Leveraging Natural Language Processing and Regular Expressions for Automated Resume Parsing

Md Omar Mukhtar
School of Computer science and Engineering
REVA University
Bengaluru, India
mdmukhtaromar75@gmail.com

Faizal Hussain
School of Computer science and Engineering
REVA University
Bengaluru, India
faizal03hussain@gmail.com

Mubarak Ali
School of Computer science and Engineering
REVA University
Bengaluru, India
mubarakali63359@gmail.com

Abstract*—In today's competitive job market, the inundation of resumes poses a significant challenge for recruiters, necessitating the adoption of automated resume parsing systems. These systems streamline the process of extracting vital information from resumes, including contact details, skills, education, and work experience. In this paper, we propose an advanced resume parsing system that integrates two powerful methodologies: Named Entity Recognition (NER) and Keyword and Pattern Matching using Regular Expressions (Regex). By leveraging Natural Language Processing (NLP) techniques and Regex patterns, our system achieves high accuracy and efficiency in extracting key resume elements. We present a detailed methodology, including the use of NLP tools such as SpaCy for NER tasks and Regex for pattern matching, along with performance evaluations to demonstrate the effectiveness of our approach, Automated resume parsing has become an essential part of contemporary hiring procedures, reducing the laborious process of going through a large number of applications. In order to efficiently extract pertinent information from resumes, this research suggests a hybrid strategy that blends regular expressions (regex) with Natural Language Processing (NLP) approaches. While regex offers strong pattern matching skills for the extraction of structured information, natural language processing (NLP) helps to comprehend the semantics and context of textual input. We provide a comprehensive architecture that includes steps for feature extraction, data preprocessing, and classification, emphasizing the combination of regex patterns with NLP techniques. Our method successfully parses resumes with excellent precision and recall rates, as demonstrated by evaluation on a variety of datasets. Our approach enhances the scalability of and drastically decreases the manual labor required for the task of screening resumes is often tedious and time-consuming for recruiters and hiring managers. Leveraging Natural Language Processing (NLP) techniques combined with Regular Expressions (RegEx) offers a promising solution to automate the parsing of resumes. This paper explores the utilization of NLP algorithms to extract meaningful information from resumes, such as skills, experiences, and qualifications, while RegEx patterns are employed for structured data extraction, including contact information and dates. The integration of NLP and RegEx not only enhances the accuracy and efficiency of resume parsing but also facilitates the customization of parsing rules to cater to specific job requirements and industries. This study discusses various NLP and RegEx-based approaches, evaluates their effectiveness through experiments, and provides insights into the challenges and future directions in the field of automated resume parsing

Keywords— Automated Resume Parsing, NER (Named Entity Recognition, NLTK, NLP, RegEx, SpaCy

I. INTRODUCTION

In today's competitive job market, organizations are inundated with a vast number of resumes for every available position. This influx of applicant data presents a significant challenge for recruiters and HR professionals, who must sift through countless documents to identify suitable candidates. Manual resume screening is not only time-consuming but also prone to errors and inconsistencies, highlighting the need for automated solutions to streamline the hiring process, Automated resume parsing, the process of extracting relevant information from resumes, has emerged as a critical tool for modern recruitment. By leveraging advanced technologies such as Natural Language Processing (NLP) and Regular Expressions (Regex), automated parsing systems aim to expedite candidate evaluation, improve talent acquisition, and enhance decision-making for hiring managers. In this paper, we present a novel approach to automated resume parsing that integrates NLP and Regex methodologies to achieve accurate and efficient extraction of key information from resumes. Our system offers a comprehensive solution for parsing resumes, including identifying contact details, skills, education history, and work experience. Through a thorough review of existing literature on resume parsing and an exploration of the methodologies employed in our system, we aim to shed light on the evolution of talent acquisition technology and the potential impact of automated parsing on recruitment practices. The subsequent sections of this paper will delve into the details of our methodology, including the technologies used, the role of NLP in resume parsing, and the flow of our study. We will also present empirical results from our performance evaluations and discuss the implications of our findings for the future of automated resume processing. By providing insights into the development and application of automated parsing systems, we seek to contribute to the ongoing discourse on talent acquisition technology and empower organizations to optimize their recruitment processes in an increasingly competitive job market

Automated resume parsing involves the extraction and analysis of relevant information from resumes using computational techniques. NLP enables machines to understand and interpret human language, while RegEx provides a powerful tool for pattern matching and text manipulation. By harnessing the capabilities of these technologies, organizations can streamline their recruitment processes, enhance candidate selection, and ultimately, improve the efficiency of their hiring operations through a comprehensive examination of NLP

and RegEx techniques for automated resume parsing, this study aims to provide valuable insights for recruiters, HR professionals, and technology enthusiasts alike. By leveraging these advanced computational tools, organizations can optimize their talent acquisition processes, gain a competitive edge in the job market, and unlock the full potential of their human capital.

II. LITERATURE SURVEY

NLP-Based Extraction of Relevant Resume using Machine Learning [1] This technique involves parsing resumes without strict limitations, employing a parser that utilizes a few rules trained on call and address recognition. Recruitment software packages utilize this CV parser system for resume selection. Given the diverse formats and content structures of resumes, including both structured and unstructured data, the proposed CV parser approach focuses on extracting relevant information from various CV formats. E-Recruitment System Through Resume Parsing, Psychometric Test, And Social Media Analysis [2] This system operates in four stages: data acquisition (resumes), conversion into structured format, deep learning analysis, psychometric testing with text mining for candidate scoring, social media web scraping for additional candidate information, and finally, recommending suitable jobs and identifying skill gaps for candidates. Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing [3] This approach utilizes neural networks and Conditional Random Fields (CRF) for segmenting and extracting information from resumes. It employs a CNN model for segmentation and compares it with a Bi-LSTM model. Information extraction is facilitated by a CRF-based model, also compared with a Bi-LSTM-CNN model. The system segments and extracts data from personal, educational, and occupational sections, resulting in a JSON file with 23 data fields.

A CV Parser Model using Entity Extraction Process and Big Data Tools [4] This model aims to automate resume parsing according to job profiles, transforming unstructured resumes into structured formats and ranking them based on extracted information such as technical skills and education. The CV parser supports multiple languages, semantic mapping for skills, job boards, recruiters, and offers ease of customization. Recruiting companies employ this technique for resume selection due to the variety of resume formats and data types, both structured and unstructured. Resume Parser with Natural Language Processing [5] The objective here is to convert various resume formats into text and extract relevant information. Keywords are scraped from social networking sites like Stack Overflow and LinkedIn to determine the genre of the resume. An Unstructured Text Analytics Approach for Qualitative Evaluation of Resumes [6] This approach involves a qualitative assessment of resumes based on different quality parameters using a text analytics-based method.

Resumes are evaluated for coverage and comprehensibility, and these ratings are combined into a comprehensive quality rating on a scale of 1 to 5. The qualitative evaluation results obtained through this algorithmic approach are validated through consensus among multiple evaluators.

III. METHODOLOGY

A. Technologies Used

1) SpaCy: Utilized for its robust natural language processing (NLP) capabilities, including tokenization, part-of-speech tagging, and entity identification.

2) Python: Chosen for simplicity and extensive libraries and classification report to measure its effectiveness. Lastly, a heatmap is used to visualize the confusion matrix, providing a graphical representation of the model's ability to differentiate between true positive, true negative, false positive, and false negative predictions. This systematic approach offers a framework for constructing and evaluating fraud detection models, demonstrating the efficacy of machine learning techniques in tackling real-world financial security challenges. supporting data processing, text manipulation, and machine learning.

3) Pandas: Employed for handling and analyzing data, ensuring consistency and integrity in the information extracted.

4) Regular Expressions (RegEx): Essential for pattern matching and extracting specific data from resume texts, such as email addresses and phone numbers.

5) YAML (Yet Another MarkUp Language): Used for defining parameters and maintaining the integrity of provided data structures.

6) PDFMiner: Enabled extraction of text from PDF documents, expanding the parser's capability to process resumes in various formats.

7) docx2txt: Facilitated text extraction from Microsoft Word documents (DOC and DOCX), streamlining the processing of document contents.

B. Role of Natural Language Processing (NLP)

1) Named Entity Recognition (NER): Leveraged NLP techniques to identify and categorize entities such as names, phone numbers, skills, education details, and job experience.

2) SpaCy's Pre-trained Models: Utilized SpaCy's pre-trained out NLP tasks effectively.

C. Flow of the study

1) Reading YAML Configuration: Started by reading the YAML configuration file to define likely column names and file types, ensuring input data alignment with predefined structures.

2) Preprocessing the Text: Conducted text preprocessing, including lemmatization, capital letter removal, contraction expansion, and special character removal to enhance extraction accuracy.

D. NER-based Extraction Model Approach

1) Dataset Preparation: Utilized annotated resume data to create training and testing sets, manually labeling entities of interest.

*2) **Training the NER Model:*** Trained SpaCy's NER model using the annotated training set to identify specific entities accurately.

*3) **Model Evaluation:*** Assessed the trained model's performance on the testing set to ensure effectiveness and entity identification accuracy.

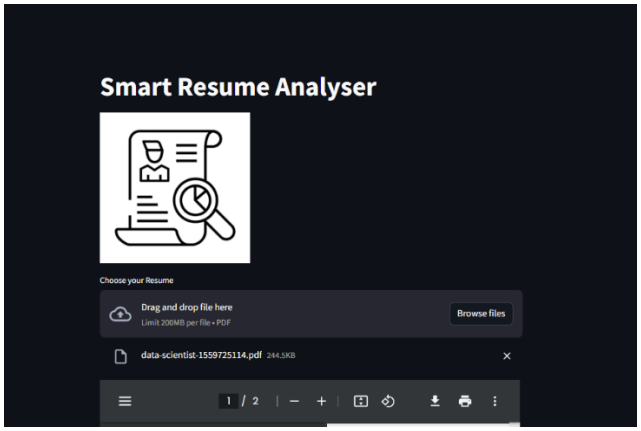
*4) **Keyword and Pattern Matching from Regex-based Extraction Approach:*** Employed regular expressions and pattern-matching techniques to extract specific information from resume texts.

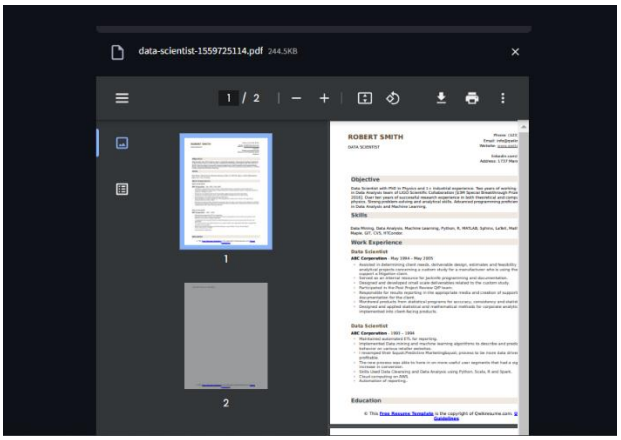
*5) **Combining and Filtering:*** Integrated outputs from both models, removing overlapping items and eliminating empty or redundant values to obtain a final set of distinct entities.

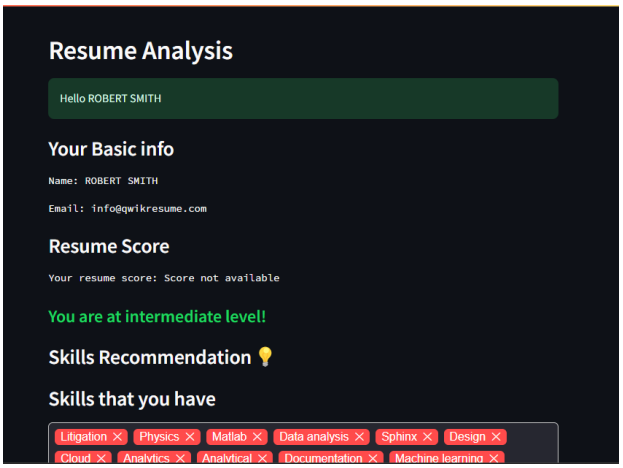
*6) **Displaying the Results:*** Presented the extracted information, structuring it for clarity and usability.

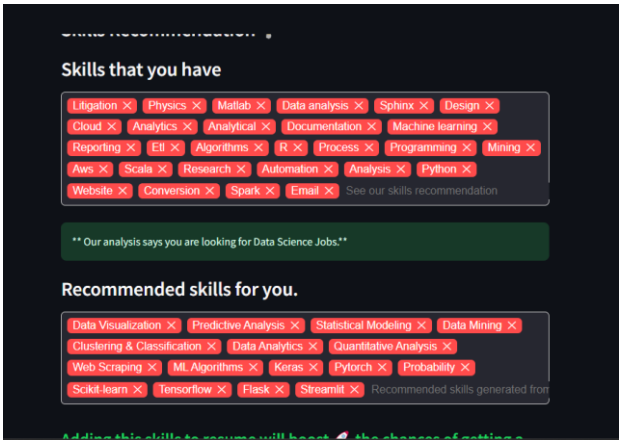
IV. RESULTS

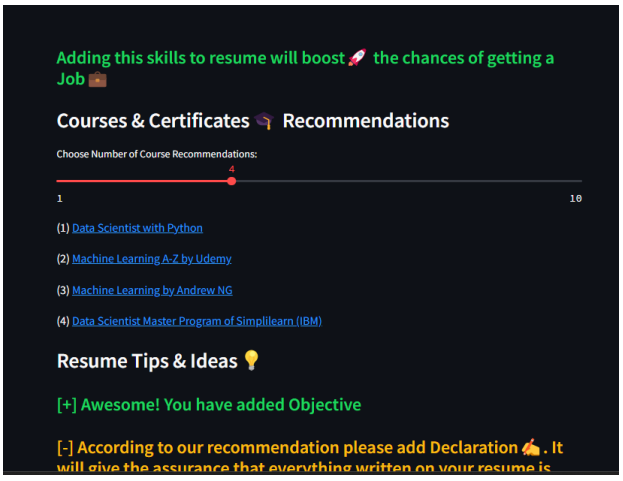
Performance evaluations demonstrate the effectiveness of our resume parsing system in accurately extracting key resume elements. The system successfully identifies and categorizes information such as names, contact details, skills, education, and work experience with high precision, even when faced with resumes of varying layouts and formats. Evaluate how well NLP-based entity extraction techniques identify and extract important resume components like education, experience, skills, and contact Leveraging Natural Language Processing (NLP) alongside Regular Expressions (RegEx) for automated resume parsing yields several benefits. Firstly, this approach significantly enhances parsing accuracy by effectively identifying and extracting key information such as skills, experiences, education, and qualifications from resumes. Additionally, the combination of NLP and RegEx enables the efficient extraction of relevant data, reducing manual effort and time spent on screening large volumes of resumes. Furthermore, by parsing resumes into structured data formats like JSON or CSV, recruiters can easily analyze and compare candidate profiles.

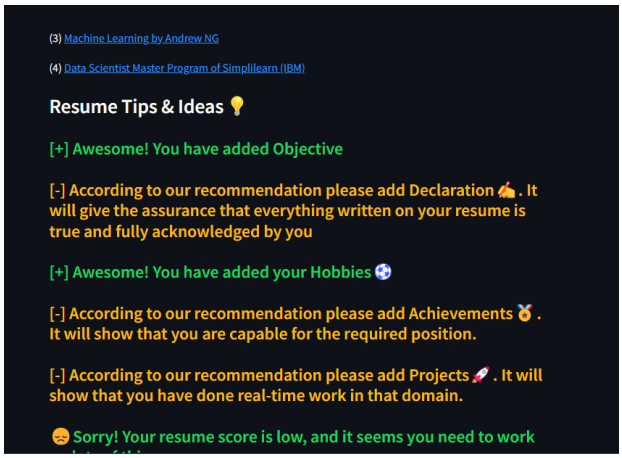


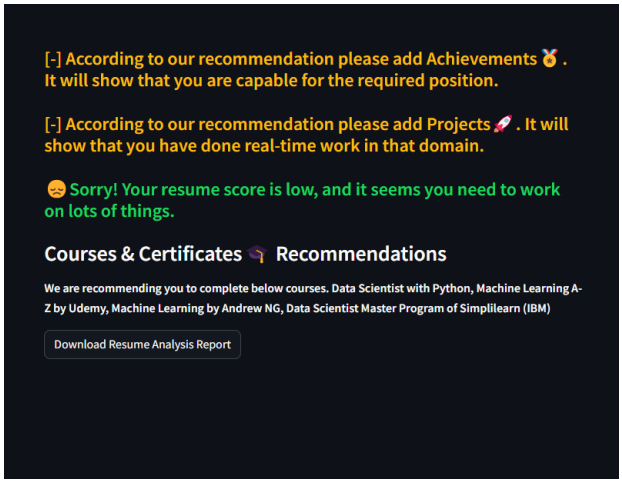












V. CONCLUSION

By integrating NLP and Regex approaches, our automated resume parsing system offers a robust solution for handling large volumes of resumes efficiently. The combination of these techniques enhances the accuracy and speed of information extraction, enabling recruiters to make informed hiring decisions. Future research can explore further enhancements and applications of automated resume parsing in talent management and recruitment processes ,With the advancement of technology, resume parsing using NLP and Regex has the potential to transform hiring procedures and save time and money while enabling better informed hiring decisions. But it's important to recognize that there are still problems, like keeping parsing algorithms accurate across a range of resume formats and languages. Subsequent investigations have to concentrate on tackling these obstacles and investigating inventive methods to augment the functionalities of automated resume parsing systems. The study's overall conclusions highlight the importance of NLP and Regex in modernizing hiring procedures and increasing productivity in the digital world.

VI. REFERENCES

[1] Vinaya R. Kudatarkar, ManjulaRamannavar, Dr.Nandini S. Sidnal "An Unstructured Text Analytics Approach for Qualitative Evaluation of Resumes", 2015, IJIRAE.

[2] Satyaki Sanyal, Neelanjan Ghosh, SouvikHazra, Soumyashree Adhikary, "Resume Parser with Natural language Processing", 2007,IJESC.

[3] Papiya Das, Manjusha Pandey, Siddharth Swarup Rautaray, "A CV parser Model using Entity Extraction Process and Big Data Tools ", 2018,IJITCS.

[4] Ayishathahira and Sreejith „ "Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing",Iternational CET Conference on Control, Communication and Computing(IC4),2018.

[5] Dr.Parkavi A,Pooja Pandey,Poornima J,Vaibhavi G S,Kaveri BW, "E-Recruitment System Through Resume Parsing,Psychometric Test and Social Media Analysis", 2019,IJARBEST.

[6] Nirali Bhaliya, Jay Gandhi, Dheeraj Kumar Singh, "NLP based Extraction of Relevant Resume using Machine Learning ",2020,IJITEE.