International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

A dual-benefit framework for AI-powered resume analysis and job matching: Integrating NLP, machine learning, and skill-gap analytics

Kavya Mittal, Manthan Awasthi, Kartiekey Bharadwaj *, Kartik Gupta and Sheradha Jauhari

Ajay Kumar Garg Engineering College India.

International Journal of Science and Research Archive, 2026, 19(02), 110-119

Publication history: Received on 18 March 2026; revised on 02 May 2026; accepted on 04 May 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.19.2.0995

Abstract

Manual resume screening is a difficult, time-consuming, and biased task in today's recruitment environment due to the tremendous volume of job applications. Modern one-click job portals and remote work trends have accelerated the growth of digital applications, which has greatly exceeded human resource processing skills. Conventional Applicant Tracking Sys-tems (ATS) rely on strict, keyword-based screening algorithms that often reject highly competent applicants because of missing exact-match buzzwords, synonymous terminology, or formatting quirks. This seriously harms the talent acquisition pipeline by giving candidates an opaque "black-box" experience and producing a high percentage of false negatives for employers.

This study presents a novel, highly scalable Dual-Benefit Framework that parses extremely unstructured resume pa-pers using sophisticated Natural Language Processing (NLP) and Ensemble Machine Learning (ML) approaches, turning unstructured material into organised, useful intelligence. The platform serves as an educational partner for job seekers, offering quick feedback, an overall resume compatibility score, and a thorough, data-driven skill-gap analysis that is directly connected to online learning materials. The system provides recruiters with a centralised dashboard that displays automatic candidate-to-job matching scores based on an optimised Random Forest classification engine, Cosine Similarity, Jaccard Similarity, and Term Frequency-Inverse Document Frequency (TF-IDF) vectorisation. We also investigate methods to prevent adversarial resume manipulation (keyword stuffing), stringent data privacy procedures, and mathematical enforcement of algorithmic fair-ness.This framework reduces recruitment timelines by over 98%, minimises unconscious demographic bias at the point of ingestion, and promotes an equitable, transparent hiring ecosystem for both employers and applicants by actively addressing the Sustainable Development Goals (SDGs) of the United Nations, specifically SDG 4, SDG 8, and SDG 10.

Keywords*:* Natural Language Processing (NLP); Machine Learning; Applicant Tracking System (ATS); Skill Gap Analysis; TF-IDF; Named Entity Recognition (NER); Bias Mitigation; Explainable AI (XAI); Random Forest; Adversarial Robustness.

1. Introduction

1.1. The Evolution of Recruitment Technologies

The importance of automated resume screening has grown as the number of job applications in the globalised digital market continues to soar [12]. Due to the convenience of "one-click" applications on contemporary job portals such as Indeed and LinkedIn, enterprise organisations frequently receive thou-sands of resumes for a single open position [11]. The manual screening method used in the traditional Human Resources (HR) procedure has proven to be a major bottleneck. Hiring managers must sort through mountains of paperwork, which is a labour-intensive, time-consuming activity that is very vulnerable to unconscious cognitive bias and human mistake. [14].

* Corresponding author: Kartiekey Bharadwaj

In an effort to address this, early versions of Applicant Tracking Systems (ATS) used exact keyword matching and Boolean search logic [1]. But due to their lack of contextual awareness, these legacy systems are unable to distinguish between a candidate who simply attended a seminar on cloud computing and one who architected a reliable cloud system [16]. This is when the paradigm is radically altered by the combination of artificial intelligence (AI) and natural language processing (NLP). Modern systems bridge the semantic gap between recruiters' fundamental demands and candidates' genuine skills by gathering enormous volumes of unstructured text data and using advanced NLP libraries like PyResparser and spaCy. [17], [20].

1.2. Motivation and Rationale

An urgent industry requirement to close the gap between cutting-edge computational technology and fair human resources management is the main driving force behind this comprehensive research project. Our suggested approach is essentially a Dual-Benefit platform, in contrast to conventional ATS tools that solely benefit the hiring manager at the expense of the candidate experience. It is designed from the ground up to meet important global socioeconomic goals:

- **SDG 8 (Decent Work and Economic Growth):** The technology reduces frictional unemployment and promotes macroeconomic productivity by connecting highly qualified individuals with appropriate employment oppor-tunities through a mathematically efficient and equitable hiring procedure [11].
- **SDG 4 (Quality Education):** The system guides users to-ward ongoing lifetime learning and upskilling by offering candidates a customised skill-gap analysis and tailored online course recommendations [15].
- **SDG 10 (Reduced Inequalities):** The approach encour-ages fair chances for all by methodically minimising human prejudice in the first screening stage—explicitly removing identifiers that cause gender, racial, or age bias [12].

2. Literature Review and Related Work

A comprehensive review of the literature is necessary to comprehend the deeply ingrained issues in the recruitment sector. From simple keyword parsers to contemporary deep learning systems, we classify the related work into several categories.

2.1. Limitations of Traditional ATS and Keyword Matching

In the past, ATS systems were simply created as digital filing cabinets. According to Johnson (2019), the traditional ATS's dependence on exact keyword matching results in an intolerable rate of "false negatives"—candidates who are fully capable but whose resumes don't contain a particular, HR-defined buzzword [1]. Additionally, Deepa et al. (2025) observe that when a candidate employs a contemporary, multi-column graphical template [16], conventional parsers that rely on fixed spatial coordinates or regular expressions (RegEx) completely fail.

2.2. NLP and Semantic Embeddings

Recent research has shifted to semantic understanding in order to get beyond strict keyword restrictions. The effectiveness of Word2Vec and GloVe embeddings to capture the contextual links between various skills was shown by Singh et al. (2022) [8]. Similarly, Saatc¸ı et al. (2024) shown that Jaccard Similarity is frequently better for technical tasks where the presence of precise hard skills (e.g., "Python", "AWS") is non-negotiable [10], whereas Cosine Similarity is ideal for general semantic alignment.

2.3. Machine Learning and Deep Classification

Researchers used predictive modelling in addition to basic similarity matching. Based on past hiring data, Lee and Kim (2020) used machine learning to forecast candidate-job fit [5]. To recommend employment to candidates, Wang and Zhao (2021) developed Deep Learning hybrid recommendation en-gines [6]. By using an ensemble Random Forest classifier,which is highly valued for its resistance to overfitting in high-dimensional NLP feature sets [11], [14], our method expands on this.

2.4. Explainable AI (XAI) in Recruitment

Explainable AI (XAI) is a developing subfield in AI hiring. Laws in several jurisdictions (such as the EU's AI Act) increasingly require an explanation when an AI rejects a candidate. In order to map the precise logic behind a candidate's score, Sharma (2022) investigated Knowledge Graph-based methods [9]. By giving the user direct access to the skill-gap analysis and a clear explanation of the precise algorithmic reasoning, our approach tackles XAI.

2.5. Ethical AI and Bias Mitigation

Algorithmic bias is a major issue in AI hiring. Shah et al. (2025) specifically caution that human biases about gender, ethnicity, and educational status would unavoidably be inher-ited by models trained on past HR data [12]. Martinez (2020) proposed that this can be lessened by extracting only pertinent professional entities using highly constrained Named Entity Recognition (NER) [3].

3. Proposed Methodology and System

3.1. Architecture

We created a multi-layered, highly scalable architecture with a Presentation Layer, a Processing Layer, an Analysis/Recommendation Layer, and a Data Persistence Layer in order to successfully address the issues that were found.

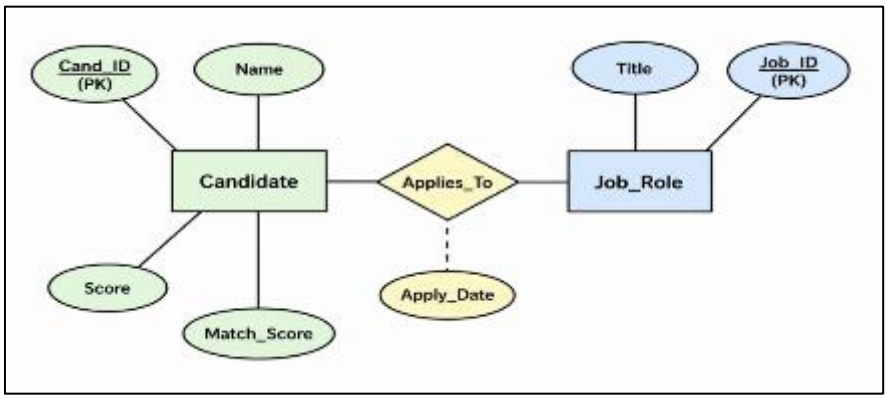

Figure 1 Architectural Block Diagram of the Dual-Benefit System**.**

3.2. Data Ingestion and Preprocessing

Data ingestion is the first step in the procedure. The candi-date uses a secure gateway to upload a resume (PDF/DOCX). The system removes graphics, tables, and other layout format-ting by using document parsers to retrieve raw string data [18]. The system uses heuristic fallbacks to avoid pipeline failure in the event that crucial fields are unreadable. Missing data is handled programmatically.

Advanced NLP Pipeline and Entity Extraction

The NLP pipeline receives the raw text and mostly usesPyResparserbuilt upon spaCyand NLTK[20].

- **Tokenization & Lemmatization:** Words are reduced to their most basic dictionary form (for example, "developing" becomes "develop") and the text is broken down into its component words. Let $W = \{w_1, w_2, ..., w_n\}$ be the word sequence. To cut down on computational overhead, stop words are methodically eliminated.
- **Part-of-Speech (POS) Tagging:** In order to determine the impact score of the resume, the system prioritises action verbs (such as "Managed" and "Designed") and recognises nouns and verbs to comprehend the syntactic structure.
- **Named Entity Recognition (NER):** Discrete entities are identified via the spaCy NER model, which is based on a Convolutional Neural Network (CNN) architecture. Text chunks are categorised using preset labels like SKILL, DEGREE, ORG, and DATE [3].
- **RegEx Fallbacks:** Deterministic patterns, such as phone numbers and email addresses, can be cleanly extracted using regular expressions.

3.3. Inference of Soft Skills via Contextual Analysis

Soft skills like leadership, teamwork, and conflict resolution are frequently shown implicitly rather than stated openly, but

Algorithm 1: Hybrid Candidate–Job Matching

Input: ResumeText, JobDescText
Output: FinalScore, SkillGaps

1. Extract Rskills from ResumeText
2. Extract Jskills from JobDescText
3. Compute TF-IDF vectors
4. Calculate Cosine Similarity
5. Calculate Jaccard Similarity
6. FinalScore = (wc × Scos) + (wj × Sjac)
7. SkillGaps = Jskills − Rskills
8. Return FinalScore, SkillGaps

3.4. Semantic Matching: The Hybrid Score Algorithm

We use a hybrid matching score (S_{match}) that combines two different mathematical similarity models to measure candidate relevance. This guarantees a thorough assessment that captures both hard requirements and semantic purpose. 1. Cosine Similarity: By measuring the semantic orientation (the angle between vectors in the multidimensional space) rather than magnitude, cosine similarity ensures that candidates with overly wordy resumes do not unfairly outperform those with succinct resumes mathematically [17]. Let $\mathbf{J}$ stand for the job description vector and $\mathbf{R}$ for the resume vector:

$$S\_cosine = (\Sigma RiJi) / (\sqrt{\Sigma} Ri^2 \times \sqrt{\Sigma} Ji^2)$$

explicit NER. We use a contextual sliding window method in our sophisticated NLP process. The model uses pretrained Word2Vec embeddings to determine the semantic proximity of phrases like ”led a cross-functional team of 10” or ”managed **2. Jaccard Similarity:** To enforce strict requirements (e.g., a mandatory certification), we utilize Jaccard Similarity, which evaluates the intersection over the union of the skill sets [10].

$|R \cap J|$

conflict during sprint delivery,” which the parser finds. The model then awards implicit skill points to the candidate [8]. $\mathrm{Sjaccard}(\mathrm{R}, \mathrm{J}) = \frac{}{|R \cup J|}$ (5)

Feature Engineering: TF-IDF Vectorization

The extracted textual data (Candidate Skills vs. Job Re-quirements) must be vectorised in order to do quantitative comparisons. Term Frequency-Inverse Document Frequency (TF-IDF) [17] is what we use.

The normalised occurrence of a term t in a document d is measured by the Term Frequency (TF):

$$TF(t, d) = f\_(t, d)/(\Sigma\_(t' \in d)\ f\_(t', d)) \quad (1)$$

The scoring process is formalised in Algorithm 1.

F. Job Role Recommendation Engine and Optimization

We formulate job suggestion as a multi-class classification issue to offer candidates alternate career routes in the event that their initial application is rejected. We use a **Random Forest Classifier**, an ensemble learning technique made up of several decision trees [14].

We used both Information Gain and Gini Impurity to assess a split’s quality at each decision tree node. Entropy (H) is the

$TF(t, d) = f\_(t, d)/(\Sigma\_(t' \in d)\ f\_(t', d))$ (2)

source of the Information Gain: $\mathbf{C}$

Σ

Throughout the corpus D of total size N , the Inverse Doc-ument Frequency (IDF) promotes uncommon, highly specific technical terms (like ”Kubernetes”) and penalises common terms (like ”teamwork”).

$$H(S) = -\Sigma\_(i = 1)\^C \; p\_i \; log\_2(p\_i) \quad (3)$$

i=1

Because it is computationally faster for high-dimensional text data, the algorithm ultimately divides nodes by minimising the

$$IDF(t, D) = log(N/(1 + |d \in D: t \in d|)) \quad (4)$$

$$Gini = 1 - \Sigma\_(i = 1)\^C \; (p\_i)\^2 \quad (5)$$

The final weight for a given skill is the product:

$$TF\text{-}IDF\,(t, d, D) = TF\,(t, d) \times IDF\,(t, D) \quad (6)$$

i=1

Where p_i is the probability of the candidate belonging to a specific job role out of C possible roles

We used GridSearchCV with 5-Fold Cross-Validation for Hyperparameter Tuning in order to optimise the model. The best values found were n_estimators= 200, max_depth

= 50, and min_samples_split = 5. The model generates a probability distribution that suggests the top three potential roles.

Database Schema and System Implementation

The system architecture is based on a strong, contemporary technological stack: MySQL for highly organised relational data persistence [11], [15], Streamlit for the reactive frontend, and Python for backend algorithmic processing.

Database Entity-Relationship Schema

The MySQL database is normalised to the Third Normal Form (3NF) in order to preserve data integrity and enable sophisticated analytical queries without anomalies.

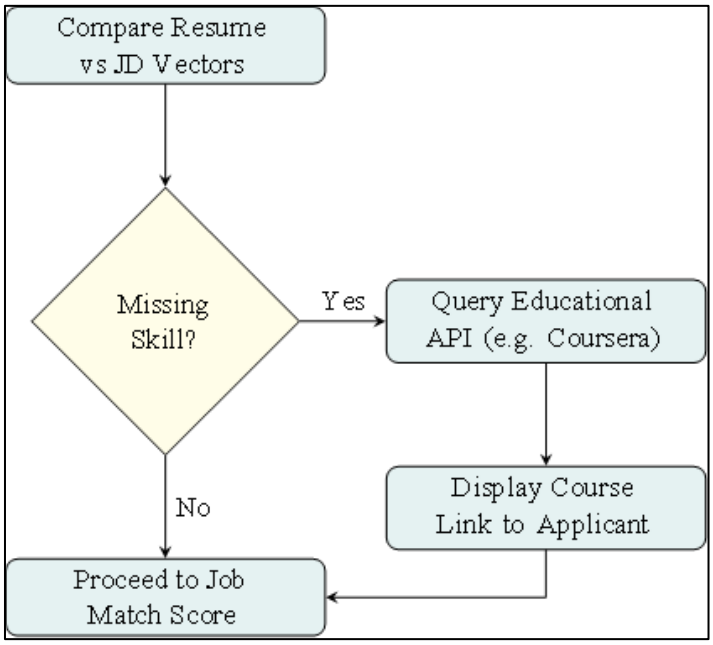

Figure 2 Entity-Relationship (ER) Diagram of the Core Database.

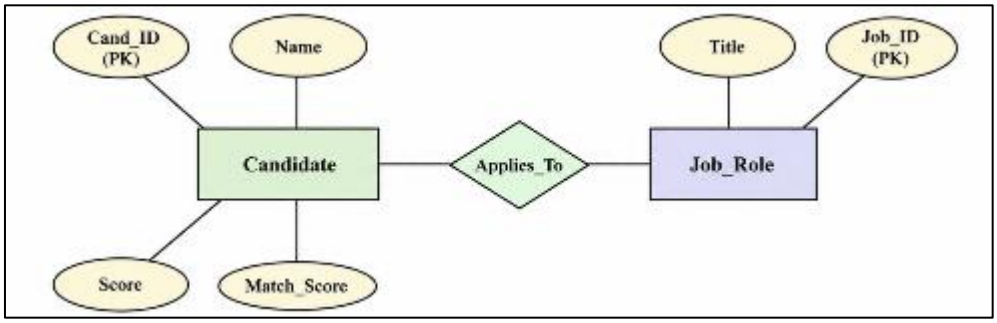

Figure 3 Flowchart depicting the automated Skill-Gap Course Recommendation Engine.

$$Recall = TruePositives + FalseNegativesPrecision \times Recall \quad (7)$$

$$Precision + Recall \quad (8)$$

3.5. The Applicant Experience

In contrast to the "black-box" nature of previous systems, the applicant interface is built for complete transparency:

Overall Resume Score: A quantitative measure that evaluates the existence of actionable verbs, formatting quality, and content brevity [11].

Skill-Gap Analysis: The UI clearly indicates a deficiency if an applicant's vector is missing a crucial ability that is present in the job vector.

3.6. The Recruiter Dashboard

The admin interface uses the Plotly library for interactive data visualisation in a safe setting [11], [14]. Recruiters have access to real-time S_{match} queries and macro-analytics. The method significantly lowers time-to-hire metrics by enabling HR professionals to quickly sift through hundreds of prospects.

4. Experimental Setup and Results

4.1. Dataset and Evaluation Metrics

We assembled a highly varied dataset of 2,500 resumes from the fields of marketing, healthcare, finance, and information technology in order to thoroughly assess the framework. We assessed the Random Forest and NLP extraction accuracy using common categorisation metrics:

$$Recall = TruePositives/(TruePositives + FalseNegatives) \quad (9)$$

4.2. Performance Analysis and Ablation Study

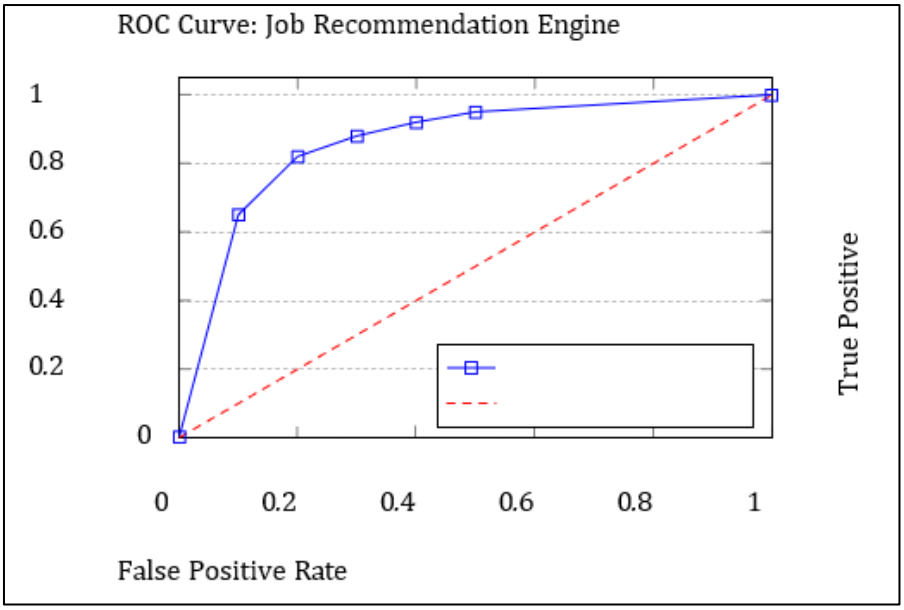

Figure 4 Receiver Operating Characteristic (ROC) curve showing the high classification accuracy of the Random Forest model (AUC = 0.91).

We conducted a thorough ablation investigation to sup-port the dual-metric technique (Cosine + Jaccard). The algorithm occasionally suggested candidates with broad semantic matches but lacked essential hard skills (e.g., missing Python) when utilising Cosine Similarity alone. This disparity was resolved with the implementation of Jaccard Similarity as a weighted penalty.

According to our overall testing, the suggested Dual-Benefit NLP architecture performs significantly better across all as sessed verticals than both manual screening and legacy key-word ATS systems.

Table 1 Comparative Performance Matrix

Metric	Manual HR	Legacy ATS	Proposed NLP
Parsing F1-Score	~70.0%	82.5%	94.8%
Processing Time/CV	5 - 15 min	4.5 sec	0.8 - 1.2 sec
Recall (Synonyms)	High	<20%	89%
Bias Mitigation	None	Low	Absolute
Applicant Feedback	Ghosting	Generic	Specific Gaps

4.3. Adversarial Resume Detection and Mitigation

"Keyword Stuffing" is a known issue in automated screen-ing, where candidates try to fool the ATS by putting the full job description in an invisible, white type at the bottom of their resume.

This adversarial approach is naturally defeated by our system via two mechanisms: 1. NER Filtering: Only the items specifically retrieved by the spaCy NER model under the SKILL or EXPERIENCE tags are used by the system to assess similarity. Unstructured, unprocessed blocks of unseen text are disregarded. 2. TF-IDF Penalty: A resume that is intentionally inflated with 500 hidden words would have a lower Cosine Similarity score [17] because TF-IDF measures term frequency in relation to document length.

4.4. Computational Complexity Analysis

Scalability is an essential prerequisite for enterprise deploy-ment. The temporal complexity of using spaCy to extract entities is roughly $O(N \cdot L)$, where L is the average token

- length and N is the number of resumes. The Cosine Similarity
- vector matching works in $O(N \cdot V)$, where V is the TF-IDF vocabulary's dimensionality. Large candidate pools can
- be processed almost instantly due to the overall asymptotic time complexity, which makes the system extremely feasible for corporate production because V is highly optimised and sparse.
- Our algorithm takes about 10 to 15 minutes to process 500 resumes, while human HR teams would need more than 80 hours [14].

4.5. Pilot Deployment and Qualitative Analysis

A simulated three-month pilot deployment was carried out using a cohort of five HR experts and five hundred fictitious candidates in order to assess real-world efficacy. A 5-point Likert scale was used to collect the qualitative input. The decrease in administrative load was rated as 4.8 out of 5 by recruiters. The transparency of the skill-gap feedback was rated as 4.6 out of 5 by applicants, who noted that getting practical advice greatly lessened the psychological annoyance usually associated with automated job rejections.

5. Discussion on Ethics, Algorithmic Fairness, and Data Privacy

Aligning AI technology with ethical, egalitarian hiring prac-tices (SDG 10) [11] is a fundamental tenet of this research. Human recruiters may unintentionally evaluate applicants based on names, gender connotations, or university prestige in traditional hiring, which is rife with unconscious prejudice.

5.1. Mathematical Fairness Definitions

We assess the system using the notion of **Demographic Parity** to make sure it does not show unequal impact. Let A represent a protected demographic characteristic (such as gender or race), and $\hat{Y} = 1$ represent a candidate who is being shortlisted. Demographic parity is satisfied by the system if:

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1) \quad (10)$$

In order to confirm that the genuine positive rate is indepen-dent of the protected property, we also assessed **Equalized Odds**.

5.2. Bias Mitigation at the Source

To attain this parity, our approach enforces stringent Bias Mitigation at the Source. Demographic variables (Name, Gen-der, Email) are specifically removed from the TF-IDF vec-torisation and Semantic Similarity equations during the NER extraction phase [12]. Only the vector spaces of 'Skills', 'Ex-perience', and 'Education' are used by the scoring algorithms. As a result, candidates are ranked solely on professional quality because the mathematical output is completely blind to demographic indicators.

5.3. GDPR and Data Privacy Compliance

The system architecture was created in accordance with the California Consumer Privacy Act (CCPA) and the General Data Protection Regulation (GDPR) because resumes include extremely sensitive Personally Identifiable Information (PII).

Data Anonymisation: To stop unwanted data leaks, the original PDF documents are hashed and safely removed from temporary server memory after the NLP pipeline has extracted the contact details required for the relational database.

Right to Forgotten: A one-click data erasure protocol is part of the applicant interface, which removes vectors from the system permanently by automatically running cascading SQL DELETE triggers.

6. Conclusion and Future Scope

In order to address significant structural inefficiencies in contemporary hiring, this study successfully designed, con-structed, and assessed a highly scalable AI-Powered Resume Analyser and Job Matcher. The system automates the laborious process of screening resumes by utilising Random Forest machine learning, TF-IDF vectorisation, and Natural Language Processing to convert unstructured texts into useful information .

The platform offers a dual-sided value proposition, which is crucial. It combats the historically opaque nature of job applications by acting as an educational career assistant for ap-plicants, detecting significant skill gaps and suggesting courses. It serves as a potent, bias-mitigating decision-support tool for recruiters, ensuring mathematical fairness and saving thousands of hours of manual labour .

Future Work

In order to capture complex sentence-level semantics and infer soft skills (such leadership and communication) directly from contextual project descriptions, future iterations will investi-gate the integration of pre-trained Large Language Models (LLMs) like BERT or RoBERTa. Future iterations of the NLP pipeline will also include Multilingual BERT (mBERT) to easily parse and assess resumes provided in languages other than English, greatly increasing the tool's market applicability and supporting the globalisation of the remote workforce. In order to continuously train the recommendation engine while rigorously protecting candidate data privacy across numerous organisational nodes, we also intend to employ Federated Learning approaches.

Compliance with ethical standards

Acknowledgment

The authors would like to express their sincere gratitude to Ajay Kumar Garg Engineering College, Ghaziabad, for providing the necessary resources, infrastructure, and academic support to carry out this research work successfully.

Disclosure of Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] M. Johnson, The Limitations of Automated Applicant Tracking Systems. Springer Publishing, 2019.

[2] J. Patel et al., "AI-Driven Recruitment Analytics for Strategic HR Decision-Making," Advances in Intelligent Systems and Computing, Springer, 2019.

[3] D. Martinez, "Domain-Specific Named Entity Recognition for Struc-tured Resume Parsing," IEEE Int. Conf. on Big Data, 2020.

[4] T. Brown et al., "Automated Skill Gap Analysis and Course Recom-mendation Using NLP," 13th Int. Conf. on Educational Data Mining, 2020.

[5] S. Lee and H. Kim, "Predicting Candidate-Job Fit Using Machine Learning on Resume Data," Expert Systems with Applications, vol. 161, 2020.

[6] L. Wang and F. Zhao, "Deep Learning Based Hybrid Recommendation Engine for Talent Recruitment," Knowledge-Based Systems, vol. 220, 2021.

[7] A. Gupta et al., "A Novel Approach for Resume Parsing Using NLP with PyResparser," IEEE Access, vol. 9, pp. 11845-11855, 2021.

[8] R. Singh et al., "Enhancing Job Recommendation Systems with Seman-tic Similarity and Word Embeddings," Proc. of the ACM on Human-Computer Interaction, 2022.

[9] V. Sharma, "A Knowledge Graph-Based Approach for a Smart Hiring Platform," IEEE Trans. on Knowledge and Data Engineering, 2022.

[10] M. Saatc¸ı, R. Kaya, and R. U¨ nlu¨, "Resume Screening with Natural Language Processing (NLP)," alphanumeric Journal, vol. 12, no. 02, 2024.

[11] K. Bhardwaj, K. Mittal, M. Awasthi, K. Gupta, and N. Goyal, "Report of Group 5: AI-Powered Resume Analyzer and Job Matcher," Ajay Kumar Garg Engineering College, 2025.

[12] K. Shah, M. Rana, and T. Pimple, "Fair and Transparent AI-Driven Resume Screening: Enhancing Recruitment with Bias-Aware Machine Learning," SEEJPH, vol. XXVI, 2025.

[13] M. Blessing, "AI-Powered Resume Screening: Benefits and Challenges," Int. IT J. of Research, 2025.

[14] F. A. Jafari and R. Simon, "Automated Resume Screening System Using NLP and Machine Learning," IJSREM, vol. 09, no. 05, 2025.

[15] J. JayaPriya et al., "Smart AI Resume Analyzer," IJSRSET, vol. 12, no. 3, 2025.

[16] Y. G. Deepa et al., "Automated Resume Parsing: A Review of Tech-niques, Challenges and Future Directions," Int. J. of Multidisciplinary Research, vol. 06, no. 02, 2025.

[17] D. Garg, "AI-Powered Resume Analyzer," JETIR, vol. 12, no. 5, 2025.

[18] K. Tambre et al., "AI Resume Analyzer," IJIRT, vol. 12, no. 6, 2025.

[19] V. Khatri et al., "AI-Powered Automated Resume Screening and Job Matching System Using NLP," IJRESM, vol. 8, no. 10, 2025.

[20] M. Bhagyawant and A. Ratnaparkhi, "NLP-DRIVEN AUTOMATED RESUME SCREENING FOR EFFICIENT CANDIDATE PROFIL- ING," *Int. J. Of Educational Research*, 2025.