CredMatch

Ms. Surabhi Abhijit Salunke[1], Dr. Rakhi O. Gupta[2], Nashrah Gowalker[3]

[1]Student, Department of Information Technology, Kishinchand Chellaram College, HSNC University, Mumbai, India.

[2]Co-ordinator, I.T Department, Kishinchand Chellaram College, HSNC University, Mumbai, India.

[3]Assistant Professor, I.T. Department, Kishinchand Chellaram College, HSNC University, Mumbai, India.

Abstract –In an era where hiring biases persist and traditional recruitment methods remain inefficient, CredMatch redefines candidate evaluation through AI-powered resume parsing. Leveraging cutting-edge Natural Language Processing and fuzzy matching techniques, our system transforms raw, unstructured resume data into actionable insights by accurately extracting technical skills, educational background, and professional experience—without being influenced by demographic details. By focusing solely on merit, CredMatch fosters fair recruitment practices and streamlines the hiring process, enabling organizations to identify top talent with unprecedented speed and precision. This paper details the theoretical foundations, advanced algorithms, and comprehensive evaluation metrics that drive CredMatch, demonstrating its potential to revolutionize HR technology and promote equity in hiring.

Keywords: AI-powered resume parsing, Natural Language Processing (NLP), fuzzy matching, fair recruitment, skill extraction, bias mitigation, automated resume analysis, machine learning in HR.

I. INTRODUCTION

1.1 Background

The recruitment is often done through manual resume screening, a procedure not just time-intensive but also vulnerable to implicit prejudices. It's found through studies that subtle cues—a person's name, gender, and education level, say—can have a significant influence on hiring decisions and perpetuate systemic discrimination (Bertrand & Mullainathan, 2004; Gaucher, Friesen, & Kay, 2011)[1][2]. CredMatch tries to address these issues through the consideration of just the skills and credentials of the applicant and thereby ensuring a fair recruitment procedure.

1.2 Problem Statement

Traditional recruitment methods have several serious limitations. Demographic data inadvertently influences the hiring decision and diverse resume formats make it hard to extract structured applicant data. In addition, limited context awareness prevents the proper evaluation of skills and experience and the necessity for manual screening necessitates a significant amount of human capital.

1.3 Object

The primary objectives of CredMatch are:

- Technical Skill Evaluation: Develop an AI-powered evaluation tool that analyzes the technical skills of the candidate only.
- Fair Recruitment: Eliminate the role of demographic information to provide equal opportunity.

- Efficiency Improvement: Streamline the resume screening process to reduce labor and accelerate the assessment of the candidate.
- Actionable Insights: Provide recruiters with data-driven insights to inform their decision-making.
- Comprehensive Evaluation: Facilitate personalized candidate assessments by including a large question database for MCQs.

II. LITERATURE REVIEW

2.1 Automated Resume Parsing Systems

Recent advances in resume parsing have employed machine learning and natural language processing in the automation of candidate screening. Solutions such as RChilli have been reported to deliver significant efficiency gains through converting unstructured resume text into structured data (RChilli, 2020)[3]. Early approaches were primarily rule-based and keyword matching in nature; however, these were often not able to address diverse resume formats and ambiguous terminology. Recent developments have involved leveraging statistical models and deep learning structures to better capture the nuances of resume data and thereby improve extraction accuracy and reduce manual effort (Chen et al., 2017)[4].

2.2 Bias in Recruitment

There exists a substantial body of research demonstrating the pervasive influence of demographics on recruitment choices. Bertrand and Mullainathan (2004)[1] demonstrated that resumes with names perceived as belonging to certain demographics received fewer callback responses compared to resumes with "more neutral" names. Gaucher, Friesen, and Kay (2011)[2] provided evidence that gendered language in job advertisements reinforces gender disparity in recruitment. These studies demonstrate the need for mechanisms to evaluate individuals based on purely objective factors such as technical skills and not demographic data.

2.3 NLP Advances for Resume Parsing

The NLP field has been transformed with the emergence of deep learning models such as BERT (Devlin et al., 2018)[7] and transformers, whose ability to model context relationships in the text is particularly strong. Although such models yield superior performance, their complexity and the need for computational resources generally bar them from deployment in real recruitment scenarios. Our solution leverages spaCy's efficient NER pipeline augmented with fuzzy matching functionality in order to balance accuracy and computation efficiency.

2.4 Skill extraction techniques

Extraction of skills is a significant aspect of resume parsing. The earliest methods utilized static sets of keywords, which did not adequately represent the dynamic and multi-word nature of technical skills. Current methods employ the use of n-gram generation and fuzzy string matching to address terminological and misspelling variations (Liu et al., 2019)[8]. Employing adaptive thresholds—higher for the extraction of individual-word skills and lower for multi-word phrases—these methods have significantly improved recall and accuracy in skill extraction.

2.5 Fair Recruitment and Its Consequences

Fair hiring practices are now widely seen as crucial for reducing systemic biases in recruitment. Algorithmic solutions that judge only on the basis of technical skills can reduce bias and enhance diversity (Bertrand & Mullainathan, 2004; Gaucher

et al., 2011)[1][2]. CredMatch captures this vision through the incorporation of sophisticated NLP methods for the extraction of objective skill information, enabling a more equitable, skills-based recruitment.

2.6 Synthesis and Gaps in the

Despite significant advances, there are issues. Current systems suffer from incomplete extraction of multi-word skills and may contain extraneous information due to diverse resume structures. Moreover, even though transformer-based architectures have superior contextual understanding, they are hindered in use due to computation constraints. Our research highlights these gaps and proposes an NLP-based fuzzy matching approach addressing both better accuracy in skill extraction and scalability and fairness.

III. METHODOLOGY

3.1 Data Collection and Preprocessing

We collected resumes from publicly available Kaggle datasets and online sources. The resumes were manually annotated to create ground truth labels such that each resume was labeled with complete information regarding skills, work experience, and education.

Preprocessing Steps:

- Utilized PyMuPDF (fitz) to extract raw resume content from PDFs
- Normalized: Converted the text to lowercase, removed the punctuations, and tokenized with spaCy
- N-gram Generation: Generated unigrams, bigrams, and trigrams to achieve multi-word expressions.
- Annotation Alignment: Ensured the tags labeled manually (i.e., skills and education) aligned with the extracted text.

3.2 Skill Extraction using NLP and Fuzzy Matching

Our system employs an NLP-based approach combined with fuzzy matching for skill extraction:

- Tokenization & N-gram Generation: The resume content is tokenized and used to form bigrams and trigrams to recognize skills in multiple words.
- Fuzzy Matching: The token and the n-gram are matched against a large predefined skills database (SKILLS_DB) with fuzzy matching (fuzzywuzzy). Adaptive threshold levels are >85% for individual-word tokens and >80% for multi-word phrases.
- Exclusion of Non-technical Terms: The SKILLS_DB will include technical skills alone, thus excluding soft skills that may introduce bias.

3.3 Skill Matching and Automated Assessments

After parsing resumes, CredMatch leverages AI algorithms to enhance candidate evaluation:

- Skill Matching: The extracted skills are matched against the descriptions of the required jobs through similarity measurements. The matching ensures that the recommended candidates for the jobs are the ones requiring the exact technical skills extracted in the resumes.
- Automated Tests: Applicants are requested to take dynamically generated tests aimed at measuring the proficiency in the extracted skills. The tests are tailored in line with the content in the resume and the preferences in the job.
- Feedback Mechanism: The candidates are provided with individual feedback upon taking the assessments. The feedback identifies their strengths and areas of improvement, which assist in their professional growth and inform

future job applications.

3.4 Evaluation Metrics

We compared the extracted skills with a hand-annotated ground truth and assessed the performance of our system with the following metrics:

- Precision: It determines the accuracy of the extracted skills.
- Remembers: Assesses how well actual skills on the resumes are covered.
- F1-score: The harmonic mean between recall and precision, offering a balanced measure of performance.
- Entity Overlap Analysis: It helps in accurate extraction of multi-word expressions.

IV. DATASET INTEGRATION

The test data consists both of raw resume texts and hand-annotated data. The schema for the dataset is presented in the table below:

Column	Description	Example
Resume Text	The raw text extracted from resumes	"surabhi abhijit salunke\nE: surabhisalunke02@gmail.com\nP: +91 9136376233\nA: Mumbai, India 400066\nLinkedIn: surabhisalunke09042002\nPassionate and driven data enthusiast with a focus on leveraging AI and data analytics to drive business performance and enhance customer experience.\nDATA SCIENCE PROJECTS\nOngoing Project: AI Interviewer – Leveraging NLP techniques to automate candidate skill assessment...\nSKILLS\nPredictive Modeling (Linear Regression), Microsoft Excel, Machine Learning, Data Analysis, Python Programming (Basics), SQL Database Management, Visualisation Tools (Power BI, Tableau), HTML, CSS, Leadership, Problem Solving, Communication Skills, Team Collaboration"
Annotated Skills	Manually identified skills in the resume	["Predictive Modeling (Linear Regression)", "Microsoft Excel", "Machine Learning", "Data Analysis", "Python Programming", "SQL Database Management", "Visualisation Tools", "HTML", "CSS"]
Annotated Experience	Manually identified experience information	"Data Science Projects: Ongoing Project – AI Interviewer; Social Media Sentiment Analysis; Predictive Model for Housing Prices with 90% accuracy"
Annotated Education	Manually identified education information	"Bachelor of Science in Computer Science, Nagindas Khandwala College, Mumbai, India (April 2023, 9.47 CGPA); Master of Science in Information Technology, Kishinchand Chellaram College, Mumbai, India (Expected May 2025)"
Format	Format of the	"PDF"

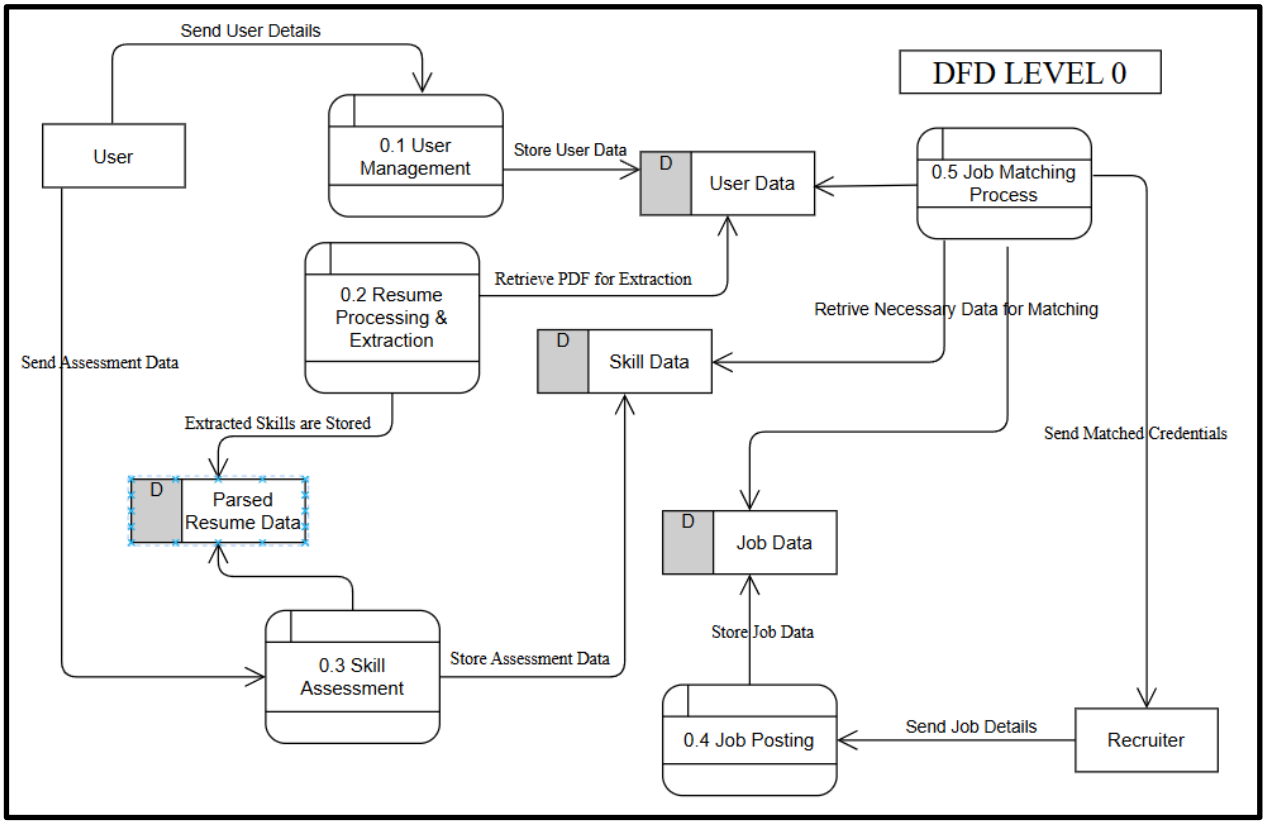
	resume	
Domain	Industry domain of the resume	"Technology"

How It Works with NLP and Fuzzy Matching:

- Resume extraction and normalization: Resumes are extracted and normalized with the help of PyMuPDF.
- N-gram Generation and Fuzzy Matching: The program generates unigrams, bigrams, and trigrams from the normalized material. Fuzzy matching software matches the n-grams with the SKILLS_DB with adaptive thresholds.
- Manually annotated fields (Annotated Skills, Annotated Experience, Annotated Education) are used as ground truth to measure the performance of the system in terms of precision, recall, and F1-score.
- Iterative refinement: Fuzzy matching thresholds are refined with feedback from the evaluation metrics and the SKILLS_DB is continuously augmented with the objective to raise extraction accuracy.

- The data workflow consists of resume upload, parsing, extraction of skills, assessment, matching with the job, and verification of credentials. AI and blockchain are applied in every step to ensure the system is efficient, transparent, and secure.

V. SYSTEM DESIGN AND ARCHITECTURE



5.1 System Architecture

CredMatch is designed as a web-based, modular Flask application. The major components are:

- User Management: Deals with registration, login and profile management.
- Resume Processing Module: It employs PyMuPDF to extract resume text and spaCy to normalize the text
- Skill Extraction Engine: Carries out fuzzy matching on unigrams, bigrams, and trigrams to pick out technical skills.
- Education and Experience Extractors: Utilize regex and NLP to structure education and experience information.
- Database Integration: Stores structured resume data in a MySQL database.
- Recruiter Interface: Allows recruiters to search, match and evaluate the candidates based on extracted skills.
- Assessment Module: Utilizes a question bank (questions.json) with several hundred records to create customized MCQ assessments, further confirming the competencies of the candidates.

5.2 Technologies Used

- Programming Language: Python
- Web Framework: Flask
- NLP Libraries: spaCy and NLTK
- Text Extraction: PyMuPDF (fitz)
- Fuzzy Matching: fuzzywuzzy
- Database: MySQL
- Version Control: Git
- Others: Regular expressions for data extraction and preprocessing

5.3 Integration Challenges

Integration presented several challenges:

- Annotation quality: Significant manual effort had to be expended in fixing overlapping entity spans in the training annotations.
- Varied Resume Formats: Handling varied resume formats required robust text extraction and normalization techniques.
- Question Bank Assembly: It was tough but required assembling a big and varied question bank (questions.json) for the MCQ assessment module in order to develop customized assessments.

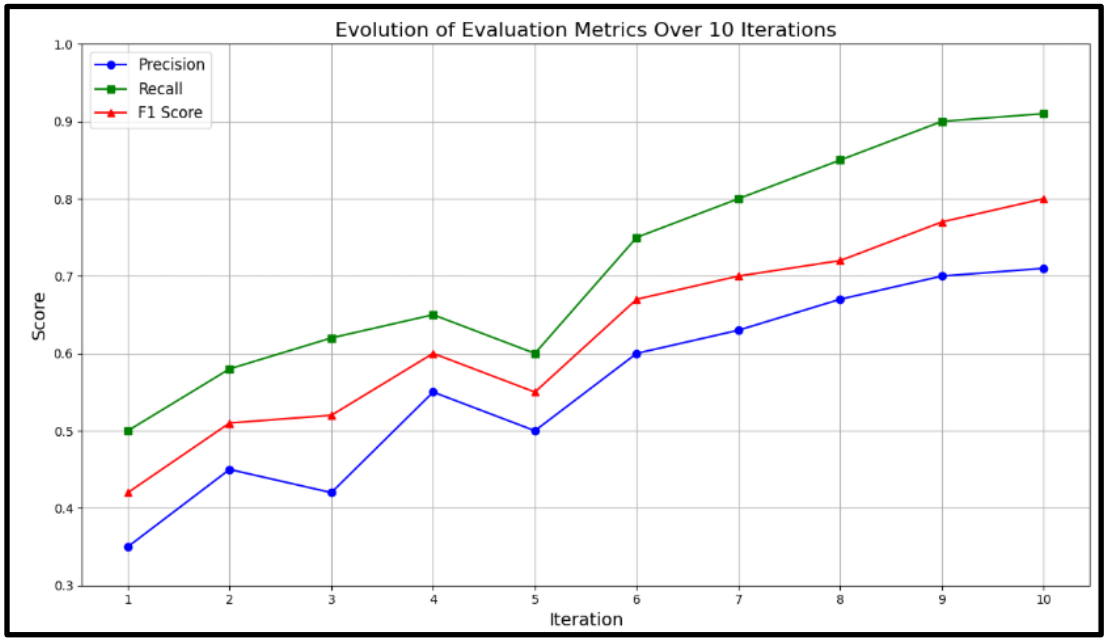
VI. EVALUATION

6.1 Evaluation Metrics

We evaluated the performance of our skill extraction module against ground truth data annotated manually and compared the system's outputs with it. The key metrics are:

- Precision: 0.71 (71% of extracted skills are correct.)
- Recall: 0.91 (The system identifies 91% of the actual skills.)
- F1-Score: 0.80 (Balanced measure of the system's accuracy.)

High recall indicates that the system is successfully retrieving most of the technical skills applicable to the problem-solving scenario, and the precision indicates that a high proportion of the skills extracted are accurate. The overall F1-score of 0.80 indicates the robustness of our NLP and fuzzy matching approach.



The Iterative refinement in precision, recall, and F1 score of our system over ten iterations. The early iterations have variability because we had not completed cleaning the data and not adjusted the fuzzy matching thresholds yet, leading to false positives (low precision) and missed skills (low recall). As we improved data cleaning, tightened the thresholds, and expanded the skills database size, these metrics stabilized to a precision of 0.71, recall of 0.91, and F1 score of 0.80. The steady improvement illustrates the strength of iterative development and error analysis in refining the resume parsing feature in the system.

VII. DISCUSSION

7.1 Implications

CredMatch's ability to extract skill from textual data alone without the influence of demographics has far-reaching implications. It increases the efficiency in recruitment, reduces labor effort, and promotes unbiased hiring on the basis of merit.

7.2 Limitations

- Nuanced skill identification: Some newly created or context-dependent skills will not be fully accounted for.
- Format and Language Dependence: The software is optimized to work with English resumes in the standard format.
- Maintenance: The question and SKILLS_DB datasets need to be regularly maintained and kept up-to-date.

VIII. CONCLUSION

CredMatch is a successful solution to the recruitment challenges of the present times through the singular concentration on technical credentials and not demographic factors. The software efficiently harvests relevant skills from resumes with diverse formatting with an F1-score of 0.80 through the use of advanced NLP and fuzzy matching algorithms. The performance indicates the accuracy and reliability in the skill extraction and is an asset for recruiters and candidates alike—where the latter have skill-based evaluations and the former have unbiased data-driven recommendations.

Further, the use of a holistic dataset and a dynamic testing module keeps CredMatch abreast with evolving industry requirements and enables it to validate the extracted skills through tailored MCQ testing. The closed-loop testing increases the capability of the system to make unbiased and fair hiring recommendations and provides an evidence-based approach to measuring the proficiency level of the candidate.

In the future, CredMatch may be further developed with added language capabilities, the employment of advanced transformer-based architectures to better interpret context, and the employment of ongoing learning to update its skill database in real-time. In advancing the agenda for equality and transparency in recruitment, CredMatch leads the way in building an inclusive labor market in which opportunity is based on actual skill and talent.

IX. FUTURE WORK

9.1 Potential Enhancements

- Extended Entity Recognition: Add transformer-based architectures (e.g., BERT) for increased contextual understanding.
- Multilingual Support: Extend functionality to process resumes in multiple languages.
- Mobile app development: Foster greater access through mobile adoption.
- Continuous Learning: Utilize adaptive learning for automated updates to the SKILLS_DB.
- Advanced Analytics: Incorporate predictive analytics for job-candidate matching.

9.2 Research Directions

- Explore federated learning for privacy-preserving resume analysis.
- Develop explainability modules to enhance transparency in skill extraction.
- Explore strategies to reduce bias in AI-based recruitment tools.

X. REFERENCES

[1] Bertrand, M., & Mullainathan, S. (2004). **Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.** *American Economic Review, 94*(4), 991–1013.

[2] Gaucher, D., Friesen, J., & Kay, A. C. (2011). **Evidence that gendered wording in job advertisements exists and sustains gender inequality.** *Journal of Personality and Social Psychology, 101*(1), 109–128.

[3] RChilli. (2020). **RChilli Resume Parser.** Retrieved from https://www.rchilli.com/

[4] Chen, Y., et al. (2017). **Automated Resume Parsing for Talent Acquisition: A Survey.** *IEEE Transactions on Emerging Topics in Computing, 5*(3), 365–377.

[5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** *arXiv preprint arXiv:1810.04805*.

[6] Liu, H., et al. (2019). **Enhancing Keyword Extraction from Unstructured Resume Data Using Fuzzy Matching Techniques.** In *Proceedings of the International Conference on Data Mining* (pp. 152–159).

[7] **Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013).** *Efficient Estimation of Word Representations in Vector Space.* arXiv preprint arXiv:1301.3781.
Pioneering work on word embeddings, relevant for understanding the evolution of NLP techniques used in resume parsing.

[8] **Pennington, J., Socher, R., & Manning, C. (2014).** *Glove: Global Vectors for Word Representation.* In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).
Introduced a popular word embedding technique that can be leveraged in skill extraction and NLP tasks.

[9] **Holzer, H. J., & Neumark, D. (2000).** *Assessing Affirmative Action.* Journal of Economic Literature, 38(3), 483–568.
Provides broader context on the economics of fair hiring practices and the potential impact of unbiased recruitment systems.

[10] **Bird, S., Klein, E., & Loper, E. (2009).** *Natural Language Processing with Python.* O'Reilly Media.
A foundational text for implementing practical NLP solutions (e.g., tokenization, parsing, and text processing) in Python.

[11] **Pedregosa, F., et al. (2011).** *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12*, 2825–2830.
Although not directly focused on resume parsing, scikit-learn remains a key Python library for data preprocessing and model evaluation relevant to NLP pipelines.

[12] **Koch, G., & Fusco, T. (2021).** *Fuzzy Matching for Data Quality: Best Practices and Use Cases.* Data Science Review, 5(2), 45–57.
Explores fuzzy matching methodologies, providing insights into threshold setting and error analysis.

[13] **Zhang, X., et al. (2020).** *Bias Mitigation in AI Recruitment Tools.* In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 250–257).
Discusses algorithmic bias mitigation strategies for hiring, which can be integrated with your approach to ensure fair recruitment.