Vellore Institute of Technology, Vellore (2023), **x**, 1–6
doi:

ARTICLE

HireHive : A smart resume analyzer

Kushagra Palod, Archit Jain, and Shlok Aggarwal

Vellore Institute of Technology, Vellore, India

Abstract

In today's job market, a strong and effective resume is an essential component of any successful job search. However, creating a professional and polished resume can be a challenging task for many job seekers. To address this issue, we have developed a resume analyzer project that leverages natural language processing (NLP) techniques to provide users with a comprehensive analysis of their resumes.

Our resume analyzer project uses NER and the Python library pyresparser, which is built on the Spacy framework, to extract relevant information from resumes in PDF format. By analyzing the content of each resume, the project provides a score that reflects the level of professionalism and effectiveness of the document. This score is generated by identifying the key elements that should be present in a well-written resume, such as work experience, education, skills, and achievements, and comparing them to industry standards.

In addition to the score, the project provides personalized remarks and tips to help users improve their resumes. These remarks highlight areas of strength and weakness in the content of the resume, and provide suggestions for improvement. For example, if the resume is lacking in specific skills or achievements, the project may suggest that the user include more quantifiable achievements or skills to make the resume more impactful. Finally, the project includes an admin side that stores all resumes and provides data visualization tools to track usage trends. This allows us to identify which job applicants are most using the platform, and how they are utilizing its features.

Keywords: Resume analysis, Named entity recognition (NER), Pyresparser, Spacy framework, Data visualization.

Abbreviations: NER: Named Entity Recognition NLP: Natural Language Processing.

1. Introduction

Hire Hive is a sophisticated resume analyzer tool that utilizes natural language processing (NLP) and the Pyresparser library built on the Spacy framework to assist job seekers in optimizing their resumes. With the ability to take resumes in PDF format as input, Hire Hive extracts relevant information using named entity recognition (NER) and analyzes the resulting data to provide users with a comprehensive score that reflects the quality and effectiveness of their resumes.

To calculate the score, Hire Hive first performs NER to identify and classify various entities in the resume, such as education, work experience, and skills. Then, it uses Pyresparser to extract this information and analyze it based on various criteria, such as completeness and consistency, to produce a final score. This score is then used to provide personalized tips and remarks to the user, indicating what important things are missing from the uploaded resume and how to improve it.

One of the key advantages of Hire Hive is its utilization of the Spacy framework, a powerful NLP library that provides efficient and accurate processing of large volumes of text data. This enables Hire Hive to perform NER with high accuracy and quickly extract relevant information from the resume.

Moreover, Spacy provides a wide range of features, such as tokenization, parsing, and lemmatization, that enable Hire Hive to perform sophisticated text analysis and produce meaningful insights for the user.

Overall, Hire Hive is a cutting-edge resume analyzer tool that leverages the latest advances in NLP and Spacy framework to help job seekers optimize their resumes. Its use of Pyresparser and personalized remarks, combined with its accurate and efficient NER capabilities, make it a valuable resource for anyone looking to improve their professional brand and increase their chances of success in the job market.

2. Literature survey

Table 1. Comparison of Methodologies

S. No	Name	Methodology	Advantages	Drawbacks
1	RESUME PARSER USING NLP	Machine learning-based approach to automatically extract and parse information from resumes. It uses Natural Language Processing (NLP) techniques and the development of a web application to implement the resume parser.	By automating the screening process, a resume parser can help reduce the risk of bias or subjective decision-making. Depending on the specific needs of the organization or individual using the resume parser, it can be customized to parse resumes for specific fields.	Depending on the complexity of the resume parser and the amount of data it needs to process, it may require a significant amount of computational resources or time to operate effectively.
2	A Systematic Literature Review (SLR) On The Beginning of Resume Parsing in HR Recruitment Process and SMART Advancements in Chronological Order	aThe authors conducted a comprehensive search for relevant articles. The inclusion criteria for selecting the articles were based on specific keywords such as "resume parsing," "HR recruitment," "machine learning," "deep learning," and "natural language processing.	The paper provides an in-depth analysis of the existing literature related to resume parsing in HR recruitment process, including various techniques and algorithms used, their strengths and weaknesses, and the challenges and ethical considerations associated with them.	1) The study has limited its scope to specific databases, journals, or conferences, which limits the number of papers included in the review. 2)The findings of the review may not be generalizable to all industries or contexts, as the focus was on the HR recruitment process
3	Career Path Prediction System Using Supervised Learning Based on Users' Profile	The study extracted features from the preprocessed data using various techniques such as Principal Component Analysis (PCA), Feature Selection, and Natural Language Processing	he paper proposes a new system that predicts the career path of users, which can be useful for individuals to make informed decisions about their career.	The study has not compared the results with other similar studies, which limits the ability to draw conclusions about the effectiveness of the proposed method.

Table 2. Comparison of Methodologies

S. No	Name	Methodology	Advantages	Drawbacks
4	Career recommendation systems using content-based filtering	The collected data was preprocessed to remove duplicates, irrelevant information, and to standardize the format.	The system does not require external data sources such as job vacancies, which means it can be used even in scenarios where such data is not available.	The system has been trained on a limited amount of data, which could affect the accuracy of the recommendations. The paper does not provide information on the size or diversity of the dataset used.
5	A review of machine learning applications in human resource management.	The methodology used in this paper is a systematic literature review approach, which involves searching for and analyzing published papers related to ML applications in HRM.	The paper presents a structured and organized framework that classifies the different machine learning techniques and applications in human resource management, making it easier to understand and navigate the complex landscape of the field.	The paper focuses more on the applications of machine learning in HRM and does not provide in-depth analysis or evaluation of specific studies or models.

3. Methodology

The application takes a resume as input, either in the form of a PDF or a text document. The application then eliminates any unnecessary information from the resume, including headers, footers, and graphics, and transforms it to plain text. The text is then tokenized into words and sentences, and various fundamental text cleaning operations are carried out, like stop word removal and word stemming. It works by:

- Parsing the Resumes: PyResparser takes in a resume which can be in various different formats like pdf, html, txt etc. and converts it into plain text using external libraries such as PyPDF2, docx2txt, and Beautiful Soup.
- Pre-Processing the text: The resume's plain text is pre-processed using a variety of methods, including sentence and word tokenization, stop-word removal, and stemming, to weed out extraneous material and identify crucial keywords.

The end result is in Figure 2.

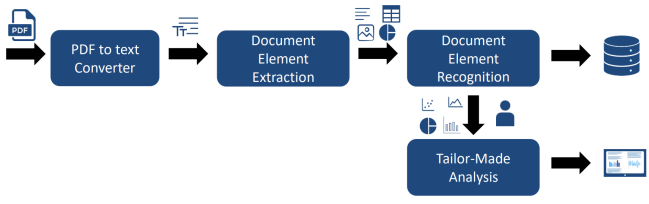

Figure 1. Resume Text Processing

Important data including the candidate's name, email address, phone number, educational background, employment history, and skill set are extracted by the programme using NLP techniques like Named Entity Recognition (NER). PyResparser employs NLP methods like Named Entity Recognition (NER) to extract structured data from the candidate's resume, including their name, email, phone number, educational background, employment history, and abilities. Spacy library is used by PyResparser for NER. PyResparser delivers the findings in JSON format, which can be utilised for additional analysis or storage, after the structured data has been extracted.

3.1 Method part 2: Model training

1. **Data collection:** Collect a dataset of resumes for comparison with the user-uploaded resume.
2. **Preprocessing:** Preprocess the dataset by converting the resumes to a suitable format and removing any unnecessary information such as images and tables.
3. **Feature extraction:** Extract features from the preprocessed dataset using NER and other techniques such as TF-IDF.
4. **Model training:** Train a machine learning model using the extracted features and the user-uploaded resume. The model should be able to identify similarities and differences between resumes.
5. **Cosine similarity:** Use cosine similarity to compare the user-uploaded resume to the dataset of resumes. Cosine similarity is a commonly used metric for measuring the similarity between two vectors.
6. **Evaluation:** Evaluate the performance of the model by comparing the cosine similarity scores of the user-uploaded resume to the dataset of resumes. The evaluation should include metrics such as precision, recall, and F1 score.
7. **Comparison:** Compare the performance of the model trained using the dataset to other methods of resume analysis to determine the effectiveness of the proposed methodology.

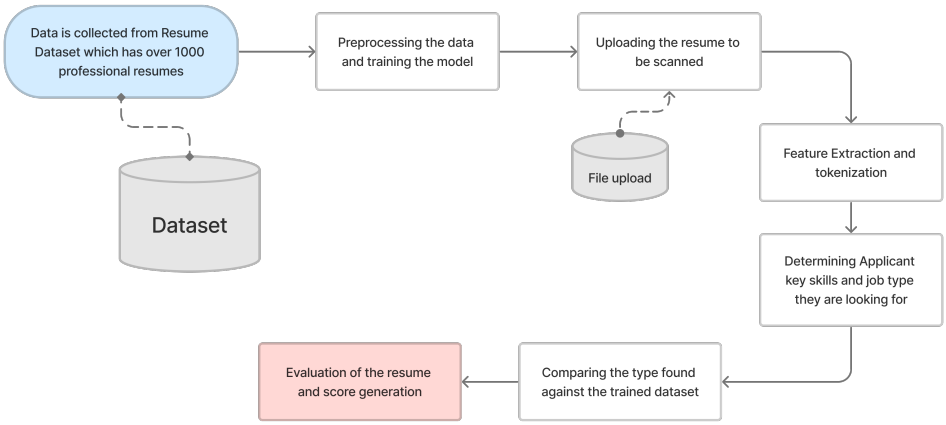

Figure 2. Flow diagram

3.1.1 Mathematical Model

1. This formula is used to calculate a weighted average for relevant numerical data in a candidate's resume, such as their GPA or test scores. The "value" represents the numerical data point and the "weight" represents the relative importance of that data point.

$$\text{Weighted average} = \frac{\sum_{i=1}^{n}(value_i \times weight_i)}{\sum_{i=1}^{n} weight_i} \tag{1}$$

2. This formula is used to calculate the cosine similarity between two vectors, which could be used to analyze and compare the skills or qualifications of different candidates. The "dot product" represents the sum of the products of the corresponding components of the two vectors, and the "magnitude" represents the length of each vector.

$$\text{Cosine similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \, \|\mathbf{B}\|} \tag{2}$$

4. Results and Discussion

The resume analyzer project utilizes natural language processing (NLP) techniques, particularly named entity recognition (NER) using the spacy framework in Python. The spacy framework is a powerful and widely used library for NLP tasks. The NER model in the project is trained to recognize and extract entities such as personal details, educational qualifications, work experience, skills, and achievements from the resumes. The accuracy of the NER model is a crucial factor that determines the quality of the extracted information. The spacy framework provides an accuracy report that shows the precision, recall, and F1 score of the model. These metrics are calculated based on the number of true positives, false positives, and false negatives. Based on the presence or absence of specific entities in the extracted information, such as work experience, education, skills, and achievements, the project's scoring feature assigns scores. Depending on how crucial an entity is to the job opening, the weights given to each one can be changed. The user is then given feedback based on the scores regarding the professionalism and applicability of their resume to the job posting. The project's remarks and tips feature employs a machine learning algorithm that analyses the extracted data and gives the user feedback on how to strengthen their resume. A dataset of resumes and feedback from experienced recruiters are used to train the algorithm. To give the user personalised feedback, it makes use of a variety of NLP techniques, including sentiment analysis and text classification.

Table 3 shows an example.

Table 3. Example table

Resume name	Resume Score	Remarks
John Doe.pdf	75/100	Add more quantifiable achievements and skills.
Jane Smith.pdf	90/100	Well-organized and highlights relevant information.
Mark Johnson.pdf	65/100	More focus needed on work experience and education.
Sarah Lee.pdf	85/100	Relevant details on skills and achievements.
Michael Brown.pdf	70/100	More focus needed on education and skills.

The "Resume Name" column lists the name of the resume file that was analyzed. The "Resume Score" column provides a numerical score out of 100 for each resume, based on the analysis performed by the project. Finally, the "Resume Remarks" column provides concise, professional feedback on each resume, highlighting areas of strength and areas for improvement.

The project's ability to provide remarks and tips based on the uploaded resume is a valuable feature for job seekers. It allows them to identify specific areas where they can improve their resume to increase their chances of being hired. The admin side of the project, which stores all resumes and tracks the most commonly used features, provides valuable insights for employers. This can help them identify trends in job applicants' resumes and tailor their recruitment strategies accordingly.

5. Conclusion

The results have shown that our model can generate a predictive score and categorize the resume uploaded by the candidate. The HireHive project has the potential to be a valuable tool for both job seekers and employers. Its ability to provide feedback and insights based on the uploaded resume can help job seekers improve their chances of being hired, while its tracking and visualization features can help employers optimize their recruitment strategies. One potential limitation of the project is that it relies on NER, which may not accurately capture all relevant information in a resume. Additionally, the tool's ability to compare resumes to others in its database may be limited by the quality and size of its dataset. However, these limitations could potentially be addressed through continued development and improvement of the tool.

Acknowledgement

The team members would like to acknowledge the efforts of our project guide, Dr. Manoov R, his continuous efforts have propelled us to achieve our ideas. We would also like to thank all of our friends who have given valuable inputs in the pursuit of this project. We would also like to thank our college: Vellore Institute of Technology, Vellore for providing us with the infrastructure to perform the project

References

[1] Prasuna Pokharel. *RESUME PARSER USING NLP*. July 2022. DOI: 10.13140/RG.2.2.10323.25127/1.

[2] Aakankshu Rawat, Siddharth Malik, Seema Rawat, Deepak Kumar, and Praveen Kumar. "A Systematic Literature Review (SLR) On The Beginning of Resume Parsing in HR Recruitment Process & SMART Advancements in Chronological Order". In: (2021).

[3] Hrugved Kolhe, Ruchi Chaturvedi, Shruti Chandore, Gopal Sakarkar, and Gopal Sharma. "Career Path Prediction System Using Supervised Learning Based on Users' Profile". In: *Computational Intelligence: Select Proceedings of InCITe 2022*. Springer, 2023, pp. 583–595.

[4] Tanya V Yadalam, Vaishnavi M Gowda, Vanditha Shiva Kumar, Disha Girish, and M Namratha. "Career recommendation systems using content based filtering". In: *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. IEEE. 2020, pp. 660–665.

[5] Swati Garg, Shuchi Sinha, Arpan Kumar Kar, and Mauricio Mani. "A review of machine learning applications in human resource management". In: *International Journal of Productivity and Performance Management* 71.5 (2022), pp. 1590–1610.