International Journal of Multidisciplinary Research and Growth Evaluation.

Automated Resume Parsing: A Review of Techniques, Challenges and Future Directions

Y Gyana Deepa [1*], Ankathi Sindhu [2], Alakuntla Shruthi [3], Bitla Neha [4]
[1] Asst. Professor, ADCE Dept., Stanley College of Engineering and Technology for Women, Hyderabad, Telangana, India
[2, 3, 4] Stanley College of Engineering and Technology for Women, Hyderabad, Telangana, India

* Corresponding Author: **Y Gyana Deepa**

Article Info
ISSN (online): 2582-7138
Volume: 06
Issue: 02
March-April 2025
Received: 21-02-2025
Accepted: 17-03-2025
Page No: 1065-1069

Abstract
Recruitment procedures have been completely transformed by the quick development of artificial intelligence (AI) and natural language processing (NLP), with automated resume parsing emerging as a crucial talent acquisition tool. In order to speed up the candidate screening process, resume parsing entails extracting, organizing, and evaluating information from resumes. A thorough examination of several resume parsing strategies is given in this review study, including rule-based strategies, machine learning models, and deep learning-based strategies like Named Entity Recognition (NER) and Transformers. We also assess well-known resume parsing tools according on their accuracy, methods, and usefulness. The study also addresses important issues like inconsistent data, multilingual parsing, and moral dilemmas with AI-driven hiring. We conclude by discussing potential avenues for future research, highlighting the necessity of increased precision, bias reduction, and greater Applicant Tracking System (ATS) integration. Researchers and developers looking to improve resume parsing technology for more impartial and effective recruiting procedures might use this review as a starting point.

Keywords: Resume Parsing, Natural Language Processing (NLP), Named Entity Recognition (NER), Applicant Tracking System (ATS)

1. Introduction

Companies receive thousands of resumes for job opportunities in today's fast-paced recruitment environment, making manual resume screening difficult and ineffective. Automated resume parsing, which uses natural language processing (NLP) and artificial intelligence (AI) to extract and organize candidate data for efficient recruiting, has become a vital option. Resume parsers assist recruiters by automatically classifying important information from resumes in different formats (PDF, DOCX, TXT), including education, work experience, skills, and personal information. Resume parsing methods have developed over time, moving from straightforward rule-based strategies to sophisticated machine learning and deep learning models. While modern approaches use Named Entity Recognition (NER), Support Vector Machines (SVM), and deep learning models like BERT to improve accuracy, traditional rule-based methods rely on predefined patterns and regular expressions. Notwithstanding these developments, resume parsing still has a number of issues, such as inconsistent data formatting, multilingual text processing, and ethical challenges with AI-based hiring biases. This review offers a thorough examination of various resume parsing methods, assessing both their advantages and disadvantages. We also examine current resume parsing tools, their relative effectiveness, and typical issues encountered by the sector. We conclude by talking about the future of resume parsing research, with an emphasis on enhancing integration with Applicant Tracking Systems (ATS), decreasing bias, and increasing accuracy.

2. Literature Review

The tabular format of the literature review summarizes research papers according to their title, year, methods, benefits, and drawbacks. Rule-based methods, machine learning models, and deep learning-based techniques like Named Entity Recognition (NER) and Transformers are among the studies examined.

By examining these techniques, we aim to provide insights into their effectiveness, accuracy, and challenges, such as handling unstructured data, multilingual parsing, and bias in AI-driven hiring systems.

Table 1

Title	Year	Approach	Advantages	Disadvantages
Recommendation for Jobs and Resume Analyzer Using NLP	2024	NLP techniques like Named Entity Recognition (NER), cosine similarity. Pyre sparser library	Resume optimization.	Parsing limitations.
Enhancing Job Recommendation Systems Using Machine Learning	2024	Collaborative filtering, content-based filtering, and deep learning techniques, Graph based methods	Tailor jobs through skill-matching.	Privacy concerns and resource demands.
Automated Resume Analysis & Skill Suggesting Website	2024	NLP for text extraction, Resume Parser, Semantic Search	Skill suggestions, scalability, automation.	Library constraints.
Resume Analyzer Using NLP	2024	NLP for text extraction, ML classifiers (SVM, Logistic Regression), Cosine Similarity	Efficient screening, personalized results.	Training data bias.
Resume Parser	2024	Machine Learning classifiers (SVM, Decision Tree), Optical Character Recognition (OCR)	Enhanced hiring, faster screening, ATS integration.	Parsing challenges and model bias.
Resume Parser Using Machine Learning	2024	NLP techniques (Regex, NLTK, Spacy)	Boosts efficiency, scalable flexibility, personalization.	Format complexity demands continuous updates.
Smart Resume Analyzer	2023	NLP techniques: Cosine Similarity, TF-IDF, NER and KNN	Exceptional ranking precision.	Training limitations and processing delays.
Resume Analysis Using Machine Learning and NLP	2023	NLP techniques like bigram, trigram, text classification ML models like KNN and SVM	Personalized feedback and skill guidance.	Formatting dependency.
Resume Building Based on it's Compatibility with Job Description	2023	NLP, ML classifiers (Logistic Regression, Decision Tree)	Custom templates reduce bias.	Refinement and integration delays.
Resume Screening Using TF-IDF	2023	NLP for text extraction, TF-IDF for ranking, Cosine Similarity for comparison	Instant results with precision processing.	Dataset scalability requires optimization.
Resume Parser Using ML and NLP	2023	Combines Machine Learning models with NLP techniques, including NER	Versatile parsing with precise classification.	Extensive labeling and format challenges.

3. Resume Parsing Techniques:

To extract structured information from unstructured resume documents, automated resume parsing uses a variety of techniques. The main resume parsing strategies used today are examined in this section.

3.1 Rule-Based Parsing:

Rule-based parsing is one of the earliest methods for extracting information from resumes. It relies on predefined patterns, templates, and keyword matching to identify specific entities such as names, contact details, education, and work experience. Regular Expressions (Regex) and phrase matching are commonly used techniques in this category.

How It Works?

Identifies patterns (such as phone numbers, email addresses, and job titles) using static rules.
It employs hardcoded templates to identify particular resume formats, uses dictionary-based matching to extract job-related keywords and competencies.

Advantages

- Simple and easy to implement.
- Works well for resumes that follow a structured format.
- Fast processing with low computational cost.

Limitations

- Struggles with unstructured and complex resume formats.
- Cannot handle synonyms, variations in job titles, or multilingual resumes.
- Requires frequent manual updates as resume styles change.

3.2 Machine learning-based parsing:

Machine learning-based resume parsing improves on rule-based approaches by using statistical models to categorize and extract pertinent information. These models are trained on labeled resume datasets, which enables them to discover trends and increase accuracy rather than depending on predetermined criteria. Naïve Bayes, a probabilistic model for text classification and keyword recognition, Random Forest & Decision Trees, which work well for structured prediction and entity recognition, and Support Vector Machines (SVM), which classify text into predefined categories like job titles and skills, are examples of common algorithms.

How It Works?

Data preprocessing is the first step in the machine learning-based resume parsing process, during which resumes are transformed into a structured format using methods including tokenization, stemming, and stopword removal. In order to comprehend textual patterns, feature extraction then finds important terms, word frequencies, n-grams, and uses Part-of-Speech (POS) tagging. During the model training phase, labeled resumes are used to train an ML classifier to correctly classify various portions. The trained model is then applied to fresh resumes in the prediction and extraction phase to extract pertinent data including experience, skills, and job titles.

Advantages

- More adaptable than rule-based methods.
- Can generalize across different resume formats.
- Learns from data patterns instead of relying on predefined rules.

Limitations

- Requires large labeled datasets for training.
- Accuracy is dependent on the quality of training data.
- Limited in understanding deep semantic meaning and context.

3.3 Deep learning-based parsing:

Neural Networks (NNs) and Natural Language Processing (NLP) approaches are used in deep learning-based resume parsing to extract structured information more accurately and contextually. These models, in contrast to conventional techniques, can handle intricate resume forms, comprehend word relationships, and identify synonyms. Named Entity Recognition (NER) models recognize particular entities like names, job titles, and abilities, whereas Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) are useful for sequence prediction and entity extraction. Furthermore, by utilizing contextual awareness, transformer-based models such as BERT, GPT, SpaCy, and Flair improve the accuracy of extraction.

How It Works?

The deep learning-based resume parsing process begins with preprocessing and tokenization, where text is converted into structured tokens using word embeddings. Next, Named Entity Recognition (NER) is applied to identify key entities such as names, addresses, job titles, and skills. To enhance accuracy, context understanding models like BERT analyze word relationships and contextual meaning. Finally, in the extraction and classification phase, the model predicts labels for each resume section, such as experience, education, and skills, ensuring precise information retrieval.

Advantages

- High accuracy in extracting structured information.
- Can handle unstructured, multilingual, and complex resumes.
- Improves over time with more data and retraining.

Limitations

- Computationally expensive and requires high processing power.
- Needs large datasets for proper training.
- Difficult to interpret how deep learning models make decisions (black-box problem).

3.4 Hybrid Approaches:

To increase precision and flexibility, hybrid resume parsing blends rule-based techniques, machine learning, and deep learning. A lot of contemporary resume parsers employ a hybrid strategy to balance out the drawbacks of each technique while utilizing its advantages.

How It Works?

While machine learning models categorize various resume sections according to learned patterns, rule-based parsing works well for straightforward extractions like emails and phone numbers. By increasing context awareness and precisely extracting complex entities like job titles, skills, and experience, deep learning techniques further improve parsing.

Advantages

- Balances accuracy and efficiency.
- Can handle structured and unstructured resumes.
- Reduces dependency on large training datasets.

Limitations

- More complex to develop and maintain.
- Requires integration of multiple technologies.

4. Evaluation Metrics:

4.1 Precision, Recall, and F1-Score

Precision: Evaluates the extracted entities' accuracy (e.g., job title, abilities). It shows the proportion of retrieved results that are accurate.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall: Measures how many relevant entities were successfully extracted. A high Recall ensures that the parser does not miss critical resume details.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

F1-Score: The harmonic mean of Precision and Recall, providing a balanced evaluation of the parser’s performance.

$$F1 - score = \frac{2 \times\ Precision \times Recall}{Precision + Recall}$$

4.2 Processing Speed & Scalability

A resume parser's efficiency is measured by how fast it can process a lot of resumes without sacrificing accuracy. For large-scale recruitment platforms and Applicant Tracking Systems (ATS), which process thousands of applications every day, processing speed and scalability are essential.

Factors affecting processing speed

- **Algorithm Complexity:** Deep learning models are slower but more accurate than rule-based parsers, which are quicker but less adaptable.
- **Computational Resources:** Processing speed is increased by cloud-based parsing programs like AWS and Google Cloud.
- **Data Preprocessing Time:** The total processing time is increased by tokenization, entity recognition, and feature extraction.

Scalability Challenges

- Handling batch processing of thousands of resumes simultaneously.
- Maintaining speed while increasing parsing depth (extracting detailed candidate insights).
- Managing real-time processing in high-volume job applications.

4.3 Handling of Different Resume Formats (PDF, DOCX, etc.)

Resumes can be submitted in a number of formats, such as HTML, TXT, DOCX, PDF, and images (scanned resumes).

No matter the format, a strong resume parser should reliably extract information.

4.3.1 Common challenges in format handling

a) **PDF Parsing Issues**
The ability of the Portable Document Format (PDF) to preserve layout consistency across various devices makes it a popular resume format. However, there are particular difficulties with PDF parsing. Some PDFs store text as images, which makes direct text extraction challenging, in contrast to text-based formats. For PDFs with selectable text, standard text extraction libraries (such as PyPDF2 and PDFMiner) perform admirably; however, they are ineffective when working with PDFs that contain images. In these situations, text must be extracted from images using Optical Character Recognition (OCR) techniques. The accuracy of the information that is extracted, however, may be impacted by OCR-based parsing errors brought about by poor image quality, font variations, or intricate layouts. Furthermore, multi-column resumes or graphical components may make extraction even more difficult, necessitating the use of sophisticated NLP techniques in order to accurately reconstruct structured data.

b) **DOCX Variability**
Tables, columns, headers, footers, embedded images, and different text styles make it difficult to parse DOCX resumes. In contrast to PDFs, which embed text as a static structure, DOCX files allow for dynamic formatting, which may cause parsing errors. For example, some resumes use sidebars for contact information, while others list work experiences in tables rather than paragraphs. Non-linear structures are frequently difficult for rule-based or keyword-matching methods to handle; in order to meaningfully interpret content, machine learning and natural language processing models are needed. Furthermore, nested elements (such as bullet points inside tables) may make extraction more difficult, requiring the use of specialized parsers or DOCX-specific libraries (like Python-docx or Apache Tika) in order to efficiently extract structured data.

c) **Scanned Resumes**
Many resumes are submitted as scanned images (JPEG, PNG, or image-based PDFs), and in order to transform the text into a machine-readable format, OCR-based extraction is needed. When it comes to digitizing such resumes, OCR tools like Tesseract, Google Vision API, and ABBYY FineReader are essential. However, skewed alignment, background noise, font clarity, and image resolution all affect OCR accuracy. Resumes with decorative fonts or low-quality scans may be misinterpreted, which could lead to entity extraction errors (e.g., misreading "Senior Manager" as "Senlor Manager"). Additionally, handwritten text, logos, and highly formatted documents may be difficult for OCR to read accurately, necessitating post-processing methods. More sophisticated AI-based OCR models use deep learning methods to better identify formatting, context, and text structure, which lowers conversion errors.

d) **Multi-Column Layouts**
Multi-column formats are used in many contemporary resumes to improve readability and appearance. The sequential processing of content by standard text extraction tools frequently results in the merging of irrelevant sections or a misinterpretation of the reading order. A two-column resume, for example, might list a candidate's work experience on the right and education on the left. However, a parser might mistakenly combine text from both columns, resulting in a misclassification. Advanced NLP techniques like document segmentation and layout detection are necessary when handling multi-column layouts. To increase parsing accuracy, some AI-powered resume parsers employ Transformer-based models (such as LayoutLM and Doc2Vec) to comprehend document structure and reading order. Before using text extraction algorithms, computer vision-based methods also assist in accurately separating various sections and detecting text alignment.

4.3.2 Technology used for format handling

PDF Parsing: Programs that handle text positions and tabular data, such as PyMuPDF, PDFPlumber, and Adobe PDF API, effectively extract text from structured PDFs. But in order to turn images into text, scanned PDFs need OCR-based techniques like Tesseract, AWS Textract, or Google Vision OCR. For intricate layouts, hybrid methods that combine machine learning and text extraction increase accuracy.

DOCX Parsing: Word documents can have their text, tables, and metadata extracted using libraries like Apache Tika and Python-docx. Data is stored in XML format in DOCX files, which facilitates structured extraction. Deep learning models and sophisticated pre-processing are necessary for handling complex layouts in order to improve accuracy.

OCR for Scanned Resumes: Scanned resumes can be converted into text using Tesseract OCR, Google Vision AI, and ABBYY FineReader. AI-driven solutions provide improved accuracy, multi-column support, and language detection. The quality of extraction is further improved by post-OCR methods like error correction and entity recognition based on natural language processing.

5. Conclusion and future scope:

Resume parsing has changed dramatically, moving from rule-based systems to AI-driven solutions that use machine learning and deep learning to increase accuracy. This review emphasized important resume parsing techniques, such as rule-based, machine learning-based, deep learning-based, and hybrid approaches, as well as the difficulties in handling a variety of resume formats, including PDF, DOCX, and scanned images. The evaluation metrics discussed—Precision, Recall, F1-score, processing speed, and scalability—are essential for evaluating the efficacy of contemporary resume parsers. Although AI-powered parsers perform better than traditional methods, they still struggle with complex layouts, multi-column formats, and unstructured content. Additionally, bias in AI-based screening is a serious concern because models trained on biased datasets may unintentionally favor some candidates over others.In order to better understand resume semantics and enhance entity recognition, future research should concentrate on improving accuracy using sophisticated NLP models like Transformers (BERT, GPT, and T5). In order to ensure impartial and equitable candidate evaluation, efforts must also be made to reduce bias in AI-based hiring. This can be achieved by training models on representative and diverse datasets. The smooth integration of resume parsers with applicant tracking systems (ATS), which facilitates effective candidate screening, ranking, and profile enrichment, is another crucial area that needs work. Future systems should also be able to handle complex and unstructured resume formats more effectively. They should use deep learning and computer vision techniques to extract information from unconventional layouts, infographics, and graphical resumes. These issues can be resolved to make next-generation resume parsers more reliable, effective, and fair, which will change the hiring process in the long run.

6. References

1. Gharat J. Recommendation for jobs and resume analyzer using NLP. International Journal of Research Publication and Reviews. 2024 Jan;5(1):1328–34. Available from: https://www.ijrpr.com/uploads/V5ISSUE1/IJRPR13639.pdf
2. Sharma R, *et al*. Enhancing job recommendation systems through machine learning: a comprehensive analysis of skill sync job recommendation. International Journal of Research Publication and Reviews. 2023;4(10):948–55. Available from: https://ijrpr.com/uploads/V4ISSUE10/IJRPR10020.pdf
3. Roy PK, Chowdhary SS, Bhatia R. A machine learning approach for automation of resume recommendation system. Procedia Computer Science. 2020;167:2310–9. doi:10.1016/j.procs.2020.03.284
4. Kashif M, Parimal Kumar KR. Resume parser using NLP. International Journal of Advanced Research in Computer and Communication Engineering. 2024 Sep;13(9):33–6. doi:10.17148/IJARCCE.2024.13905
5. Pokharel P. Resume parser using NLP. ResearchGate. [Preprint]. 2022. Available from: https://www.researchgate.net/publication/361772014_RESUME_PARSER
6. Brindashree BV, Pushphavath TP. HR analytics: resume parsing using NER and candidate hiring prediction using machine learning model. Semantic Scholar. [Preprint]. Available from: https://www.semanticscholar.org/paper/HR-Analytics%3A-Resume-Parsing-Using-NER-and-Hiring-Brindashree
7. Sowjanya Y, *et al*. Smart resume analyser. International Journal of Research in Engineering and Science. 2023 Mar;11(3):409–18. Available from: https://www.ijres.org/papers/Volume-11/Issue-3/1103409418.pdf
8. Reza MT, Zaman MS. Analysing CV/resume using natural language processing and machine learning. ResearchGate. [Preprint]. 2022. Available from: https://www.researchgate.net/publication/365299910_Analyzing_CVresume_using_natural_language_processing_and_machine_learning
9. Jivtode A, *et al*. Resume analysis using machine learning and natural language processing. International Research Journal of Modernization in Engineering, Technology and Science. 2023 May;5(5):5757–61. Available from: https://www.irjmets.com/uploadedfiles/paper/volume_5/issue_5_may_2023/45601/final/fin_irjmets1685017650.pdf
10. Lokesh S, *et al*. Resume screening and recommendation system using machine learning approaches. Computer Science & Engineering: An International Journal. 2022 Feb;12(1):1–6. doi:10.5121/cseij.2022.12101
11. Mohd Sadiq SZ, Ayub JA, Narsayya GR, Ayyas MA. Intelligent hiring with resume parser and ranking using natural language processing and machine learning. 1Library.net. [Preprint]. Available from: https://1library.net/document/y43jn00z-intelligent-hiring-ranking-natural-language-processing-machine-learning.html.