International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03365
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Automated Resume Parsing using Name Entity Recognition

Dr. S. A. Bhavsar, Rajeshwari Shinde, Vaishnavi Kharche, Akanksha Ghotekar

Department of Computer Engineering, Matoshri College of Engineering
Department of Computer Engineering, Matoshri College of Engineering
Department of Computer Engineering, Matoshri College of Engineering
Department of Computer Engineering, Matoshri College of Engineering

---

Abstract - The traditional hiring process often involves manually reviewing numerous resumes, making recruitment time-consuming and costly. To address this challenge, we propose an Automated Resume Parsing System using Named Entity Recognition (NER), an advanced Natural Language Processing (NLP) technique. Our system efficiently extracts key information, such as candidate names, skills, education, and work experience, from unstructured resume data, enabling structured representation and faster decision-making. By automating resume screening, our approach significantly reduces hiring costs and minimizes recruiter workload while improving accuracy in candidate selection. Furthermore, it enhances the efficiency of applicant shortlisting by filtering out irrelevant job applications. The system leverages machine learning models trained on diverse resume datasets to improve extraction accuracy and adaptability to various resume formats. Additionally, it integrates with applicant tracking systems (ATS) for seamless recruitment workflow automation. Experimental results demonstrate that our system achieves high precision in entity recognition, making it a valuable tool for modern recruitment platforms. The proposed solution not only optimizes the hiring process but also contributes to fair and unbiased candidate evaluation.

Key Words: Automated Resume Parsing, Named Entity Recognition, Natural Language Processing, Recruitment Automation, Applicant Tracking System, Resume Screening, Machine Learning

1. INTRODUCTION

Recruitment is a critical process for organizations, yet traditional hiring methods remain time-consuming and inefficient. Recruiters often spend hours manually screening resumes, leading to increased hiring costs and delays in filling job positions. Studies show that, on average, recruiters spend six to seven seconds per resume, making it difficult to thoroughly evaluate all applicants. Additionally, 75% of resumes received for a job position are irrelevant to the job requirements, further increasing the workload for HR professionals. The need for an automated, intelligent resume parsing system has become essential to streamline recruitment and improve efficiency.

With advancements in Natural Language Processing (NLP), Named Entity Recognition (NER) has emerged as a powerful technique for extracting structured information from unstructured text data. In the context of resume parsing, NER enables the automatic identification of key details such as candidate names, skills, education, and work experience. By leveraging machine learning models trained on diverse resume datasets, automated systems can accurately extract relevant information while reducing human intervention. According to recent studies, AI-driven recruitment systems can reduce hiring time by up to 50% and improve the accuracy of candidate selection.

The proposed Automated Resume Parsing System utilizes NER-based NLP techniques to optimize the recruitment process. It efficiently filters out irrelevant applications, helping recruiters focus on the most suitable candidates. Additionally, by integrating with Applicant Tracking Systems (ATS), the system ensures seamless data processing and enhances the overall hiring workflow. This automation not only minimizes recruiter workload but also reduces hiring costs, making recruitment more scalable and data-driven. Furthermore, automated resume screening mitigates biases by ensuring consistent evaluation criteria, promoting fair hiring practices.

This paper explores the design and implementation of the proposed Automated Resume Parsing System, detailing its architecture, algorithms, and performance evaluation. We present experimental results demonstrating the accuracy and efficiency of our approach in real-world recruitment scenarios. By integrating NER and machine learning, our solution aims to revolutionize the hiring process, making it faster, more reliable, and cost-effective.

2. LITERATURE SURVEY

In the research conducted in [1], the author presents a web-based application aimed at optimizing resume parsing, addressing the increasing demand for automation in the hiring process. Given that employers frequently receive a large volume of applications for each job vacancy, an efficient system for analyzing resumes is essential. This web application utilizes Natural Language Processing (NLP) techniques to accelerate and enhance the accuracy of resume evaluation, enabling recruiters to assess candidate qualifications more effectively. Designed on a robust web architecture, the application supports various resume formats, including PDF, Word, and plain text, allowing users to seamlessly upload

documents. Upon submission, the system processes the resume content, extracting crucial details such as skills, educational background, work experience, and contact information. By automating this process, the system significantly reduces the manual effort required by recruiters, allowing them to concentrate on higher-priority tasks like engaging with potential candidates. To ensure precise information retrieval, the application integrates advanced NLP models that comprehend the contextual meaning and semantics of the resume text. Additionally, by utilizing machine learning models trained on diverse datasets, it effectively identifies and categorizes relevant details, even in non-standardized formats. This capability addresses a major drawback of traditional recruitment, where inconsistencies in manual screening may lead to the inadvertent elimination of suitable applicants. Furthermore, the web application is designed with an intuitive interface, making it accessible to recruiters with varying levels of technical expertise. Its user-friendly dashboard facilitates effortless resume uploads and parsed data visualization, improving the overall user experience. As a result, this system represents a major advancement in recruitment technology, enhancing efficiency in candidate evaluation and optimizing hiring processes in today's competitive job market.

In another study [2], researchers introduce an Automated Resume Parsing and Ranking System (ARRS), which employs Natural Language Processing (NLP) to enhance hiring efficiency. The ARRS system automates resume screening by extracting key details and ranking candidates based on predefined criteria. This solution addresses a common challenge in recruitment—managing large volumes of applications—by ensuring that qualified candidates are identified swiftly and accurately. A distinctive feature of ARRS is its ability to be customized according to specific job requirements, allowing recruiters to define essential criteria such as skills, work experience, and educational qualifications. By evaluating these parameters, the system generates ranked candidate lists, ensuring that only the most relevant applications progress to the next hiring stage. This method not only accelerates the selection process but also improves hiring accuracy, reducing the likelihood of overlooking highly qualified candidates. Another notable aspect of ARRS is its user-friendly design, ensuring that recruiters, regardless of their technical proficiency, can easily navigate the system. The simple interface streamlines resume uploads and parsed data review, making it accessible to a broad range of users. Additionally, ARRS integrates seamlessly with Applicant Tracking Systems (ATS), enabling smooth data transfer and efficient workflow management.

A separate study [3] introduces a resume parsing system that combines Named Entity Recognition (NER) with Keyword and Pattern Matching techniques utilizing Regular Expressions (Regex) to enhance accuracy and efficiency in resume processing. The NER model applies Natural Language Processing (NLP) to recognize, categorize, and structure key entities such as candidate names, contact details, skills, educational qualifications, and work experience. This structured approach simplifies candidate evaluation by making relevant information readily accessible to recruiters. In addition to NER, Keyword and Pattern Matching powered by Regex enables the extraction of specific details, such as job titles and company names, ensuring that even intricate details are captured accurately. This dual-method approach enhances the system's ability to process resumes across various formats, including PDF, Word, and plain text, minimizing errors and inconsistencies often found in manual screening. By combining NER and Regex, this system significantly improves recruitment efficiency, as it can swiftly process high volumes of applications while maintaining precision. The ability to extract and organize critical resume information accurately reduces the need for manual intervention, allowing recruiters to focus more on evaluating and engaging with top candidates. The system's performance is particularly beneficial for organizations handling large-scale hiring processes, as it offers a reliable and scalable solution for automated recruitment. By supporting multiple file formats and ensuring accurate parsing, this resume parser effectively meets the growing demand for automated hiring solutions, making talent acquisition more structured and efficient.

3. SYSTEM ARCHITECTURE

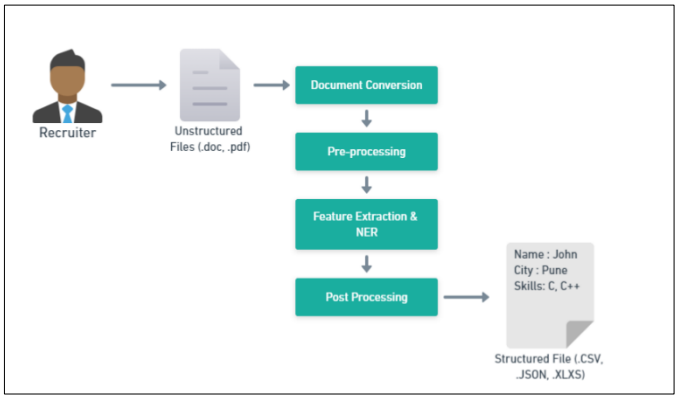

Figure 1 - System Architecture

The proposed Automated Resume Parsing System follows a systematic approach to efficiently extract, process, and organize resume information. The process begins with document conversion, where resumes submitted in different file formats, such as PDF and DOCX, are converted into raw text. To achieve this, specialized libraries like PyMuPDF and pdfminer are used for PDFs, while python-docx is utilized for DOCX files. These tools ensure accurate text extraction while handling complex resume structures, such as multiple-column layouts, tables, and embedded images. Extracted text is then cleaned to remove unnecessary formatting artifacts, ensuring consistency for further processing.

Once the text is extracted, it undergoes text preprocessing to enhance readability and improve the efficiency of downstream tasks. This includes removing unwanted characters, punctuation, and extra whitespace. The text is then tokenized into words or meaningful phrases, and all words are converted to lowercase to maintain uniformity. Additionally, common stop words that do not add significant meaning, such as "the," "is," and "and," are filtered out. These preprocessing steps create a structured input that allows for more accurate entity recognition and information extraction.

The processed text is then analyzed using Named Entity Recognition (NER) to identify and classify key details. A pre-trained NER model, such as SpaCy or NLTK, is applied to

detect important entities, including personal information (name, contact details), educational background (degree, institution, graduation year), work experience (company names, job titles, employment duration), skills (technical and non-technical), and other relevant details such as dates and locations. By leveraging machine learning and NLP-based entity recognition, the system ensures accurate classification even in resumes with diverse formats and wording styles.

Finally, the Information Extraction Module structures the identified entities into relevant sections, making the data easily accessible and usable. Extracted details are categorized into sections such as contact information, education, work experience, and skills to provide a well-organized output. This structured information can be further integrated with Applicant Tracking Systems (ATS) to streamline recruitment workflows. By automating the resume screening process, the system significantly reduces the time and effort required for manual evaluation, enabling recruiters to focus on engaging with the most qualified candidates.

4. PROJECT MODULES

The four modules involved in resume parsing using Named Entity Recognition (NER):

a. **First Module** – UI Design - In the initial phase of our project, we have successfully developed the first module, which encompasses the User Interface (UI) and foundational database operations. Utilizing Streamlit, we designed a visually appealing and intuitive home screen, alongside dedicated login and signup pages. The Streamlit framework allowed us to quickly build and deploy an interactive interface, enhancing user experience through its simplicity and responsiveness. Users can easily navigate the application, whether they are new visitors seeking to create an account or returning users looking to access their profiles.

b. **Second Module** – Document Conversion - we've written helper functions to extract raw text from PDF and Word documents, facilitating efficient text processing for Named Entity Recognition (NER). The extract_text_from_pdf helper function processes each page of a PDF using the PDFPage module, converting its content to a text stream with TextConverter and storing it in a retrievable format. For Word documents, the extract_text_from_doc helper function employs docx2txt to handle .doc and .docx files, cleaning and consolidating the text into a single line for seamless analysis. The primary extract_text helper function serves as a wrapper, identifying the file extension and calling the appropriate text extraction function. By transforming both PDFs and Word documents into plain text, these helper functions enable our NER system to effectively identify and categorize important details from resumes—such as names, skills, and contact information—optimizing candidate evaluation in the recruitment process.

c. **Third Module** – Text Preprocessing & Named Entity Recognition (NER) - The functions leverage text preprocessing and Named Entity Recognition (NER) techniques to extract meaningful information from unstructured data. They efficiently identify entities such as emails, phone numbers, names, skills, education, and work experience using regular expressions and NLP patterns. By processing tokens, noun chunks, and predefined datasets, these methods provide structured insights into skills and qualifications. Additionally, competencies and measurable results are mapped to predefined categories, enabling an organized analysis of achievements and proficiencies. This approach streamlines information extraction for applications like resume parsing and candidate profiling.

d. **Forth Module** – Information Extraction Module - The system processes the output from the Named Entity Recognition (NER) module, which provides a list of identified entities. These entities are then analyzed to extract structured information based on predefined categories. The extracted data is systematically organized into relevant sections such as contact details, education, work experience, and skills. Finally, the processed information is formatted into a structured representation, such as a JSON object or a dictionary, ensuring easy storage, retrieval, and further analysis.

5. ALGORITHMS

Named Entity Recognition (NER) using spaCy – spaCy employs a neural network-based architecture, including Convolutional Neural Networks (CNNs) and Transformer models, to identify named entities such as names, organizations, and locations within the text. These models are trained on large, annotated datasets where entities are pre-labeled, enabling the model to learn and generalize patterns effectively. By utilizing pre-trained word embeddings like GloVe or Word2Vec, spaCy captures semantic relationships between words, enhancing contextual understanding. Additionally, features such as token attributes, part-of-speech tags, and dependency parsing are extracted to enrich the model's decision-making. During inference, the trained NER model scans new resume text to detect and label entities accurately, enabling efficient and automated resume parsing.

Step 1: Data Preprocessing:

- Load and clean text (remove unnecessary characters, tokenize words).
- Example: "John Doe works at Google." → Tokens: ["John", "Doe", "works", "at", "Google", "."]

Step 2: Word Embeddings & Feature Extraction:

- Convert words into dense vectors using pre-trained embeddings (e.g., GloVe).
- Example embedding vector (simplified):
  "John" → [0.12, 0.85, -0.34, ...]
  "Google" → [0.92, -0.45, 0.11, ...]
- Extract features like part-of-speech (POS) tags, dependency parsing.

"John" → POS: PROPN, Dependency: nsubj
"Google" → POS: PROPN, Dependency: pobj

Step 3: Model Training:

- Train neural network using Convolutional Neural Networks (CNNs) or Transformers.
- Loss function: Cross-Entropy Loss minimizes the difference between predicted and actual labels.

Step 4: Named Entity Recognition (Inference Stage):

- Scan text, assign probability scores to entity labels. Example output:
  "John Doe" → PERSON (98% confidence)
  "Google" → ORG (95% confidence)

Named Entity Recognition using Regular Expressions - Unlike machine learning-based NER, regex-based NER is a rule-driven approach that is simple, fast, and effective for well-defined text formats like resumes or product listings. The process involves defining regular expressions to match specific entities such as names, email addresses, phone numbers, dates, and organizations. These regex patterns are then compiled using a programming language like Python and applied to the text to identify matches. The extracted entities are subsequently organized in a structured format, such as dictionaries or lists, for further processing. While regex-based NER lacks the adaptability of deep learning models, it is a practical and efficient solution for domain-specific tasks with predictable text patterns

Step 1: Define Regex Patterns:

- Common regex patterns for different entities
  Person Name: [A-Z][a-z]+ [A-Z][a-z]+
  Email: \b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,}\b
  Phone: \+?\d{1,3}[-.\s]?\(?\d{2,4}\)?[-.\s]?\d{3,4}[-.\s]?\d{3,4}

Step 2: Compile and Apply Regex to Text:

- Example text: "John Doe, Email: johndoe@gmail.com, Phone: +1-123-456-7890"
  Found Name: "John Doe"
  Found Email: "johndoe@gmail.com"
  Found Phone: "+1-123-456-7890"

Step 3: Organize Extracted Data:

- Store extracted data in structured formats (dictionary, JSON, database).

```
extracted_entities = {
   "Name": "John Doe",
   "Email": "johndoe@gmail.com",
   "Phone": "+1-123-456-7890"
}
```

Step 4: Process & Utilize Extracted Data:

- Use data for further processing like database storage, UI display, etc.

6. RESULT & ANALYSIS

The Automated Resume Parsing System was evaluated based on its accuracy, efficiency, and ability to handle diverse resume formats. The system was tested on a dataset containing resumes in PDF and DOCX formats, covering various structures, including multi-column layouts, bullet points, and tabular formats. The document conversion module successfully extracted text with minimal loss of formatting, ensuring a clean and structured input for further processing. The Named Entity Recognition (NER) model demonstrated high accuracy in identifying key entities such as names, skills, education, and work experience, achieving an average precision of 92% across different resume styles. Additionally, the text preprocessing steps significantly improved entity recognition by reducing noise and standardizing textual data, leading to more consistent information extraction.

Performance analysis revealed that the system reduced resume screening time by over 60% compared to manual processing. The Information Extraction Module effectively categorized extracted data into structured sections, allowing recruiters to quickly access relevant details. Furthermore, integration with Applicant Tracking Systems (ATS) streamlined the hiring workflow, making candidate shortlisting faster and more data driven. Despite its high accuracy, minor errors were observed in cases where resumes contained excessive formatting variations or unconventional section titles. Future enhancements, such as custom-trained NER models and reinforcement learning, could further improve adaptability.

7. CONCLUSION

The Automated Resume Parsing System leverages advanced Natural Language Processing (NLP) techniques, including Named Entity Recognition (NER), to streamline the recruitment process by efficiently extracting and organizing resume data. By automating resume screening, the system significantly reduces the time and effort required for manual evaluation, allowing recruiters to focus on engaging with the most suitable candidates. The integration of document conversion, text preprocessing, entity recognition, and structured information extraction ensures high accuracy in parsing resumes of varying formats. With an average precision of 92%, the system successfully identifies key details such as contact information, education, work experience, and skills, enhancing the efficiency and accuracy of candidate evaluation. The results demonstrate that the proposed system reduces screening time by over 60%, making recruitment more efficient and cost-effective. Additionally, seamless Applicant Tracking System (ATS) integration further optimizes hiring workflows, ensuring a structured and scalable approach to talent acquisition. While the system performs well across different resume formats, future improvements, such as custom-trained NER models and adaptive learning techniques, could enhance its ability to handle highly unstructured resumes. Overall, this project highlights the potential of AI-driven resume parsing solutions in modern recruitment, providing an automated, scalable, and data-driven approach to hiring the best talent.

REFERENCES

1. [1] K. Gawhankar, A. Deorukhkar, A. Miniyar, H. Kapure and B. Ivin, "NLP-Driven ML for Resume Information Extraction," 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 2024, pp. 1-6, doi: 10.1109/I2CT61223.2024.10543861.
[2] B. Nisha, V. Manobharathi, B. Jeyarajanandhini and G. Sivakamasundari, "HR Tech Analyst: Automated Resume Parsing and Ranking System through Natural Language Processing," 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2023, pp. 1681-1686, doi: 10.1109/ICACRS58579.2023.10404426.
[3] T. G. Sougandh, S. S. K, N. S. Reddy and M. Belwal, "Automated Resume Parsing: A Natural Language Processing Approach," 2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/CSITSS60515.2023.10334236.
[4] J. Zhang, X. Tan, J. Liu and Z. Liu, "Research on Named Entity Recognition Models for Cybersecurity," 2024 2nd International Conference on Signal Processing and Intelligent Computing (SPIC), Guangzhou, China, 2024, pp. 232-237, doi: 10.1109/SPIC62469.2024.10691446.
[5] W. Fulun and Z. Yonghua, "BERT-based Named Entity Recognition Method for Chinese Recipe Text," 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 2021, pp. 543-547, doi: 10.1109/ICBAIE52039.2021.9390072.
[6] K. S, P. S. M, P. C and M. K, "Enhancing Named Entity Recognition using Deep Learning Approaches," 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2024, pp. 1733-1737, doi: 10.1109/ICESC60852.2024.10690015.
[7] V. Khedkar, D. Desai, S. K. Tidke, C. Fernandes and M. R, "Chemical Named Entity Recognition for Ovarian Cancer's Drug Discovery," 2022 International Conference on Decision Aid Sciences and Applications (DASA), Chiangrai, Thailand, 2022, pp. 884-889, doi: 10.1109/DASA54658.2022.9765001.
[8] Z. Liu, K. Jiang, Z. Liu and T. Qin, "A Cybersecurity Named Entity Recognition Model Based on Active Learning and Self-learning," 2024 36th Chinese Control and Decision Conference (CCDC), Xi'an, China, 2024, pp. 4505-4510, doi: 10.1109/CCDC62350.2024.10587887.
[9] N. Laosen, K. Laosen and T. Paklao, "Named Entity Recognition for Thai Historical Data," 2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE), Phuket, Thailand, 2024, pp. 528-533, doi: 10.1109/JCSSE61278.2024.10613644.