score	Interpretability	Computational Efficiency
Logistic Regression	0.94	0.93	0.95	0.94	High	High
Naïve Bayes	0.87	0.86	0.88	0.87	Moderate	Very High
SVM	0.92	0.91	0.93	0.92	Moderate	Moderate

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Rando m Forest	0.91	0.90	0.9 2	0.9 1	Moderat e	Moderat e
Baggi ng	0.89	0.88	0.9 0	0.8 9	Moderat e	Moderat e

The comparative analysis indicates that Logistic Regression achieved the highest overall balance of Accuracy, Precision, Recall, and F1-score while maintaining strong interpretability and computational efficiency. SVM demonstrated competitive predictive capability but required higher computational resources. Random Forest provided robust classification performance but introduced additional model complexity. Naïve Bayes offered very fast execution but comparatively lower contextual precision, while Bagging improved stability with moderate performance improvements.

G. Final Model Selection

Based on comparative evaluation, Logistic Regression was selected as the primary classification algorithm for the proposed AI-Powered Resume Screening System because it provided the most balanced combination of predictive performance, deployment practicality, interpretability, and scalability.

1) It achieved the highest overall Accuracy and F1score.
2) It performed effectively in sparse and highdimensional textual environments.
3) It maintained lower computational complexity compared to more advanced models.
4) It provided highly interpretable outputs, improving recruiter transparency and trust [11].
5) It integrated efficiently with TF-IDF and semantic similarity techniques for enhanced resume-job matching.

Although SVM and Random Forest demonstrated strong predictive capability, their increased complexity and computational demands made them less practical for scalable real-time recruitment systems. Naïve Bayes offered speed but lower contextual precision, while Bagging improved stability with additional computational overhead. Therefore, Logistic Regression was identified as the most practical and effective Machine Learning model for the proposed intelligent recruitment framework.

IX. RESULTS AND DISCUSSION

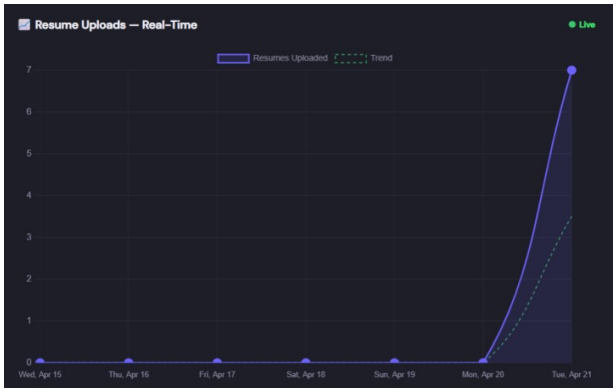


Fig. 2. Real-Time Resume Upload Trend Analysis

Fig. 2 illustrates the real-time resume upload activity monitored through the proposed recruitment dashboard over a weekly period. The graph shows minimal resume submission from April 15 to April 20, followed by a significant increase on April 21, where total uploads sharply rose. This sudden growth indicates increased candidate participation and recruitment engagement. The trend analysis helps recruiters monitor application volume, identify peak submission periods, and improve resource planning for efficient resume screening. Overall, this dashboard output enhances real-time recruitment monitoring and supports data-driven hiring management.

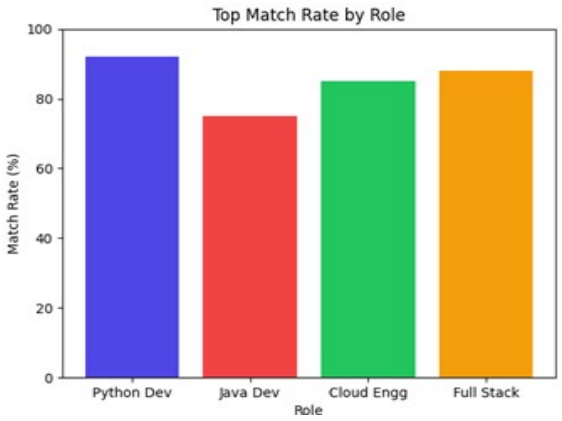


Fig. 3. Top Match Rate by Role Analysis

Fig. 3 illustrates the candidate-job matching performance across different technical roles within the proposed AIPowered Resume Screening System. The graph compares match rates for Python Developer, Java Developer, Cloud Engineer, and Full Stack roles based on resume-job semantic analysis. Among all evaluated roles, Python Developer achieved the highest match rate, followed by Full Stack and Cloud Engineer, while Java Developer showed comparatively lower alignment. These results indicate that the system effectively identifies role-specific candidate suitability by analyzing skill relevance and contextual resume compatibility. This graphical output supports recruiters in understanding role-wise candidate distribution and improves decision-making by highlighting positions with stronger applicant-job alignment.

Fig. 4. Skill Category Distribution Analysis

Fig. 4 illustrates the distribution of candidate skills across multiple technical categories within the proposed recruitment framework. The pie chart represents skill segmentation in Frontend, Backend, Cloud, Database, Machine Learning (ML), and DevOps domains based on uploaded resume data. Among the analyzed categories, Machine Learning and Backend skills constitute a significant portion of candidate profiles, indicating higher applicant concentration in these technical areas. Frontend and DevOps also demonstrate notable representation, while Cloud and Database skills appear comparatively lower. This distribution analysis enables recruiters to understand domain-wise candidate availability, identify skill concentration trends, and optimize hiring strategies according to organizational technical requirements. Overall, the skill category dashboard enhances recruitment intelligence by supporting data-driven workforce planning and targeted candidate evaluation.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Fig. 5. System Performance Analytics Evaluation

Fig. 5 illustrates the multidimensional performance evaluation of the proposed AI-Powered Resume Screening System across key operational parameters including Speed, Accuracy, Coverage, Ranking, Diversity, and Efficiency. The radar chart demonstrates that the system achieves particularly strong performance in Accuracy and Efficiency, indicating reliable candidate classification and optimized recruitment processing. Ranking performance is also notably high, reflecting effective candidate prioritization and shortlist generation. In contrast, Coverage and Diversity show comparatively moderate values, suggesting opportunities for broader candidate inclusion and enhanced dataset variability. Speed performance remains balanced, supporting practical realtime implementation. This comprehensive performance visualization enables recruiters and system developers to assess operational strengths, identify optimization opportunities, and improve recruitment intelligence through data-driven system refinement.

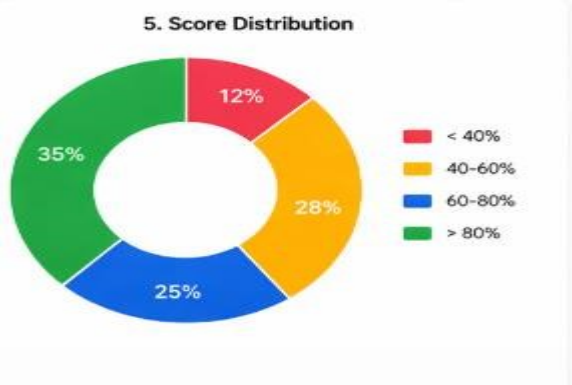


Fig. 6. Candidate Score Distribution Analysis

Fig. 6 illustrates the score distribution of evaluated candidates within the proposed AI-Powered Resume Screening System based on final assessment performance.

The chart categorizes candidates into four score ranges: below 40%, 40–60%, 60–80%, and above 80%. A significant proportion of candidates achieved scores above 80%, representing the largest segment and indicating strong candidate suitability within the evaluated recruitment pool. Candidates scoring between 60–80% and 40–60% also constitute notable portions, reflecting moderate to good performance levels. The smallest percentage falls below 40%, suggesting that only a limited number of applicants demonstrated low suitability. This score distribution analysis enables recruiters to understand candidate quality segmentation, streamline shortlisting decisions, and prioritize high-performing applicants efficiently. Overall, the graphical output strengthens recruitment intelligence by providing clear performancebased candidate categorization for optimized hiring decisions.

X. CONCLUSION

The proposed AI-Powered Resume Screening System provides an intelligent and scalable recruitment framework designed to automate candidate evaluation, improve screening precision, and enhance recruiter decision-making through integrated Artificial Intelligence technologies.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

By combining Natural Language Processing (NLP) [9], Machine Learning (ML) [7], semantic similarity analysis, GPT-assisted interview generation [3], [4], dashboard analytics, and MongoDB-based data management [8], the system successfully addresses major limitations associated with traditional manual resume screening and basic Applicant Tracking Systems.

The implementation of resume parsing, preprocessing, TF-IDF-based feature extraction, and cosine similarity significantly improved candidate-job matching accuracy by evaluating both technical relevance and contextual compatibility. Comparative analysis of multiple Machine Learning algorithms demonstrated that Logistic Regression offered the most balanced combination of predictive performance, computational efficiency, scalability, and interpretability, making it the most suitable classification model for the proposed framework. Experimental results further confirmed that TF-IDF with contextual enhancement improved feature representation quality and strengthened classification precision.

Beyond resume screening, the integration of GPT-based interview intelligence and cognitive assessment modules expanded the system into a more comprehensive recruitment ecosystem capable of evaluating technical knowledge, problem-solving ability, and candidate readiness. Dashboard analytics and real-time visualization further improved recruiter efficiency by providing centralized monitoring of resume uploads, skill distributions, role-based match rates, system performance, and candidate score segmentation.

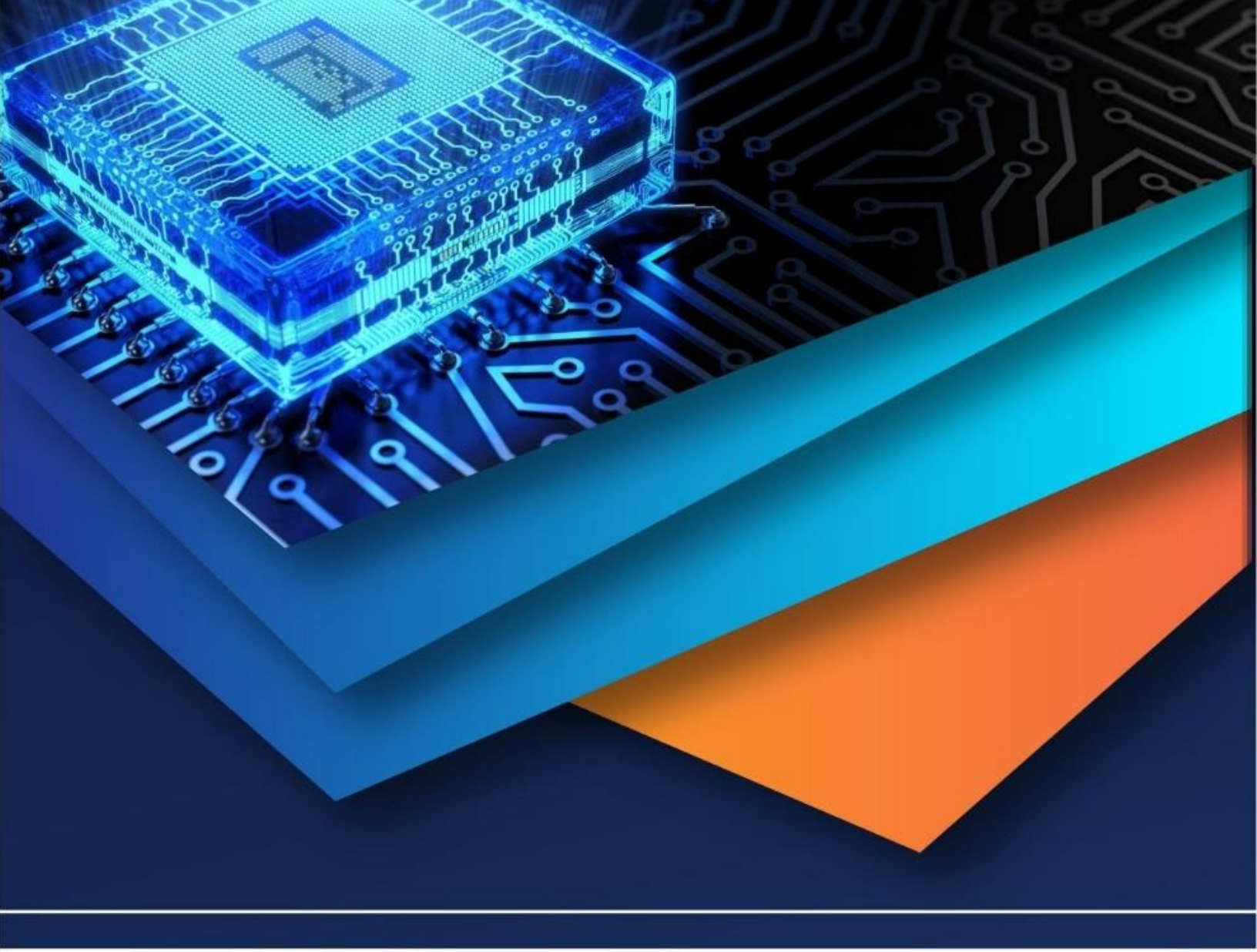
The overall results demonstrate that the proposed system significantly improves recruitment speed, candidate shortlisting accuracy, operational transparency, and datadriven hiring effectiveness. By reducing repetitive manual effort and enabling intelligent candidate evaluation, the framework offers a practical solution for modern digital recruitment environments.

In conclusion, the developed recruitment framework establishes a robust foundation for automated hiring by integrating intelligent classification, semantic analysis, adaptive interview support, and scalable analytics into a unified platform. Its balanced combination of automation, accuracy, and practical usability makes it a highly effective recruitment solution capable of supporting evolving workforce acquisition needs in modern organizations.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pretraining of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.
[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in Proc. Int. Conf. Learning Representations (ICLR Workshop), 2013.
[3] T. Brown et al., “Language models are few-shot learners,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 1877–1901.
[4] OpenAI, “GPT-4 Technical Report,” arXiv:2303.08774, 2023.
[5] S. Zhang, R. Liu, and J. Wang, “Artificial intelligence-based recruitment systems: A survey of automated resume screening techniques,” IEEE Access, vol. 10, pp. 112345–112360, 2022.
[6] A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
[7] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[8] MongoDB Inc., “MongoDB Atlas documentation,” 2024.
[9] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. Sebastopol, CA, USA: O’Reilly Media, 2023 ed.
[10] R. Smith and P. Anderson, “Semantic similarity analysis for intelligent candidate-job matching using NLP,” IEEE Access, vol. 11, pp. 45678–45692, 2023.
[11] K. Johnson, M. Patel, and L. Chen, “Automated hiring frameworks using machine learning and recruiter analytics,” in Proc. IEEE Int. Conf. Data Science and Advanced Analytics, 2024, pp. 214–221.
[12] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[13] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. draft, 2023.
[14] N. Kumar and S. Gupta, “Real-time dashboard analytics for AIenabled recruitment systems,” International Journal of Intelligent Systems and Applications, vol. 16, no. 2, pp. 55–68, 2024.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)