Journal of Human Centered Technology

https://www.humentech.utm.my E-ISSN: 2821-3467

SmartResume: A Resume Parser Application using Natural Language Processing

Muhammad Alif Imran Mohammad Fadzir[1], Shakirah Hashim[2]*

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, 40000, Malaysia

*Corresponding Author shakirahhashim@uitm.edu.my

Cite: https://doi.org/10.11113/humentech.v5n1.118 Research Article

Abstract:
In today's dynamic business environment, Information and Communication Technology (ICT) serves as a pivotal force driving innovation, particularly in human resource management. One notable advancement is the emergence of smart resume applications, which are reshaping traditional recruitment practices. These systems utilize advanced technologies such as artificial intelligence (AI), machine learning, and natural language processing (NLP) to facilitate the automation of resume screening and candidate evaluation processes. This paper presents the development of an Automated Resume Parsing designed to reduce the operational load on human resource professionals. By utilizing NLP techniques, the system extracts key candidate information including name, contact details, educational background, and skills from the resume documents. A custom Named Entity Recognition (NER) model is employed to enhance the accuracy and relevance of extracted data. The model was trained using the SpaCy NLP framework to achieve an overall accuracy of 92.4% and an F1-score of 0.90. The extracted data are presented through an interactive web interface for HR personnel, enabling structured and efficient review of applicant information. The results demonstrate that integrating NLP and machine learning in recruitment systems can significantly enhance automation, consistency, and fairness in candidate evaluation processes. The system's effectiveness was demonstrated through successful extraction and structured presentation of applicant information, aligned with organizational hiring criteria. Beyond technical functionality, this study highlights the potential of such human-centered technologies to enhance decision-making, increase recruitment efficiency, and transform the recruiter-applicant interaction by fostering transparency and fairness in the hiring process.

Keywords: Natural language processing (NLP); Resume parser; Named entity recognition (NER)

1. INTRODUCTION

Selecting the right candidate for a vacancy can be an arduous and time-intensive process. Extracting information from the resume is still highly challenging because they are highly heterogeneous in format and content. Resumes come in a wide variety of formats, ranging from simple text documents to highly stylized templates with complex layouts. Additionally, the amount and kind of data included in the resumes also differ depending on the experience and the requirements of the position. Performing the process manually may lead to inaccuracies and human errors. Alternatively, some organizations provide an online form for job applicants to fill in with the required details. While this solution is convenient for the organization, it can be tedious for candidates to re-enter information they have already included in their resumes.

Thus, resume data parsing powered by Natural Language Processing (NLP) techniques is a critical component to solving these problems. NLP can be broadly defined as the branch of Artificial Intelligent (AI) that has to do with the spoken and written languages of humans. It pertains to how to reduce language processing to machinery that will be of significance to the human being. Uses of NLP are numerous and varied, such as language translation (1), voice recognition (2), abstracting (3), summarizing (4), and opinion mining (5). The contribution of NLP in the case of resume parsing is extremely important because it enhances the ability to extract relevant information from resumes with greater accuracy and efficiency.

This study aims to develop a web-based Resume Parser System that will harness NLP for effective information extraction from a resume. The system is intended to address the challenges associated with manual data extraction from a resume. Accordingly, the objectives of the study are as follows:

- To design an NLP model capable of identifying and extracting key information from the resumes.
- To develop a web-based resume parser system that applies NLP techniques to process the resume in various formats.
- To evaluate the functionality of the system in ensuring efficient and accurate information extraction from the resume.

Received 01 July 2025; Accepted 13 November 2025
Available online 06 February 2026

2. BACKGROUND OF THE STUDY

On the 1980s, a rule-based parsing systems were trailblazers where the systems would identify a particular keyword, pattern, or a combination of both from the resume based on predefined rules. However, this type of parsing is rigid which limits the adaptability of the patterns and often leads to inaccurate results (6). In addition, it was structurally oriented to extract only major information, such as education and work experience, and had no capability of dealing with difficult linguistic constructions. In this persuasion, development related to resume parsing started fully in the 2000s and progressed before the 2010s, incorporating machine learning techniques. Algorithms such as Naive Bayes and Support Vector Machines were one-of-a-kind in boosting resume parsing procedures with much more accuracy and swiftness (7). Furthermore, in the course of this period, the invention of the Part-of-Speech (POS) tagging was handy in adding to the understanding of the context through the determination of the grammatical roles of words that help to come out with proper sentence structures.

Since 2010, resume parsing performance has improved significantly due to advancements in algorithms such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), which have enhanced capabilities in information extraction. For example, text classification for resume parsing using CNN variants has been evaluated with respect to accuracy and efficiency (8). This is proven further by the capabilities of modern parsing systems to process diverse resume formats, including PDF, Word, and plain text (9) Other NLP models such as Bidirectional Encoder Representations from Transformer (BERT) (10), Named Entity Recognition (NER) (11, 12), and Term Frequency-Inverse Document Frequency (TF-IDF) (13) are capable of efficiently recognizing and extracting a candidate's skills and qualifications from resumes. Furthermore, by employing an advanced Large Language Model (LLM), accurate predictions can be achieved due to its ability to capture deeper semantic meanings from the nodes and edges of the Graph Neural Network (GNN) extracted from resumes.

A work in Gaur et al. (14) proposed a semi-supervised deep learning model for classifying educational institutions in resume parsing. The model was initially trained on unlabelled data using a deep neural network, followed by the application of a correction module and subsequent retraining to enhance classification accuracy. Also using deep learning (15) proposed an automated resume parser which has an ability to rank the candidate's score using heuristic calculations. It achieved an information extraction accuracy of 93%, which is comparable to human-level performance. Meanwhile,(16) proposed a resume parser that extracts skills from the resume and uses them to recommend relevant jobs to the candidate, based on the cosine similarity calculation. To improve the accuracy of resume parsing, a paper in (10) has incorporated pre-trained architecture BERT where it suggests three features to enhance the parsing process from rich text documents which are capitalization, shrift size, and font styles.

Additionally, a distance-based algorithms such as Jaccard distance, Euclidean distance and Cosine similarity are commonly used to compare the similarity between vectors. This approach helps recommend the most suitable job opportunities based on the information extracted from resumes. For instance, (17) utilized NLP in combination with distance-based algorithms to develop a job recommendation system that matches resumes to suitable job roles. Their findings support the use of similarity measures such as Euclidean and Cosine similarity for effective candidate-job matching.

2.1 Named Entity Recognition (NER)

Named Entity Recognition uses statistical models and machine learning techniques that classify words into predefined categories (18). It extracts entities from the text and performs the identification of the names of the following entities: a person's name, organization, locations, dates, etc. This is appropriate for resume parsing as it can be exploited to extract the important details from the resume and automatically identify the keywords accordingly.

However, NER techniques also face several challenges, particularly related to ambiguity and context dependency, which stem from the inherently ambiguous nature of entities in natural language (19). Effective entity recognition often necessitates domain-specific knowledge and tailored model training to achieve optimal performance. Furthermore, recognizing and processing entities across multiple languages presents additional challenges due to the diversity of linguistic structures, grammar rules, and writing systems involved.

In light of these challenges, this paper proposes a custom NER model using SpaCy, a comprehensive Python library for text processing, tokenization, and parsing. The model is tailored to recognize specific entities relevant to resumes, such as names, contact information, locations, work experience, skills, education, and more. While the focus of this study is on a custom NER model developed using the spaCy framework, alternative lightweight approaches such as rule-based extraction using regular expressions and libraries like NLTK and TextBlob have also been explored in prior works (6, 7). These methods rely on manually defined linguistic patterns or keyword matching to identify entities such as names or contact information. In contrast, NER approach adopted in this work enables context-aware entity recognition and generalizes better across diverse resume formats. The integration of spaCy's statistical model and custom-trained annotations allows the system to capture both lexical and contextual dependencies, achieving higher extraction accuracy compared to rigid rule-based parsing.

3. METHODOLOGY

The overall system architecture of the automated resume parser is illustrated in Figure 1, which outlines the main components: (i) Text Extraction and Preprocessing, (ii) Entity Recognition using NLP, and (iii) Data Visualization and Interface. The proposed system automates the extraction of structured information from resumes submitted by candidates through the web-based platform. Upon uploading file, the system processes the document and outputs structured data such as name, contact information, education, skills, and work experience.

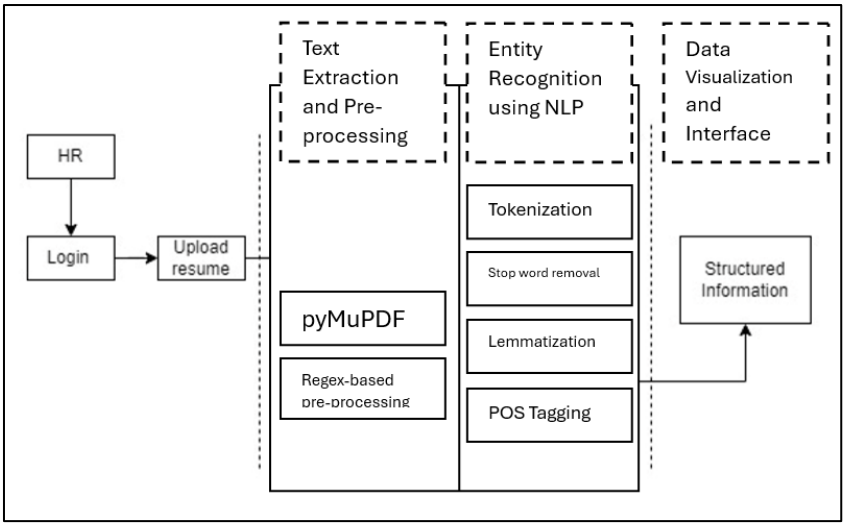

Figure 1. Architecture of Resume Parser System.

i) Text Extraction and Pre-processing

The text filtering module employs the pyMuPDF library (Figure 2) due to its high accuracy in preserving text layout during extraction from multi-page PDF resumes, which is essential for consistent parsing of sections such as Experience and Education. After extraction, the text is cleaned using regex-based preprocessing (20) to remove special characters, bullet points, and redundant whitespace.

ii) Entity Recognition using NLP

The textual data then undergoes several NLP preprocessing steps:

- Tokenization; conducted using the NLTK word tokenizer to split sentences into word-level tokens.
- Stop word Removal; eliminates non-informative words using NLTK's predefined English stop word list.
- Lemmatization; standardizes tokens to their base form to reduce morphological variations.
- Part-of-Speech (POS) Tagging; assists in identifying proper nouns and skill names for improved entity recognition.

This preprocessing pipeline ensures that the input text is clean, consistent, and semantically structured for the Named Entity Recognition model. For this work, SpacyNER is implemented to locate and classify these entities into predefined categories such as locations, organizations, skills, experience, email and contact number which can be displayed or used for further processing. The custom NER model was trained using the annotated dataset to detect specific entity types relevant to resume data. The dataset used for developing and evaluating the proposed NLP-based resume parser consisted of 500 anonymized resumes collected from open-access resume repositories and publicly available Kaggle datasets.

```
def extract_text_from_pdf(file_path):
    # Open the PDF file
    document = fitz.open(file_path)
    text = ""
    for page_num in range(len(document)):
        page = document.load_page(page_num)
        text += page.get_text()
    return text
```

Figure 2. Code snippet for extracting text from a PDF-formatted resume.

iii) Data Visualization and Interface

The model outputs entities such as person, email, phone, education, experience, location, and skills, which are then structured into JSON format for downstream processing. The extracted information is rendered back to the HR personnel through an HTML template. The system supports multiple file formats (PDF and DOCX) and can handle resumes up to 10 MB in size. All data processing occurs locally to ensure privacy and data confidentiality.

Figure 3 (a) shows the interface for the user to upload the resume. The uploaded resume can be viewed to show that the file has been correctly selected. Next, Figure 3 (b) illustrates the extracted information from the resume, presented in a user-friendly format for HR personnel.

(a) (b)

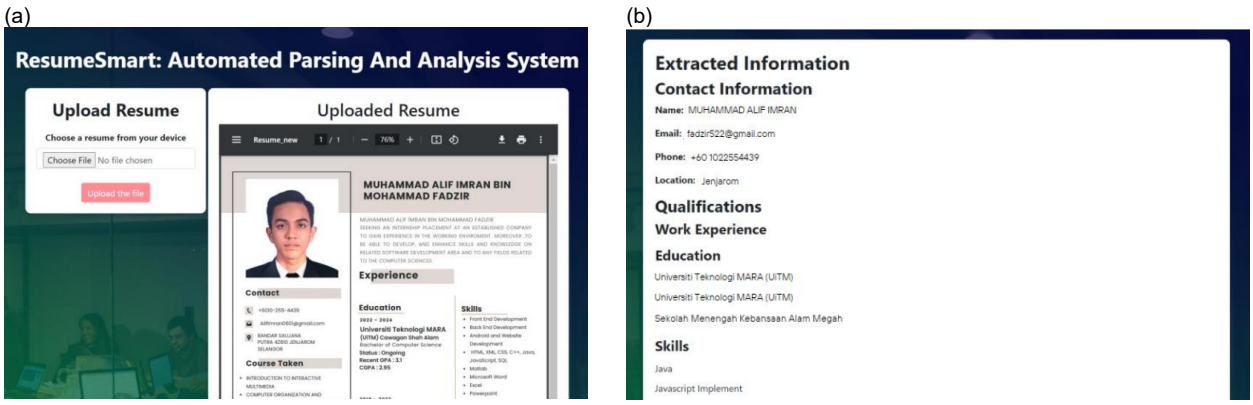

Figure 3. Interface to upload resume and interface extracted information

4. RESULTS AND DISCUSSION

To evaluate the performance of the NER model, standard NLP evaluation metrics were used:

- F1-Score: harmonic mean of precision and recall, providing a balanced metric.
- Accuracy: proportion of correctly extracted entities over the total.

The model achieved an overall accuracy of 92.4%, with an F1-score of 0.90 for the skills and experience entities, demonstrating reliable performance for HR application use cases. Meanwhile, test cases are designed to systematically verify the functionality of the Resume Parser System. Test case specifies inputs, expected outcomes, and the conditions under which testing will be performed. Users are required to validate the test case presented in Table 1.

Table 1. Test case of the Resume Parser System.

Test Scenario	Steps	Predicted Results	Actual Results	Status
Register New Account	1. Enter username, password and cofirm password. 2. Click on “Register” button.	Login page will be displayed	Login page is displayed	Successful
Log into system	1. Enter username, password and cofirm password. 2. Click on “Login” button.	Upload Resume Page will be displayed	Upload Resume Page is displayed	Successful
Upload Resume	1. Browse a resume file Click on “Upload the File” button	Resume file will be uploaded and displayed	The uploaded resume is displayed	Successful
Display Extracted Information	n/a	Automatically display the extracted information from the resume	The information extracted from the resume is displayed	Successful

As shown in Table 1, a functionality test was conducted using four test cases. The actual results were compared to the expected results, and a test was considered successful if both matched. In this study, all four test cases successfully passed the functionality test.

5. CONCLUSION

A web-based application to parse a resume has been developed to ease the HR departments and recruiters to extract key information from resumes using NLP techniques. By integrating a custom SpaCy Named Entity Recognition (NER) model trained on an annotated dataset, the system achieved a high accuracy of 92.4%, effectively extracting key attributes such as skills, experience, and education. The results confirm that NER-based approaches are more reliable for entity-level extraction compared to conventional rule-based methods. The system’s web interface further enhances usability by allowing HR personnel to easily upload, process, and view structured candidate profiles.

In future work, the custom NER model will be trained on a larger set of Malaysian resume datasets to enhance its performance. A feedback loop where the model’s predictions are periodically reviewed and corrected will be implemented to further improve learning and accuracy. Additionally, the application of transfer learning techniques using domain-specific NER models is expected to enhance the recognition and classification of entities in resumes, leading to more precise and reliable outcomes.

AUTHORSHIP CONTRIBUTION STATEMENT

Muhammad Alif Imran Mohammad Fadzir: writing - original draft, formal analysis, methodology, project administration. Shakirah Hashim: conceptualization, validation, supervision, writing - review & editing

DATA AVAILABILITY

Data is openly available in a public repository.

AI DECLARATION

The use of ChatGPT and Grammarly in this manuscript was used for grammar checking and paraphrasing. The authors reviewed and edited the content as needed and took full responsibility for the final publication.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no competing interests.

ACKNOWLEDGMENT

This study was supported by Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA. The authors sincerely appreciate the institutional support that contributed to the success of this research.

REFERENCES

(1) Dalayli F. Use of NLP techniques in translation by ChatGPT: Case study. Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC). 2023;19–25. https://aclanthology.org/2023.contents-1.3/
(2) Thakur A, Ahuja L, Vashisth R, Simon R. NLP & AI speech recognition: An analytical review. 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE. 2023;1390–1396.
(3) Lachmy R, Pyatkin V, Manevich A, Tsarfaty R. Draw me a flower: Processing and grounding abstraction in natural language. Trans Assoc Comput Linguist. 2022;10:1341–1356. https://doi.org/10.1162/tacl_a_00522.
(4) Kadhim EA, Feizi-Derakhshi MR, Aghdasi HS. Advanced text summarization model incorporating nlp techniques and feature-based scoring. IEEE Access. 2025;(99):1. https://doi.org/10.1109/ACCESS.2025.3528830.
(5) Kumaragurubaran T, Vaishnavi C, Vidhiya SB. Psychological-based opinion mining for fake review detection using advanced NLP and machine learning techniques. 2025 Global Conference in Emerging Technology (GINOTECH). IEEE. 2025;1–6. https://doi.org/10.1109/GINOTECH63460.2025.11077032.
(6) Rawat A, Malik S, Rawat S, Kumar D, Kumar P. A systematic literature review (SLR) on the beginning of resume parsing in HR recruitment process & SMART advancements in chronological order. 2021. https://doi.org/10.21203/rs.3.rs-570370/v1.
(7) Sanyal S, Hazra S, Ghosh N, Adhikary S. Resume parser with natural language processing. Comp Sci. 2017.
(8) Liu J, Shen Y, Zhang Y, Krishnamoorthy S. Resume parsing based on multi-label classification using neural network models. Proceedings of the 6th International Conference on Big Data and Computing. 2021;177–185. https://doi.org/10.1145/3469968.3469998.
(9) Aggarwal A, Jain S, Jha S, Singh VP. Resume Screening. Int J Res Appl Sci Eng Technol. 2022;10. https://doi.org/10.22214/ijraset.2022.43037.
(10) Bocharova MY, Malakhov EV. CapStyleBERT: Incorporating capitalization and style information into BERT for enhanced resumes parsing. ACM International Conference Proceeding Series. Association for Computing Machinery. 2024;249–254. https://doi.org/10.1145/3651781.3651820.
(11) Narendra GO, Hashwanth S. Named entity recognition-based resume parser and summarizer. Int J Adv Res Sci Commun Technol. 2022;2:728–735. https://doi.org/10.48175/IJARSCT-3029.
(12) Gautam G, Sharma D, Chaturvedi V. Automating talent acquisition assisted by resume parsing system for enhanced job matching. 2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN). 2025;278–281. https://doi.org/10.1109/cictn64563.2025.10932337.
(13) Mariappan P, Krishna K, Sahoo J, Preetham S. Smart resume screening and job represents recommendation system. available at SSRN 5141402. 2025.
(14) Gaur B, Saluja GS, Sivakumar HB, Singh S. Semi-supervised deep learning based named entity recognition model to parse education section of resumes. Neural Comput Appl. 2021;33:5705–5718. https://doi.org/10.1007/s00521-020-05351-2.
(15) Thangaramya K, Logeswari G, Gajendran S, Roselind JD, Ahirwar N. Automated resume parsing and ranking using natural language processing. 2024 3rd International Conference on Artificial Intelligence for Internet of Things (AIIoT). 2024;1–6. https://doi.org/10.1109/AIIoT58432.2024.10574696.
(16) Chandak AV, Pandey H, Rushiya G, Sharma H. Resume parser and job recommendation system using machine learning. 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC). 2024;157–162. https://doi.org/10.1109/ESIC60604.2024.10481635.

(17) Artajaya H, Julieta, Giancarlos J, Moniaga J V, Chowanda A. Job recommendation system based on resume using natural language processing and distance-based algorithm. 2024 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS). 2024;1–6. https://doi.org/10.1109/AIMS61812.2024.10512474.
(18) Eiselen R, Bukula A. IsiXhosa named entity recognition resources. ACM Trans Asian Low-Resour Lang Inf Process. 2022;22(2):1–9. https://doi.org/10.1145/3531478.
(19) Khetan V, Wetherley E, Eneva E, Sengupta S, Fano AE. Knowledge graph anchored information-extraction for domain-specific insights. 2021;arXiv:2104.08936. https://doi.org/10.48550/arXiv.2104.08936.
(20) Sougandh TG, Reddy NS, Belwal M. Automated resume parsing: A natural language processing approach. 2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS). IEEE. 2023;1–6. https://doi.org/10.1109/CSITSS60515.2023.10334236.