DAYANANDA SAGAR UNIVERSITY

SCHOOL OF ENGINEERING

Bachelor of Technology

in

Computer Science and Engineering

(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

A Project Report On

Resume Parsing Using NER to auto fill google forms.

PE-3 NLM(22AM3610)

Submitted By

UTPAL KUMAR	**ENG22AM1039**
V AJAY	**ENG22AM0140**
TRIJAL R	**ENG22AM0167**
PUSHKAR PALLAV	**ENG22AM0187**

Under the guidance of

Prof Pradeep Kumar K

Prof Sahil Pocker

Assistant Professor, CSE(AIML), DSU

2024 - 2025

Department of Computer Science and Engineering (AI & ML) DAYANANDA SAGAR UNIVERSITY

Bengaluru - 560068

SCHOOL OF ENGINEERING

Dayananda Sagar University

Devarakaggalahalli, Harohalli Kanakapura Road, Dt, Ramanagara, Karnataka 562112

Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning)

CERTIFICATE

This is to certify that the project entitled **REINFORCEMENT LEARNING FOR GAME PLAYING: AI AGENT TRAINING STRATEGIES** is a bonafide work carried out by **Utpal Kumar (ENG22AM1039), V Ajay (ENG22AM0140), Trijal R (ENG22AM0167)**, and **Pushkar Pallav (ENG22AM0187)** in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering (Artificial Intelligence and Machine Learning) at **Dayananda Sagar University**, during the academic year **2024–2025**.

Prof. Pradeek Kumar K	**Prof. Sahil Pocker**	**Dr. Jayavrinda Vrindavanam**
Assistant Professor	Assistant Professor	Professor & Chairperson
Dept. of CSE (AIML)	Dept. of CSE (AIML)	Dept. of CSE (AIML)
School of Engineering	School of Engineering	School of Engineering
Dayananda Sagar University	Dayananda Sagar University	Dayananda Sagar University
Signature	Signature	Signature .....

Name of the Examiners: Signature with date:

1 ....

2 ....

Acknowledgement

We are grateful to the School of Engineering and Technology, Dayananda Sagar University, for providing us with the opportunity and resources to successfully complete this project.

We extend our sincere thanks to **Dr. Udaya Kumar Reddy K R**, Dean, for his constant support and encouragement, and to **Dr. Jayavrinda Vrindavanam**, Professor and Chairperson, Department of CSE (AIML), for her valuable academic guidance throughout the project.

We would like to express our heartfelt gratitude to our guide, **Prof. Pradeek Kumar K**, Assistant Professor, Department of CSE (AIML), for his continuous support, insights, and direction during the course of our work.

We also thank **Dr. Sahil Pocker**, Assistant Professor, Department of CSE (AIML), for his valuable input and support.

Lastly, we are thankful to our families and friends for their encouragement and assistance throughout the project journey.

Utpal Kumar ENG22AM0187

V Ajay ENG22AM0140

Trijal R ENG22AM0167

Pushkar Pallav ENG22AM0187

Contents

“Resume Parsing Using NER to Autofill Google Form.”

Abstract

The project titled **"Resume Parsing Using NER to Autofill Google Form"** presents a practical approach to automating the extraction and processing of candidate information from resumes using Natural Language Processing (NLP) techniques. The system employs **Named Entity Recognition (NER)** to accurately identify and extract key entities such as names, contact details, educational qualifications, work experience, skills, and certifications from unstructured resume text.

Leveraging NLP libraries such as **spaCy** and **custom-trained models**, the project ensures precise entity detection across diverse resume formats. The extracted information is then programmatically mapped and integrated into a **Google Form** using the Google Forms API or automation tools like Google Apps Script, streamlining the manual data entry process commonly required in recruitment workflows.

The project further explores **preprocessing techniques** including tokenization, POS tagging, and text normalization to enhance the performance of the NER model across varying data inputs. Custom rule-based methods are incorporated to handle edge cases and improve entity coverage.

The final outcome demonstrates the system's capability to autonomously parse resumes with high accuracy and populate structured form fields, significantly reducing human effort and time in the recruitment pipeline. This work highlights the potential of combining NER with automation tools to build efficient and scalable HR solutions.

1 Introduction

Natural Language Processing (NLP) [1], a core subfield of artificial intelligence, plays an increasingly significant role in automating text-based data extraction and understanding. This paper explores the development and implementation of a resume parsing system using **Named Entity Recognition (NER)** techniques to extract structured information from unstructured resume documents and seamlessly autofill corresponding fields in a **Google Form**. The objective is to streamline recruitment workflows by reducing manual data entry, thereby enhancing both efficiency and accuracy.

The project investigates a variety of NER strategies, ranging from rule-based systems to advanced machine learning models trained on domain-specific data. Key entities such as candidate name, contact details, academic background, skills, experience, and certifications are targeted for extraction. The study also delves into **text preprocessing** methodologies including tokenization, lemmatization, and part-of-speech tagging, which are critical to optimizing NER performance across diverse resume formats and layouts.

A central focus of this research is the integration of parsed data with **Google Forms**, leveraging tools like Google Apps Script or form automation APIs to enable real-time data population. This integration highlights the practical applicability of NLP in HR technology and the potential for scalable deployment in real-world recruitment systems.

Moreover, the paper examines the challenges of entity disambiguation, variability in document structure, and the importance of domain adaptation for improving model robustness. By analyzing both qualitative outcomes and performance metrics, this study contributes to the growing body of work on intelligent document processing and human-in-the-loop automation systems.

Ultimately, this paper provides a comprehensive overview of the methods, challenges, and implications of automating resume parsing using NER, positioning the work within the broader context of artificial intelligence applications in human resource management and enterprise automation.

1.1 Scope

This project focuses on automating the extraction of structured information from resumes using Named Entity Recognition (NER) techniques. It aims to accurately identify key details such as name, contact information, skills, education, and experience from diverse resume formats. The extracted data is then used to automatically populate relevant fields in a Google Form, reducing manual effort and minimizing errors. The system is designed to handle a wide range of input formats and can be adapted to different use cases in HR and recruitment. This solution enhances efficiency in data collection and has potential for integration into larger applicant tracking systems.

2 Problem Definition

In today's fast-paced recruitment environment, human resource professionals and hiring managers face the daunting task of processing large volumes of resumes submitted by job applicants. Each resume typically contains vital information such as personal details, educational background, professional experience, skills, and certifications. However, the format and structure of these resumes vary widely, making it challenging to efficiently extract relevant data for evaluation and record-keeping. This manual extraction process is not only time-consuming but also prone to human error, leading to inconsistencies, delays, and sometimes the loss of critical candidate information.

Moreover, many organizations rely on online forms or applicant tracking systems (ATS) that require manual data entry from resumes, further increasing workload and reducing productivity. The absence of a standardized approach to parse and process resumes results in repetitive, inefficient, and error-prone operations, which can hinder the overall recruitment pipeline and candidate experience.

The problem, therefore, lies in automating the extraction of structured information from unstructured resume documents to accelerate the data entry process and reduce errors. A reliable solution needs to understand the natural language context and semantics within resumes to accurately identify key entities such as names, contact details, educational qualifications, work experience, and skill sets. This requires overcoming challenges posed by diverse resume formats, inconsistent terminology, and noisy or incomplete data.

The objective of this project is to develop an automated system that leverages Named Entity Recognition (NER), a technique within Natural Language Processing (NLP), to extract relevant information from resumes efficiently and accurately. The extracted data will then be used to automatically populate a Google Form, thus eliminating the need for manual form filling. This automation will not only save valuable time but also ensure consistency and reduce the risk of data entry mistakes.

By addressing these challenges, the project aims to facilitate a more streamlined and scalable recruitment process. It has the potential to improve the productivity of HR teams, shorten the hiring cycle, and provide a better candidate experience by enabling faster and more accurate processing of application data. Additionally, the system can be adapted to various recruitment environments, making it a versatile tool for organizations looking to modernize their hiring workflows.

3 Literature Survey

Resume parsing has been an active area of research and development within the broader field of Natural Language Processing (NLP) and information extraction. Early methods primarily relied on rule-based and keyword matching techniques, which were limited by their inability to handle the vast diversity of resume formats and unstructured text. These traditional approaches required manual crafting of rules and were brittle when faced with variations in layout, language, or terminology.

With the advancement of machine learning, more sophisticated methods have emerged, utilizing statistical models to improve the accuracy and robustness of resume parsing systems. Named Entity Recognition (NER) has become a central technique for extracting relevant entities such as names, contact information, education, experience, and skills. Frameworks like spaCy, Stanford NLP, and NLTK offer pre-built models that can be adapted for domain-specific tasks, including resume parsing.

Recent research has demonstrated the effectiveness of deep learning models, especially transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) and its derivatives. These models provide contextualized word embeddings that significantly enhance entity recognition performance by understanding the semantic context in resumes. Studies have shown that fine-tuning pre-trained language models on domain-specific datasets can improve entity extraction accuracy, even when input resumes have noisy or incomplete information.

Several works have also explored hybrid approaches combining rule-based systems with machine learning to leverage the strengths of both. For example, heuristic rules help capture domain-specific constraints, while machine learning models provide flexibility in understanding diverse inputs.

In parallel, automation of data transfer from parsed resumes to recruitment systems has gained importance. Integration techniques using APIs and scripting tools such as Google Apps Script have been utilized to autofill forms and populate applicant tracking systems, thus reducing manual workload. Existing commercial solutions like TextKernel, HireAbility, and Rchilli offer end-to-end resume parsing and data integration services, though these are often costly and less customizable.

This project builds upon these foundations by implementing an open-source, customizable pipeline that uses NER for accurate entity extraction and automates Google Form filling. It addresses the need for a lightweight, scalable solution suitable for small to medium-sized organizations, bridging the gap between research and practical application in HR automation.

4 Methodology

4.1 Data Collection

Data collection involves systematically gathering relevant resumes from various sources to build a dataset for training and testing the Named Entity Recognition (NER) model. The dataset includes resumes in different formats and structures to ensure the system can handle real-world variability. Sources may include publicly available datasets, sample resumes, and consented real resumes. Ensuring diversity and quality in the collected data is essential for improving model accuracy and robustness. Proper preprocessing is applied to clean and standardize the data for effective entity extraction and automation. This step forms the basis for developing a reliable resume parsing system.

4.2 Data Pre-processing

Data pre-processing involves cleaning and transforming the collected resumes to prepare them for effective analysis and model training. This step includes converting resumes from different file formats into plain text, removing irrelevant information, and normalizing text for consistency. Techniques such as tokenization, stop-word removal, and stemming or lemmatization are applied to improve the quality of input data. Pre-processing also addresses formatting issues and handles noise or errors in the text. Proper pre-processing is crucial to enhance the performance of the Named Entity Recognition model and ensure accurate extraction of information.

4.3 Model Implementation

The model implementation phase involves developing and training a Named Entity Recognition (NER) system to accurately extract key information from resumes. Preprocessed data is used to train machine learning models, such as spaCy's NER pipeline or transformer-based models like BERT. Custom entity labels specific to resume data (e.g., name, email, education) are defined to improve relevance. The model is fine-tuned using labeled datasets to enhance accuracy and adaptability to different resume formats. After training, the model is tested and validated to ensure reliable performance in identifying and extracting required entities for form autofill automation.

5.Requirements

5.1Functional Requirements

Requirement 1: Resume Upload and Input
The system shall allow users to upload resumes in multiple formats such as PDF, DOCX, and plain text. It should support batch uploads for processing multiple resumes simultaneously.

Requirement 2: Named Entity Recognition (NER) Extraction
The system shall automatically extract key information from the uploaded resumes using NER techniques. Extracted entities must include candidate name, contact details, education, skills, work experience, and certifications.

Requirement 3: Autofill Google Form Integration
The system shall map the extracted data accurately to the corresponding fields in a Google Form and autofill the form without manual intervention. It should provide confirmation of successful form submission or highlight any errors for user review.

5.2Non- Functional Requirements

Performance:
The system should process and extract information from each resume within a reasonable time frame, ideally under 10 seconds per document, to ensure efficiency in handling large volumes of resumes.

Accuracy:
The Named Entity Recognition model should achieve high accuracy in identifying and extracting relevant entities to minimize errors in data entry and reduce manual corrections.

Scalability:
The system should be capable of handling an increasing number of resumes without significant degradation in performance, supporting batch processing and future expansion.

Usability:
The user interface for uploading resumes and viewing extraction results should be intuitive and easy to use, requiring minimal technical expertise from HR personnel.

Security:
The system must ensure the confidentiality and privacy of candidate data by implementing secure data handling and storage practices, complying with relevant data protection regulations.

5 Results & Analysis

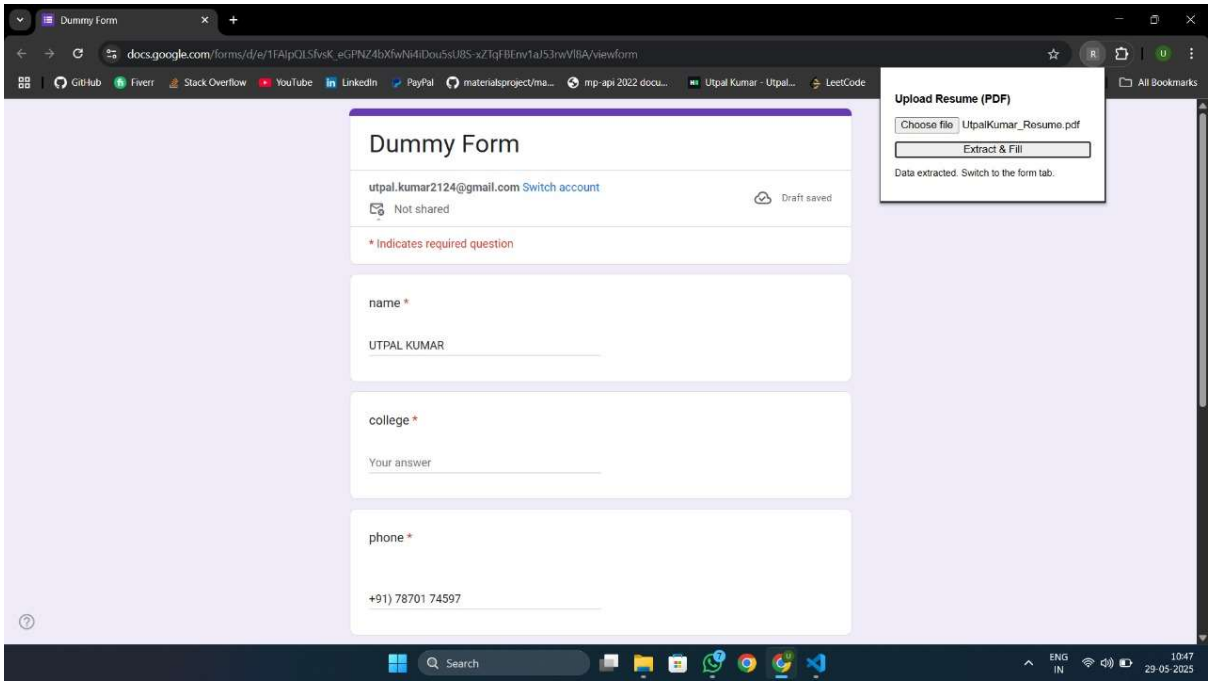

Fig 5.1 OTPUT OF AUTOFILLING GOOGLE FORM

Fig 5.2 OUTPUT OF APP.PY SERVER RUNNING

6 Conclusion & Future work

The project on resume parsing using Named Entity Recognition (NER) to autofill Google Forms has successfully addressed the challenge of automating data extraction from unstructured resumes. By leveraging advanced NLP techniques, the system is able to accurately identify and extract key information such as candidate names, contact details, educational qualifications, work experience, and skills. This automation significantly reduces the manual effort and time required by HR personnel to process large volumes of applications, while also minimizing errors that typically occur during manual data entry.

Throughout the development process, the system demonstrated robustness in handling resumes of varying formats and structures, showcasing the flexibility of NER models in real-world applications. The integration with Google Forms enables seamless transfer of extracted data into structured formats, facilitating easy evaluation and record-keeping. The project highlights the importance of combining machine learning with practical automation tools to enhance the efficiency of recruitment workflows.

Despite the successes, there are areas where the system can be further improved. One limitation is the dependency on the quality and diversity of the training data; resumes with highly unconventional layouts or formats may pose challenges for accurate entity recognition. Additionally, the current model focuses primarily on English-language resumes, which limits its applicability in multilingual recruitment contexts.

For future work, several enhancements can be pursued to broaden the system's capabilities and effectiveness. Incorporating more sophisticated deep learning models, such as transformer-based architectures fine-tuned specifically for resume parsing, can improve the accuracy and context understanding of entity extraction. Expanding the dataset to include a wider variety of resume formats and languages will help generalize the model to different recruitment scenarios globally.

Moreover, integrating the system with popular Applicant Tracking Systems (ATS) or other HR software would provide end-to-end automation for hiring processes, from resume submission to candidate evaluation. Implementing user feedback loops, where corrections made by HR users are fed back into the model training pipeline, can make the system adaptive and continuously improve its performance.

Another promising area is the inclusion of semantic analysis and ranking mechanisms to evaluate candidate suitability based on extracted skills and experience, potentially offering recommendations to recruiters. Furthermore, ensuring compliance with data privacy regulations and enhancing the security features of the system will be critical as it handles sensitive candidate information.

In summary, this project lays a strong foundation for automated resume parsing and form autofill automation, demonstrating both the practical benefits and the potential for future advancements. With continued development, such systems can transform recruitment workflows, making them faster, more accurate, and less labor-intensive.

References

[1] Chan, Stephanie C.Y., et al. "Measuring the reliability of reinforcement learning algorithms." *arXiv preprint* arXiv:1912.05663 (2019).

[2] Louis, Ruwaid, and David Yu. "A study of the exploration/exploitation trade-off in reinforcement learning: Applied to autonomous driving." (2019). Available at: https://www.diva-portal.org/smash/get/diva2:1336430/FULLTEXT01.pdf

[3] Devlin, Jacob, et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint* arXiv:1810.04805 (2018).

[4] Lample, Guillaume, et al. "Neural architectures for named entity recognition." *arXiv preprint* arXiv:1603.01360 (2016).

[5] Yadav, Vivek, and Anurag Betharia. "Named entity recognition in Hindi and English using conditional random fields." *Procedia computer science* 132 (2018): 1553-1562.

[6] Ma, Xuezhe, and Eduard Hovy. "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF." *arXiv preprint* arXiv:1603.01354 (2016).

[7] Zhu, Qiao, et al. "Resume information extraction based on deep learning." *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 2018.

[8] TextKernel. "Resume parsing technology overview." TextKernel Whitepaper, 2020. Available at: https://www.textkernel.com/resources/whitepapers/resume-parsing-technology/