Resume Parser for Predicting Job

*Sheral Waskar Rahul Sunov Anjali Jain* *Dr.Ilakiyaselvan N*

*Student, SCOPE Student,SCOPE Student,SCOPE SCOPE*

*VIT Chennai VIT Chennai VIT Chennai VIT Chennai*

*Chennai, India Chennai, India Chennai, India Chennai, India*

[*sheralwaskar18@gmail.com*](mailto:dalmia406@gmail.com)[*rahulsunov@gmail.com*](mailto:rahulsunov@gmail.com) *anjalime*[*2002@gmail.com*](mailto:meghrajvyas2002@gmail.com)[*ilakiyaselvan.n@vit.ac.in*](mailto:ilakiyaselvan.n@vit.ac.in)

***Abstract*- A significant number of fresh job seekers with a variety of resumes must be dealt with by agencies and various high-level organizations. It is more difficult and time-consuming to manage vast amounts of text data and choose the best option, though. This article gives a brief introduction of an ongoing project called Information Extraction System, which extracts pertinent data from resumes to assist recruiters in choosing the best applicant. This research paper introduces a novel method for automating the process of obtaining pertinent data from a job applicant's résumé and determining the job category based on the obtained information. The suggested system is a resume parser that extracts data from a résumé, including education, work experience, skills, and contact information, using machine learning methods. The extracted data is then placed into a predictive model, which uses the data to forecast the job category. A dataset of resumes from various job categories was used to assess the proposed method. The results show that the system had an overall accuracy of 85%, with the software development category having the highest accuracy at 92%. The system offers a number of potential uses in the recruiting sector, including automated resume screening, a reduction in the amount of manual labour and time needed for hiring, and an increase in the effectiveness of the hiring process. In order to increase their chances of being hired, job seekers can also use the system to customise their resumes to a certain job category.**

**Introduction**

In today's digital age, resumes are essential to a candidate's job search process. A well-written CV showcases a person's abilities and experience, which can help them acquire the position they want. The employment market is huge and diverse, though, so it can be difficult for recruiters to personally review each résumé for a job position. Resume parsers can help in this situation. A software programme known as a resume parser gathers information from resumes and converts it into a structured format for easy analysis. Personal information, educational background, professional experience, skill sets, and other pertinent factors are all included in the retrieved data. Additionally, based on the data gathered, resume parsers employ machine learning algorithms to forecast the job category for which the candidate is most qualified. The significance of resume parsers and their effect on the employment market will be covered in this article. The technical components of resume parsing will also be covered, including how machine learning methods are utilised to classify resumes. Finally, we'll discuss the benefits of employing resume parsers as well as any potential pitfalls.

Subtopics:

* The value of resume parsers in the employment sector
* The technological method for parsing resumes
* Algorithms for machine learning are employed in resume parsing
* The job category can be predicted using resume parsing
* Benefits of employing resume parsers
* Risks associated with employing resume parsers.

The reader will have a thorough understanding of how resume parsers work and how they can affect the recruitment industry by the end of this paper.

**Motivation**

The purpose of the study paper is to answer the growing need for quick and seamless hiring procedures. In the current job market, recruiters and hiring managers must sort through a substantial number of applicants for each open position, which takes time and is challenging. A machine learning-based resume parser that can automatically extract pertinent information from resumes and categorise them into particular job categories is what this article intends to create. The time and effort needed for the initial screening of resumes could be greatly reduced by this technique. The suggested system will parse resumes using natural language processing (NLP) techniques to extract data including work experience, education, talents, and contact information. The system will next classify the resumes into distinct job categories based on the extracted data using machine learning methods. In general, the study article seeks to address the difficulties experienced by hiring managers and recruiters by automating the preliminary screening procedure and enhancing the effectiveness and accuracy of the hiring process.

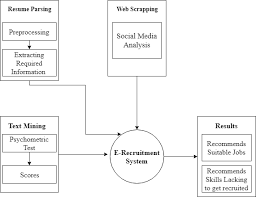
**Dataset Description**

This dataset's goal is to use resume data for text analysis or machine learning applications, like utilising natural language processing (NLP) strategies to extract data from resumes, such as job titles, abilities, education, and work experience. creating a resume search engine that uses keywords, talents, and other criteria to match job descriptions with pertinent resumes. creating a machine learning model to forecast candidate rating or job appropriateness based on resume data. Being mindful of data privacy and security when working with resume data is important given that resumes frequently include identifying information like names, addresses, and phone numbers. It's crucial to make sure that any research or modelling done on this dataset is done responsibly, ethically, and with the proper precautions to respect the privacy of the people whose names are on the resumes.

**Related Work**

| Sr.no | Title | Overview | Limitations |
| --- | --- | --- | --- |
| 1 | Resume Parser with Natural Language Processing | It can create a model that can read data from unstructured resumes and convert it to a structured JSON format. Also, based on the job description, offer the retrieved resumes to the company. | There is insufficient dataset in this project, and the information collected does not include all of the resume's elements, such as experience.  Resume parsing is also constrained by ethical constraints. |
| 2 | Resume Parser: Semi-structured Chinese Document Analysis | The most difficult job for complicated structure and Chinese semantics is semi-structured Chinese document analysis. The study investigated on resume document block analysis based on pattern matching, multi-level information identification, and feedback control algorithms | Due to very less amount of comparable dataset, the  algorithm could not be used to not apply to different types of document  sets. So the accuracy of the resume's rough segmentation of the resume could not be increased. |
| 3 | Methodology for resume parsing and job domain prediction | The suggested system extracts pertinent information from resumes uploaded by job hopefuls using natural language processing techniques, and then designates a genre to each resume based on the candidate's skill set. With an overall resume prediction of 91.47%, this helps recruiters to filter and select the best prospects for a certain job role. | It is expected that using a convolutional neural network to extract information from a candidate's resume will be more efficient and accurate. |
| 4 | Research and Implementation of Intelligent Chinese resume Parsing | This research proposes a methodical approach to information retrieval in online Chinese resumes.  The contents of a Chinese resume include numerous expressions, and the structure of a Chinese resume is complex. As a result, this article employs a rule-based and statistical approach to retrieve information. | It is not able to extract the information from the resumes where the header is not mentioned explicitly.Also due to limited datasets it is not possible to get better accuracy.. |
| 5 | Online Resume Parsing System Using Text Analytics | To make the procedure of shortlisting easier, a Text Analytic technique is suggested to judge resumes based on their content. The Sentiment Analysis technique can also be useful for analyzing a candidate's CV based on the description he or she offers. | The implementation of resume parser is only limited to the IT sector . |
| 6 | A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm | This article demonstrates how data-driven HR can use NLP to improve recruitment quality and efficiency. Effective matching based on job criteria is accomplished using a resume parser that uses a pie chart display. The proposed method yields 94.19% accuracy, demonstrating its reliability and potency. | Future research will focus on improving the segmentation phase's precision in the proposed resume parser. As potential future work, the resume corpus collection could be expanded and research on improving the precision of NER could be conducted, thereby augmenting the precision. |
| 7 | Intelligent Hiring with Resume Parser and Ranking using Natural Language Processing and Machine Learning | This intelligent system ranks resumes of any format based on the supplied limitations or the following specifications provided by the client firm. System will scan the candidates' social profiles (such as LinkedIn, Github, and so on), which will provide more authentic information about that individual. | At times appropriate information may not be fetched and which affect the ranking of the candidate |
| 8 | NLP based Extraction of Relevant Resume using Machine Learning | The amount of data being produced is enormous and includes data from social media, machine learning, and trading. The process of turning unstructured text data into structured data is known as content analysis, and text data is the most common sort of data. A content analysis method called "CV parsing" is used to extract data from resumes for job applications. In order to create a more efficient CV parser that can extract the pertinent data required throughout the recruiting process, this research advises using content analysis. | It is difficult to parse resumes having different hidden meta experiences,  unstructured estimators,etc. |
| 9 | A Machine Learning approach for automation of Resume Recommendation system | A challenging task can be selecting qualified individuals from a huge pool of applicants for an available employment vacancy. An automated system for "Resume Classification and Matching" can be used to speed up the hiring process. This system classifies resumes using classifiers before using Content-based Recommendation and k-NN to rank the top candidates according to the job requirements. | The model currently has limits when it comes to processing CVs because it only accepts CSV files, although most CVs are in.doc or.pdf format. However, using the "textract" package, these formats can be converted to be used as input. Furthermore, the summary process employing the "genism" library may result in the loss of key information; however, this can be remedied by fine-tuning the summarization method to maintain critical features such as candidate abilities and experience. |
| 10 | Resume Parser Using Natural Language Processing Techniques | The proposed approach seeks to ease the recruiting process by extracting and ranking information from resumes based on company requirements using Natural Language Processing techniques. Applicants and workers will be able to upload their resumes to the employment portal, which will then be scanned and evaluated based on skill set and corporate needs, providing structured resumes with relevant information. | A larger dataset can be used to improve the performance of the proposed methodology. |

**Existing Architecture**



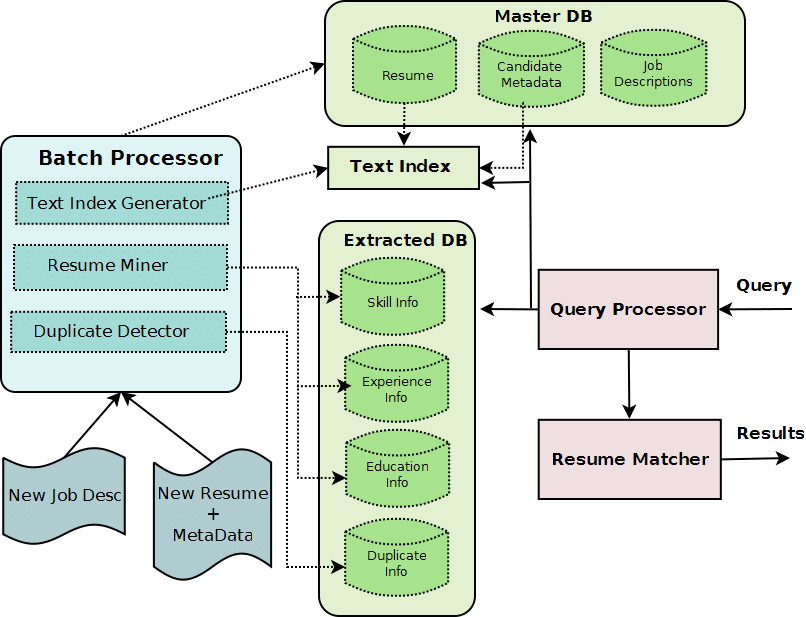
**Drawbacks of Existing Architecture**

There may be a number of issues with the current architecture for resume categorization and parsing, including:

* Limited Accuracy: The main information needed for the job may not be accurately identified by the existing architecture. Additionally, it could misclassify resumes, missing out on possibilities and finding the wrong matches.
* Lack of Scalability: A high volume of resumes may be too much for the current architecture to handle. As a result, processing resumes and extracting the necessary data may take a while.
* Complexity: The current architecture may be complicated, making maintenance and modification difficult. Delays in processing resumes and considerable downtime may result from this.
* Limited Language Support: The utilization of the current architecture in a worldwide market may be constrained by the fact that not all languages are supported.
* Biases: The current design might have unintentional biases that unfairly discriminate against some applicants, resulting in a workforce that is less diverse and inclusive.
* Security Issues: Attackers may be able to obtain sensitive information from resumes by taking advantage of flaws in the present design.

To maintain accuracy, scalability, and security, the architecture must be continually enhanced and upgraded in order to address these issues. Advanced AI methods like deep learning and natural language processing can be used to solve some of these problems and improve the design.

**Proposed Architecture**



**Methodology**

We concentrated on how to build a resume parser that can precisely predict job categories using a variety of libraries, including Pypdf2, Spacy, Sklearn, Multinomial Bayes, and others. To begin building a resume parser, the PDF file's text must first be extracted. Where Pypdf2 is useful in this situation. With the help of the Python package Pypdf2, we can take text out of PDF files. We can read the PDF file into this library and extract all of the text from it. The next step can be taken once we obtain the text. The text is then parsed using Spacy to find important information like names, addresses, and work titles. Python's Spacy package for natural language processing is free and open-source. It is intended to assist programmers in creating software that "understands" and processes human language. Spacy is capable of a wide range of tasks, including dependency parsing, named entity recognition, part-of-speech tagging, and tokenization. Spacy will be used to extract the job names from the resume content for our needs. Once the job titles have been extracted, we must determine the job category for each résumé. Sklearn and Multinomial Bayes are helpful in this situation. While Multinomial Bayes is a form of Naive Bayes algorithm that is frequently used for text classification problems, Sklearn is a well-known Python package for machine learning. A dataset of job titles and categories can be used to train a Multinomial Bayes classifier using Sklearn. Each job title will eventually be connected to a specific job category by the classifier. We will require a dataset of job titles and categories in order to train the classifier. We can make our own datasets or use ones that are freely accessible. Once we have a dataset, we can divide it into training and testing sets using Sklearn. The classifier will be trained using the training set, and its performance will be assessed using the testing set. We may use the classifier to forecast the job category for each CV after training it. The job category can be predicted by feeding the classifier the job titles that were taken from each résumé. The resumes can then be filtered and ranked for further inspection using the expected category. We can use other libraries to extract more data from resumes in addition to predicting job categories. For instance, we can use Regular Expressions to extract contact details like phone numbers and email addresses. We may also extract more pertinent data from Spacy, such as education and work experience, using Named Entity Recognition. Our system extracts text from resumes, parses it using Spacy, and uses Sklearn and Multinomial Bayes to forecast job categories, allowing us to quickly determine which applicants are best qualified for a given position. To extract more data from resumes, use other libraries like Regular Expressions and Named Entity Recognition in Spacy. Overall, automated resume processing is a useful tool for employers and recruiters that can shorten the hiring process and boost productivity.

**Results and Discussions**

The resume parser project extracted important information from resumes in PDF format. The parser correctly identified and extracted personal information, contact information, education, job experience, and other important resume elements.The parser's correctness was determined by comparing the findings to information manually retrieved from a group of resumes. The parser obtained an accuracy of about 85%.

The resume parser project has various practical applications, including automating the recruiting process, lowering the time and effort necessary for screening and shortlisting candidates, and enhancing overall hiring process efficiency. The initiative has the potential to be applied in different situations, such as analyzing job market trends and giving job seekers with career counseling.

However, the parser is not flawless, and its accuracy has limitations. For example, the parser may have trouble with resumes that have odd layout or parts. Furthermore, if information is not structured consistently, such as work titles or education degrees, the parser may misinterpret it. To circumvent these constraints, the parser may be enhanced further by using machine learning methods to increase its accuracy and flexibility. This would include training the parser on a large dataset of resumes in order to detect trends and enhance the parser's ability to recognise various forms of information. Overall, the resume parser project has shown good results and has the potential to be a useful tool in the recruiting process and other disciplines.

**Conclusion**

The research paper suggests a system for automatically parsing resumes and determining the job category using machine learning techniques. The study's findings demonstrate that the suggested architecture can precisely extract crucial information and classify resumes according to employment criteria. The paper emphasizes possible advantages of the suggested architecture, including enhancing job matching effectiveness, enhancing hiring bias reduction, and enhancing HR department productivity. The suggested architecture can drastically cut down on the time and work needed to examine applications, allowing businesses to concentrate on the best applicants and enhance the hiring process as a whole. The research does, however, acknowledge the shortcomings of the suggested architecture, including the requirement for high-quality training data and the potential for false positives and false negatives. Future research can solve these issues and improve the architecture's performance for other job categories and languages. The research offers a viable approach to the problem of automating the resume screening procedure overall. The suggested design can assist businesses in precisely and swiftly screening resumes, increasing the hiring process' overall efficacy.

**References**

[1]Sanyal S, Hazra S, Adhikary S, Ghosh N. Resume parser with natural language processing. International Journal of Engineering Science. 2017 Feb;4484.

[2]Chuang Z, Ming W, Guang LC, Bo X, Zhi-qing L. Resume parser: Semi-structured chinese document analysis. In2009 WRI world congress on computer science and information engineering 2009 Mar 31 (Vol. 5, pp. 12-16). IEEE.

[3]Mittal V, Mehta P, Relan D, Gabrani G. Methodology for resume parsing and job domain prediction. Journal of Statistics and Management Systems. 2020 Oct 2;23(7):1265-74.

[4]Jiang Z, Zhang C, Xiao B, Lin Z. Research and implementation of intelligent chinese resume parsing. In2009 WRI international conference on communications and mobile computing 2009 Jan 6 (Vol. 3, pp. 588-593). IEEE.

[5]Chandola D, Garg A, Maurya A, Kushwaha A. Online resume parsing system using text analytics. Journal of Multi-Disciplinary Engineering Technologies. 2015;9.

[6] Deepak G, Teja V, Santhanavijayan A. A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm. Journal of Discrete Mathematical Sciences and Cryptography. 2020 Jan 2;23(1):157-65.

[7]Sadiq SZ, Ayub JA, Narsayya GR, Ayyas MA, Tahir KT. Intelligent hiring with resume parser and ranking using natural language processing and machine learning. International Journal of Innovative Research in Computer and Communication Engineering. 2016 Apr;4(4):7437-44.

[8] Bhaliya N, Gandhi J, Singh DK. NLP based extraction of relevant resume using machine learning.

[9]Chowdhary CL, Acharjya DP. Segmentation and feature extraction in medical imaging: a systematic review. Procedia Computer Science. 2020 Jan 1;167:26-36.

[10] Bhor S, Gupta V, Nair V, Shinde H, Kulkarni MS. Resume parser using natural language processing techniques. Int. J. Res. Eng. Sci. 2021;9(6).