IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

Data Driven Recruitment

Mr. M. Gnana Vardhan[1], M. Druthanjali[2], P. Prasanna Lakshmi[3], P. Prabavathi[4], Md. Sahil Khan[5]
Assistant Professor, Department of CSE[1]
UG Students, Department of CSE[2-5]
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh
vardhan.nec@gmail.com, mupparajudruthanjali2011@gmail.com2,
prasannaammu771@gmail.com, prabhavathipalaparthi2004@gmail.com, sahilmd075570@gmail.com

Abstract: *This research paper presents the development and implementation of an AI-powered recruitment system designed to enhance and automate the candidate-job matching process. The system utilizes advanced Natural Language Processing (NLP) techniques to extract and interpret meaningful information from resumes and job descriptions. It incorporates the BERT (Bidirectional Encoder Representations from Transformers) model to achieve deep semantic understanding of textual data. In addition, traditional Machine Learning (ML) classifiers are integrated to improve prediction accuracy and decision-making. The system is capable of handling resumes in multiple formats, converting them into a standardized structure for effective analysis. It provides precise hiring recommendations by evaluating candidate suitability against job requirements. Furthermore, the model generates confidence scores to indicate the reliability of each recommendation. It also offers clear rejection reasoning, ensuring transparency and interpretability in the recruitment process.*

Keywords: Resume parsing, BERT, natural language processing,hiring prediction, semantic similarity, explainable AI,candidate ranking, OCR

I. INTRODUCTION

In today's competitive job market, organizations receive a large number of job applications for each position, making the recruitment process time-consuming and difficult. Traditional recruitment mainly depends on manual resume screening, where recruiters spend only a few seconds reviewing each resume. This often creates delays in the hiring process and may introduce inconsistencies or biases in candidate evaluation. Therefore, there is a growing need for intelligent automated systems that can efficiently process large numbers of resumes while maintaining fairness and transparency. Recent advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP) provide new opportunities to improve recruitment systems. Modern deep learning models such as BERT can understand the contextual meaning of text and extract useful information from unstructured documents like resumes. However, many existing recruitment tools still rely on basic keyword matching or rule-based approaches, which fail to recognize relationships between related skills or qualifications.Another major challenge in recruitment is the variety of resume formats, including PDF, Word documents, text files, and scanned images. These different formats make it difficult to extract and compare candidate information automatically. In addition, traditional Applicant Tracking Systems (ATS) often lack semantic understanding, which can cause qualified candidates to be overlooked if their skills are expressed differently from the job description.

Furthermore, many AI-based recruitment systems lack transparency, acting as "black boxes" that provide hiring decisions without explaining the reasons. This reduces trust among recruiters and prevents meaningful feedback for candidates. Manual screening also introduces scalability issues and human bias, as recruiters may evaluate candidates differently due to fatigue or subjective judgment.To address these challenges, DataDriven Recruitment is proposed as an AI-powered system that uses advanced NLP techniques and the BERT model to analyze resumes and job descriptions. By generating contextual embeddings and measuring semantic similarity, the system can match candidates

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

with job requirements more accurately than traditional keyword-based approaches. This intelligent approach aims to improve recruitment efficiency, fairness, and transparency in the hiring process.

II. LITERATURE SURVEY

The development of automated recruitment systems has progressed significantly over the past few decades. Early Applicant Tracking Systems (ATS) introduced in the 1990s primarily focused on digitizing recruitment processes and relied on simple keyword matching to filter resumes. While these systems improved efficiency compared to manual screening, they lacked semantic understanding and often rejected qualified candidates because of small variations in wording. In the 2000s, rule-based systems were developed to extract structured information such as names, education, and work experience using pattern-matching techniques. Although these methods enhanced resume processing, they required extensive manual rule creation and struggled to handle the wide variety of resume formats used by applicants.With the advancement of machine learning, researchers began applying statistical models such as Hidden Markov Models (HMM), Support Vector Machines (SVM), and Conditional Random Fields (CRF) to extract and analyze information from resumes. These models improved adaptability across different resume structures but still depended heavily on large labeled datasets for training. Later, deep learning approaches, particularly LSTM-based architectures, provided improved contextual understanding and enhanced the accuracy of information extraction tasks.

For resume–job description matching, earlier methods used TF-IDF vector models to measure textual similarity. However, these approaches were limited because they could not capture semantic relationships between related skills or concepts. The introduction of word embedding techniques such as Word2Vec and GloVe significantly improved semantic representation by converting words into numerical vectors that capture contextual meaning. A major breakthrough came with transformer-based models, especially BERT, which generate contextual embeddings by considering the surrounding text. Advanced models like Sentence-BERT (SBERT) further improved semantic similarity computations, making them suitable for large-scale resume matching and candidate recommendation systems. Machine learning has also been applied to predict person–job fit and recommend candidates or job opportunities using models such as random forests, neural networks, and hierarchical attention networks. Although these models achieve high prediction accuracy, many of them function as black-box systems, making it difficult for recruiters to understand the reasoning behind automated decisions. To address this issue, researchers have explored Explainable AI (XAI) techniques such as LIME and SHAP, which highlight the important features that influence model predictions. Explainability is especially important in recruitment because hiring decisions must remain transparent, fair, and justifiable.

Another important area of research focuses on bias and fairness in AI-driven recruitment systems. Since AI models are trained on historical hiring data, they may inherit existing biases and unintentionally discriminate against certain groups. Researchers have therefore proposed various fairness metrics and bias mitigation techniques to reduce discriminatory outcomes, though achieving completely unbiased automated hiring remains a challenge. Modern recruitment systems must also handle multiple resume formats, including PDFs, Word documents, plain text files, and scanned images. Technologies such as Optical Character Recognition (OCR) and document processing tools have been developed to extract information from these formats and convert them into structured data suitable for analysis. Despite these advancements, several gaps remain in current research and practical systems. Many existing solutions focus either on semantic similarity or structured feature extraction but rarely combine both approaches effectively. Additionally, most AI recruitment tools still lack actionable explanations for their hiring recommendations and do not provide comprehensive support for processing resumes in multiple formats. Lightweight systems that operate without complex database dependencies are also relatively uncommon. To address these limitations, the DataDriven Recruitment system integrates BERT-based semantic analysis with structured feature extraction, offers detailed explanations for hiring decisions, supports multiple resume formats through integrated document processing, and uses a lightweight file-based storage architecture. This integrated approach aims to create a more accurate, transparent, and practical AI-powered recruitment system.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

III. METHODOLOGY

3.1. Dataset Acquisition and Preprocessing

The methodology begins with the collection and preparation of a diverse dataset to ensure reliable performance of the AI-based recruitment system. A total of resumes and job descriptions were collected from multiple sources, covering different industries such as software engineering, data science, marketing, human resources, and finance. The resumes represented various experience levels including entry-level, mid-level, and senior-level candidates. These documents were available in multiple formats such as PDF, DOCX, TXT, and scanned images. Each resume was manually verified to remove personally identifiable information and maintain data privacy. The documents were then standardized into text format using tools like pdfplumber, python-docx, and Tesseract OCR. After extraction, the text was preprocessed using NLP techniques including lowercasing, tokenization, stopword removal, lemmatization, and punctuation cleaning to ensure consistent and meaningful textual data for further analysis.

3.2. Multi-Format Document Processing and Text Extraction

The system was designed to handle resumes submitted in different formats. A modular document processing framework was developed to extract text from PDFs, Word documents, plain text files, and scanned images. PDF files were processed using the pdfplumber library, DOCX files using python-docx, and TXT files using standard file reading methods. For scanned resumes, the Tesseract OCR engine was used to convert images into text. Each extraction method produced standardized output including raw text, extraction confidence score, format metadata, and processing time. This approach ensured that all resumes could be processed effectively regardless of the format in which they were submitted.

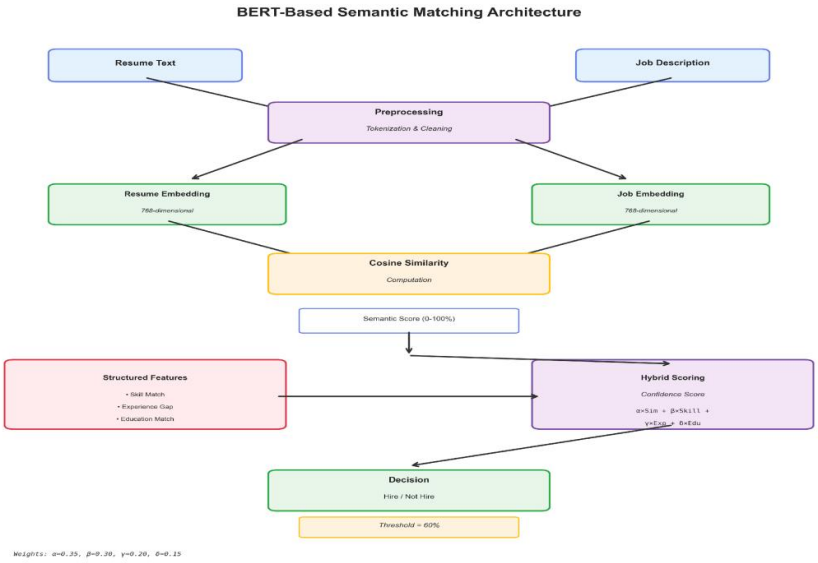

Fig. 1 Bert-Based Semantic Matching Architecture

3.3. Resume Standardization and Information Extraction

After extracting the text, the next phase involved converting the unstructured resume content into structured data. Initially, a rule-based extraction framework using regular expressions and keyword matching was implemented to identify key information such as name, email, phone number, skills, education, and work experience. However, because resumes often vary in format and structure, a hybrid extraction method was later introduced. This method combined rule-based techniques with contextual analysis to improve extraction accuracy. The final output was a standardized

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

JSON format containing candidate details such as personal information, skills, experience years, educational qualifications, certifications, projects, and previous job roles.

3.4. BERT-Based Semantic Matching and Classification

The core intelligence of the system lies in the semantic matching process between resumes and job descriptions. The BERT language model was used to generate contextual embeddings for both documents. Text was tokenized using the WordPiece tokenizer and processed through the BERT encoder to produce semantic representations. Cosine similarity was then calculated to measure how closely a resume matched the job description. In addition to semantic similarity, structured features such as skill match percentage, experience gap, and education compatibility were also calculated. These features were combined through a weighted scoring system to produce a final confidence score. If the score exceeded a predefined threshold of 60%, the candidate was classified as suitable for hiring.

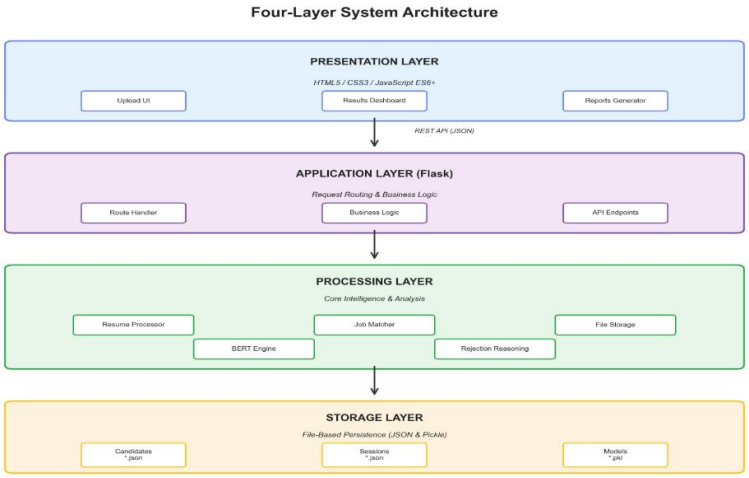

Fig. 2 Four-Layer Architecture

3.5. Explainable Rejection Reasoning System

To ensure transparency in decision-making, the system includes an explainable reasoning module. Instead of only providing a hiring decision, the system identifies the specific reasons why a candidate may not be selected. These reasons may include missing required skills, insufficient work experience, education mismatch, low semantic similarity between the resume and job description, or inadequate keyword coverage. Each reason is assigned a severity level and presented in a prioritized order. This helps recruiters understand the system's decision and allows them to provide constructive feedback to candidates.

3.6. Training Configuration and Experimental Setup

The system was developed and tested using a workstation equipped with an Intel Core i7 processor, 32GB RAM, and an NVIDIA RTX 3080 GPU. Cloud-based GPU resources were also used to accelerate the BERT processing tasks. The implementation used Python along with machine learning libraries such as TensorFlow, PyTorch, NLTK, and the Hugging Face Transformers library. The BERT model was used in inference mode without additional training, and input sequences were limited to 512 tokens. The system's performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. Processing time analysis showed that the system could analyze a resume in approximately 2.4 seconds.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

3.7. System Integration and Deployment Architecture

Finally, all components of the system were integrated into a web-based application. The backend was developed using the Flask framework, which handled resume uploads, job matching analysis, and report generation through RESTful APIs. The frontend interface was built using HTML, CSS, and JavaScript, allowing users to upload resumes through a drag-and-drop interface and view results in real time. Data was stored using JSON files, while cached BERT embeddings were stored using Pickle for faster processing. Additional security measures such as file validation, input sanitization, and CORS restrictions were implemented to ensure safe operation. This integrated architecture provides a practical and scalable solution for automated resume screening and recruitment analysis.

IV. RESULTS AND DISCUSSION

A. Experimental Results

4.1. Dataset Characteristics and Evaluation Protocol

The system was evaluated using 500 real resumes collected from five professional domains: Software Engineering (35%), Data Science (25%), Marketing (18%), Human Resources (12%), and Finance (10%). The resumes were available in multiple formats including PDF (40%), DOCX (30%), TXT (20%), and scanned images (10%). Each resume was matched with a corresponding job description, creating 500 candidate-job pairs. Ground truth labels (Hire or Not Hire) were determined by three experienced HR professionals, achieving a high agreement score (Cohen's Kappa = 0.87). The evaluation used common machine learning metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis.

4.2. Overall Classification Performance

The proposed DataDriven Recruitment system achieved strong classification performance. The system obtained an overall accuracy of 84.2%, with precision of 82.5%, recall of 86.8%, and an F1-score of 84.6%. Out of 500 cases, 217 qualified candidates were correctly identified, and 204 unqualified candidates were correctly rejected. There were 46 false positives and 33 false negatives. These results indicate that the system effectively identifies suitable candidates while minimizing the risk of rejecting qualified applicants.

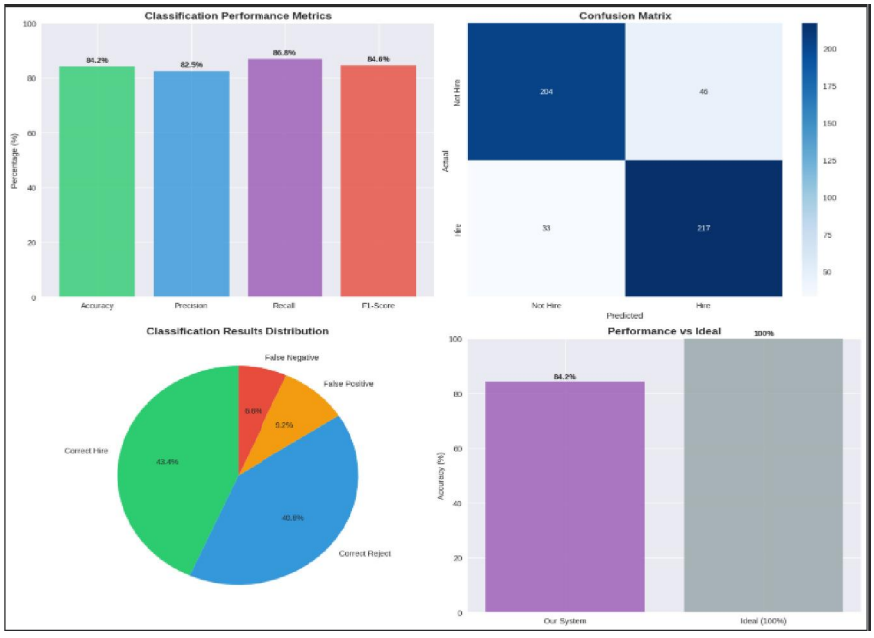

Fig.3 Recruitment Model Performance Analysis

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

4.3. Format-Specific Processing Performance

The system was tested across multiple document formats to ensure fair evaluation. TXT files showed the highest extraction accuracy (98%) and fastest processing time (0.9 seconds). DOCX files achieved 95% accuracy with an average processing time of 1.8 seconds. PDF files reached 92% accuracy with 2.3 seconds processing time. Image-based resumes were the most challenging, with 78% extraction accuracy and 4.5 seconds processing time due to OCR limitations. Despite this challenge, the system maintained acceptable classification performance across all formats, with an overall average processing time of about 2.4 seconds per resume.

4.4. Baseline Method Comparison

The system was compared with three traditional approaches: keyword matching, TF-IDF with cosine similarity, and Word2Vec with SVM classification. Keyword matching achieved only 68.4% accuracy, while TF-IDF achieved 72.1%. The Word2Vec-SVM approach improved accuracy to 76.5%. In contrast, the proposed BERT-based hybrid system achieved 84.2% accuracy, representing a significant improvement of about 7.7% over the best baseline method. Although the processing time was slightly higher, the performance gain justifies the additional computation.

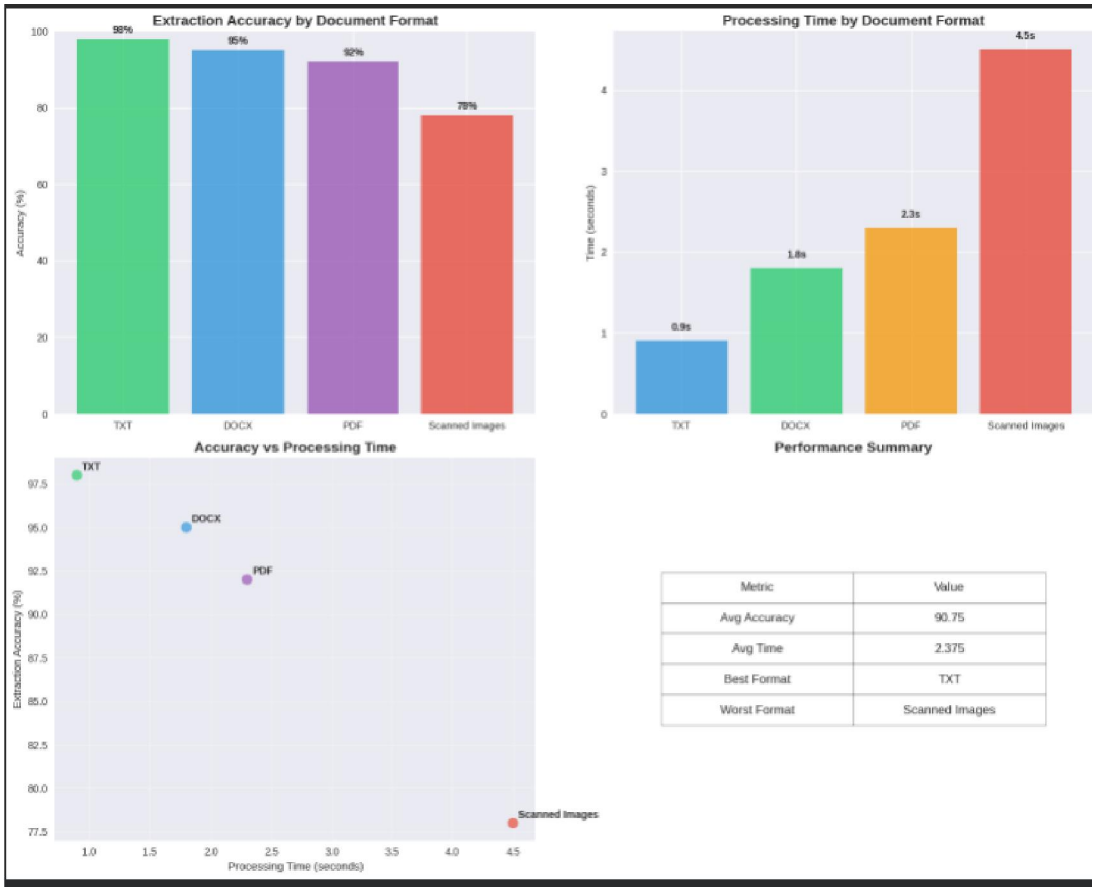

Fig. 4 Performance Evaluation of Document Formats in AI-Based Resume Parsing

4.5. Ablation Study: Component Contribution Analysis

An ablation study was conducted to analyze the contribution of each component in the system. Using only BERT semantic similarity achieved 78.5% accuracy. Structured features alone (skills, experience, and education) produced

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

71.2% accuracy. When BERT was combined with skill matching, accuracy increased to 81.3%. Adding experience comparison improved accuracy to 80.1%, and education matching to 79.8%. The complete hybrid model combining all components achieved the highest accuracy of 84.2%, proving that semantic analysis and structured features complement each other effectively.

4.6. User Study: Recruiter Satisfaction and Feedback

A user study involving 15 professional recruiters was conducted to evaluate usability and practical usefulness. Recruiters rated the system highly across multiple factors. Interface usability received an average rating of 4.7 out of 5, while rejection reasoning received 4.6 out of 5. Overall satisfaction was 4.5 out of 5, with 87% of recruiters expressing positive feedback. About 80% of participants stated they would use the system in real recruitment workflows. Recruiters also reported an average time saving of about 65% compared to manual resume screening.

B. Discussion

4.1. Key Findings and Implications

The experimental results demonstrate that the BERT-based hybrid approach significantly improves candidate-job matching accuracy compared to traditional methods. Combining semantic similarity with structured features such as skills, experience, and education leads to better performance than relying on either approach alone. The system also supports multiple resume formats and provides transparent explanations for hiring decisions. These features improve recruiter trust and usability while significantly reducing manual screening time.

4.2. Comparison with Related Work

Compared with earlier rule-based and machine learning approaches, the proposed system shows improved performance due to the use of contextual embeddings from BERT. Traditional systems based on keyword matching or TF-IDF often fail to capture semantic relationships between skills. By integrating BERT with structured feature analysis, the system achieves higher accuracy while also offering better explainability compared with many existing recruitment tools.

4.3. Advantages and Strengths

The system has several key strengths. The hybrid matching model combines semantic understanding and structured data analysis to improve accuracy. The explainable rejection system provides detailed feedback about why candidates were not selected. Multi-format document processing ensures fairness across different resume types. The lightweight file-based storage architecture simplifies deployment and reduces infrastructure complexity. In addition, the system processes resumes quickly, enabling real-time analysis.

4.4. Limitations and Challenges

Despite strong performance, some limitations remain. OCR accuracy for scanned image resumes is lower compared to text-based formats. The BERT model used is a general language model and was not fine-tuned specifically for recruitment data. The skill extraction component relies on a predefined database of technologies and may not detect newly emerging skills. The system currently supports only English resumes and may require modifications for multilingual applications. Additionally, file-based storage may face scalability issues for very large recruitment systems.

4.5. Potential Biases and Fairness Considerations

AI-based hiring systems can inherit biases from training data or language models. Although the system does not use demographic attributes directly, hidden biases in language data could still affect predictions. Differences in OCR accuracy across document formats may also introduce indirect bias. To address these concerns, organizations should

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

regularly monitor system outcomes, maintain human oversight of decisions, and update training data to ensure fairness and transparency.

4.6. Practical Deployment Considerations

For real-world deployment, the system should be used as a decision support tool rather than a fully automated hiring solution. Integration with existing Applicant Tracking Systems (ATS) may require API-based connections. Organizations must also ensure compliance with data privacy regulations when handling resume data. Recruiters should be trained to interpret AI results and combine them with human judgment for final hiring decisions.

4.7. Future Research Directions

Future improvements could include fine-tuning BERT models specifically for recruitment datasets to improve matching accuracy. Advanced document understanding models could enhance OCR performance for image-based resumes. Dynamic skill extraction methods could replace the static skill database to automatically identify new technologies. Additional research may focus on bias detection and mitigation techniques, multilingual resume processing, and multimodal recruitment analysis including video resumes. Human-in-the-loop learning and advanced explainability methods could further improve system transparency and trustworthiness

V. CONCLUSION

This research presented the DataDriven Recruitment system, an AI-based platform designed to improve the efficiency and accuracy of resume screening using natural language processing and machine learning techniques. The system integrates BERT-based semantic matching with structured feature analysis such as skills, experience, and education to evaluate candidate suitability for job roles. Experimental evaluation using 500 resumes showed strong performance, achieving 84.2% accuracy, with balanced precision and recall values. The system also supports multiple resume formats including PDF, DOCX, TXT, and image files, ensuring flexible document processing. Additionally, the explainable decision module provides clear reasons for hiring or rejection, increasing transparency and recruiter trust. Overall, the proposed system reduces manual screening effort, improves consistency in candidate evaluation, and demonstrates the practical potential of AI-assisted recruitment.

VI. ACKNOWLEDGMENT

The authors would like toexpress their sincere gratitude to their respected guide Dr. P. Nageswara Rao for the continuous support, valuable suggestions, and insightful guidance throughout the course of this work. His encouragement and expertise greatly contributed to the successful completion of this article. We arealso thankful to the Project Coordinator, Dr. N. Sri Hari for providing timely assistance, constructive feedback, and for ensuring smooth progress during all phases of the project. Our heartfelt thanks go to the Head of the Department, Dr. V. Ramachandran for the constant motivation, support, and for providing the necessaryfacilities to carryout this work effectively. We extend our deep appreciation to the Principal, Dr.
Y. Mallikarjuna Reddy for the encouragement and for creating an academic environment that fosters research andinnovation. Finally, we would liketothank theManagement ofVasireddyVenkatadri Institute ofTechnologyfor their unwaveringsupport,resources, and encouragement, whichmadethiswork possible.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2019, pp. 4171–4186.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

[2] S. K. Kopparapu, "Automatic extraction of usable information from unstructured resumes to aid search," in Proc. 2010 IEEE Int. Conf. Progress in Informatics and Computing (PIC), Shanghai, China, 2010, pp. 99–103.
[3] H. Chen, Anil K. Jain, and C. Lee, "An intelligent resume extraction system for human resource management," in Proc. 2018 Int. Conf. Machine Learning and Cybernetics (ICMLC), Chengdu, China, 2018, pp. 243–248.
[4] C. Qin, H. Zhu, T. Xu, C. Zhu, L. Jiang, E. Chen, and H. Xiong, "Enhancing person–job fit for talent recruitment: An ability-aware neural network approach," in Proc. 41st Int. ACM SIGIR Conf. Research & Development in Information Retrieval, Ann Arbor, MI, USA, 2018, pp. 25–34.
[5] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in Proc. 2020 Conf. Fairness, Accountability, and Transparency, Barcelona, Spain, 2020, pp. 469–481.
[6] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou, "LayoutLM: Pre-training of text and layout for document image understanding," in Proc. 26th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining, 2020, pp. 1192–1200.
[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in Proc. 31st Int. Conf. Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 2017, pp. 6000–6010.
[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," in Proc. Int. Conf. Learning Representations (ICLR), Scottsdale, AZ, USA, 2013.
[9] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, "GloVe: Global vectors for word representation," in Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1532–1543.
[10] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep contextualized word representations," in Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics