International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

A Review of Resume Analysis and Job Description Matching Using Machine Learning

Swanand Modak[1*], Prasanna Shinde[2], Aniket Tiwari[3], Sonali Nalamwar[4]
[1]Department of Computer Engineering
AISSMS College of Engineering
Pune, India
dev.swanandmodak@gmail.com
[2]Department of Computer Engineering
AISSMS College of Engineering
Pune, India
dev.prasanna0102@gmail.com
[3]Department of Computer Engineering
AISSMS College of Engineering
Pune, India
tiwarianiket475@gmail.com
[4]Faculty of Computer Engineering
AISSMS College of Engineering
Pune, India
srnalamwar@aissmscoe.com

Abstract— In the contemporary job market, the effective matching of resumes to job descriptions is a critical facet of talent acquisition. This research paper provides a comprehensive review of the advancements, methodologies, and challenges associated with leveraging machine learning (ML) and natural language processing (NLP) techniques for resume analysis and job description matching. The study surveys the existing literature, synthesizes key findings, and presents a taxonomy of approaches employed in the field. The paper begins by elucidating the significance of efficient resume-job description matching in enhancing the recruitment process. It then delves into the foundational principles of machine learning as applied to human resource management, emphasizing the role of natural language processing, pattern recognition, and semantic analysis in extracting relevant information from resumes and job descriptions. The review encompasses an in-depth analysis of various machine learning algorithms and models utilized in resume parsing, including but not limited to neural networks, support vector machines (SVM), and ensemble methods. Moreover, the paper investigates the incorporation of deep learning architectures, such as convolutional neural networks and recurrent neural networks, for more nuanced feature extraction and representation. Key challenges and limitations associated with current methodologies are thoroughly examined, addressing issues such as the need for large, diverse datasets for robust training. The paper concludes with a discussion on future research directions and emerging trends in the realm of resume analysis and job description matching. This research contributes to the existing body of knowledge by offering a comprehensive synthesis of the current state of machine learning applications in resume analysis and job description matching, providing valuable insights for researchers, practitioners, and industry professionals seeking to optimize talent acquisition processes

Keywords- Resume analysis, job description matching, machine learning, natural language processing, pattern recognition, semantic analysis, classificatio.

I. INTRODUCTION

In the contemporary landscape of workforce dynamics, the intricate task of aligning resumes with job descriptions assumes a pivotal role in talent acquisition. The evolution of recruitment practices in the digital age has accentuated the demand for sophisticated tools capable of efficiently navigating through voluminous resumes and dynamically evolving job requirements. In response to this imperative, the integration of machine learning methodologies has emerged as a transformative force, wielding the potential to markedly enhance the efficacy of candidate identification and selection processes.

This paper undertakes a meticulous examination of the application of machine learning paradigms in the analysis of resumes and the precise alignment of candidates with job descriptions. The burgeoning intersection of artificial intelligence and human resource management has witnessed a paradigmatic shift in recent years, as organizations increasingly turn to advanced technologies to automate and optimize their recruitment workflows. The magnitude of this convergence is underscored by its capacity to redefine traditional approaches, ushering in a more streamlined, data-driven, and efficacious model for talent acquisition.

Recognizing the nuances inherent in the process of resume-job description matching becomes imperative for realizing a recruitment process marked by precision and efficiency. Traditional methods often grapple with the intricacies posed by the voluminous textual information embedded in resumes and the fluid nature of contemporary job requirements. In this context, the application of machine learning, with its innate capacity to discern patterns within extensive datasets, extract

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
***Article Received:** 25 November 2023 **Revised:** 12 December 2023 **Accepted:** 30 January 2024*

semantically rich information, and adapt to evolving contextual demands, stands as a promising solution.

This paper embarks on a scholarly exploration to elucidate the foundational principles, methodological intricacies, and attendant challenges associated with the integration of machine learning in the realms of resume analysis and job description matching. Through this scholarly endeavor, we aspire to contribute meaningfully to the ongoing discourse surrounding the infusion of artificial intelligence into human resource management, discerning its transformative potential in sculpting the trajectory of modern talent acquisition practices.

II. BACKGROUND

A. Traditional Recruitment Process

In today's rapidly evolving job market, efficient and effective talent acquisition has become a critical challenge for organizations. Especially the "war for talent" and the amount of applications for open positions lead to new dimensions in the processing and the selection of candidate profiles [1]. In a recruitment process, the HR department of the organization invites the candidates for interviews for any positions in the organization based on the resumes of the candidates. This HR department traditionally manually evaluates the resumes of the candidates and then decides whether the individual is qualified for the position or not. The department then conducts the interview of the qualified candidates and selects the candidate fit for the position based on the talent the candidate exhibits and also their personality. In the case of mass recruiters and large companies, the task of maintaining the data and details about each candidate and their interview becomes a tedious one [2]. This traditional approach, often reliant on manual resume screening and keyword matching, suffers from several limitations: [1]

- Inefficiency: Sifting through a large volume of resumes is time-consuming and laborious, leading to potential oversights of qualified candidates.
- Subjectivity: Human review can be susceptible to bias based on perceived candidate demographics, leading to unfair assessments and missed opportunities
- Inaccuracy: Traditional keyword matching often fails to capture the full extent of a candidate's skills and experience, resulting in an incomplete understanding of their fit for a role.

B. Challenges in Resume Analysis and Job Matching

In [3] the researchers have identified the following key challenges in the candidate recruitment process:

- Separating the right candidates from the pool: Only the qualified candidates should be considered for the further steps in recruitment. The efficiency of this step can be increased if we can extract the skillset of the candidate and directly compare it with the skillset the job requires. The correct candidates can be immediately filtered in this step itself but the volumes of resumes which are submitted for any job position make it humanly impossible to screen all the resumes.
- Making sense of the resumes: The resumes in the market are not uniform in their structure nor do they have any specific format. The resumes are usually semi structured in their nature. The resumes are sent to the recruiter in varying document formats like .doc, .pdf, .odt, .jpeg, etc. This makes the extraction of information from the resumes very difficult. The process hence may become error prone and a qualified candidate may get missed in it. [4].
- Mapping the candidate information to the job description: We eventually need to map the resumes to the job description and select the fit one for the job position or for further steps in the recruitment.

C. Automation in Recruitment

The task of determining whether a candidate is suitable for a job position requires human intelligence as we need to take into account various factors while making the decision. There are many factors which influence the decision of the recruiters. In these there are many factors whose evaluation cannot be automated such as soft skills of the candidate, accountability of the candidate etc. However we can automate a part of the process and lower the number of candidates the recruiter needs to evaluate [5]. A considerable amount of work has to be done in extracting the information from the resumes of the candidates. This is a very mundane yet very important task and it is natural to assume that it can be offloaded to computer software. However the traditional software fails in making a sense of the words written in a resume. A resume is nothing but a collection of characters and spaces for traditional software. Traditional software fails to understand the semantic meaning of the words written in the resume [6].

Machine learning (ML) offers a transformative potential to address these challenges and revolutionize the recruitment process. By leveraging sophisticated algorithms and natural language processing (NLP) techniques, ML models can automatically analyze resumes and job descriptions, extract relevant information, and perform accurate and objective candidate-job matching. Using ML techniques results in the quick shortlisting of candidates and hence results in saving time and money [7] of the recruiter as well as the candidate

D. Key Tasks in Resume Analysis and Job Matching

In all the papers we analyzed, a common pattern of solving the problem of resume analysis and job description matching using ML, we could identify a pattern of approaching the problem. This pattern can be encapsulated in the following steps which are also the key tasks:

- Resume parsing: Extracting text from resumes of varying formats, types and structures
- Skills extraction: Identifying and classifying the skills and expertise mentioned in the resume, often relying on NLP techniques like named entity recognition and document clustering from the textual data obtained after parsing the resume.
- Job description analysis: Understanding the requirements and responsibilities outlined in the job description, extracting key skills and qualifications.
- Similarity measures: Developing algorithms to quantify the degree of fit between a candidate's profile and the requirements of a particular job
- Matching models: Building ML models that predict the suitability of a candidate for a specific position

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
***Article Received:** 25 November 2023 **Revised:** 12 December 2023 **Accepted:** 30 January 2024*

based on their skills, experience, and other relevant factors.

In the following section we have reviewed various papers pertaining to the topic at hand. We have divided the section into two parts. The first part describes the ideas and the techniques used by the authors of each paper in parsing the resumes and extracting the information out of them. The next part describes the techniques the authors used to actually classify the extraction and match it with the available job descriptions.

III. Related Work

A. Resume Parsing and data extraction

For a trained human, reading a resume or job description and understanding its relevance is not a difficult task. In contrast, a computer system that parses resumes needs to be continuously trained and adapted to deal with the expressive nature of the human language [6].

In [1] the researchers have observed that most of the resumes come to the recruiter in the .pdf format. The researchers take advantage of the structured nature of the resume to extract various segments of data by leveraging layout information. Individual section was processed using independent processors. Named Entity Recognition (NER) was used to label the locations, institutes, dates etc. from the extracted text in individual sections. Stochastic Neighbourhood Embedding (t-SNE) was used for dimensionality reduction and also clustering related skills

In [5] the researchers have used a unique idea of considering the LinkedIn format of resumes as the standard format of resumes. A parser which converts the resumes in different formats into the standard (LinkedIn) format was written. A 100% accuracy was obtained in converting the nonstandard format resumes to standard format. The data in the standard format was then converted to HTML format to maintain the style properties of the extracted data which would have been lost.

A pattern matching model for unstructured document parsing is used in [8] by the researchers. Regular expressions were used to parse the data in the resumes and extract the name, email-id, phone numbers etc. The data of Belthangady staff members was used for extraction. In their system, the stopwords are removed by and stemming is done using the Porter stemmer. This preprocessed data is given to the regular expression module for data extraction.

An ontology based model can be used to parse the resumes and convert them into aontological structural model as demonstrated in [9]. The system they proposed works using a semantic web approach. First they convert the resumes into an HTML format and then remove the HTML tags. Then they implement a sentence end algorithm which indicates that the sentence has ended. Words are split and they are compared with the words in the ontology knowledge base using which the category of each sentence is found out. Using some defined rules, the information is extracted from each sentence

In [10] the researchers use a two-step process for extracting the segment information from the resumes. First they built a classification model to predict whether a text line in the resume is a heading or not. The heading is classified into various heading categories. In the last step the segment under each header is extracted.

In [11] the researchers used a Hidden Markov Model (HMM) to split a resume into various segments. This paper only concentrated on segmentation personal information extraction and educational information extraction. For the classification of the personal information from its block the researchers used SVM multiclass classification because the words in the personal information block are independent. The extraction of information from other blocks is done using the HMM algorithm.

In [4] the researchers have created a segmentation model using CNN. This model segments the information contained in the resume into four segments namely, personal, educational, occupational and others. A pretrained Glove model is used for word embedding in CNN. Then the model is sequence labeled using a CRF based Bi-LSTM-CNN model. The precision, recall and F score values are given in the following table.

In [12] the researchers used the NER module from the Stanford CoreNLP library for extraction of candidate information from the resume.

B. Resume Classification and Job Description Matching

In [6] the researchers used six approaches as baseline models for matching the job descriptions with the information extracted from the resumes. In their work they made the pairs of resumes and job description as ¡r, j¿ and gave it a label as 0 (Not Match) or 1 (Match). A dataset of 3,809 job descriptions and 1,314 resumes was used which effectively created 5,005,026 ¡r, j¿ tuples. Out of these pairs, few pairs annotated manually (Dataset A) and others were annotated using label propagation (Dataset B). The baseline models they used are: (1) Word n-grams, (2) TF-IDF, (3) Bag of Words, (4) Bag of Means, (5) Doc2Vec and (6) CNN. The researchers suggest a novel method of deep Siamese Networks for classification of the resumes. The deep Siamese network consists of a pair of identical CNN that contains repeating convolution, max pooling and leaky rectified linear unit layers with a fully connected layer at the top [Matching resumes to jobs via deep deep siamese networks]. The accuracy of these approaches is given in table 1.

In [13] the researchers used a dataset of 5,298,912 publicly available employee profiles scraped from the internet. They use a novel method of extracting job transitions from the resumes and represent this information as a graph with given rules. The researchers divided the data into three sets, Set I which has data where the current position of the candidate is either in one of the most frequent 100 universities or one of the most frequent 100 companies, Set II where the current position is one of the most frequent 100 companies and Set III where the current position of the candidate is one of the most frequent 25 companies. They use Decision Table - Naive Bayes (DTNB) algorithm to create a predictive model which predicts which job the employee will be transitioning to next. Accuracy of the baseline model and DTNB algorithm is given in table 2.

In [14], the researchers scraped candidate profile data and also job description data from LinkedIn using an API. 95% resume data was used as training data and the remaining 5% was used as test data in the experimentation that the researchers performed. The researchers used the same dataset for experimenting with multiple classification algorithms for classifying the data. The various algorithms used were: (1) CNN, (2) TF-IDF in conjunction with Naive Bayes, (3) TF-IDF in conjunction with SVM and (4) TF-IDF in conjunction with XGB. The precision, recall and F1 scores of all the algorithms are summarized in Table 3

In [End to end resume parsing and finding candidates for a job] the researchers converted 1000 resumes which were in various formats to LinkedIn format which they decided to be the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
***Article Received:** 25 November 2023 **Revised:** 12 December 2023 **Accepted:** 30 January 2024*

standard format for the study. Along with 715 resumes in the standard format the dataset comprised 1715 resumes. The conversion of non-standard format resumes to the standard format was done with 97% accuracy and was performed using heuristic based methods to grab the segments from the raw text extracted. BERT was used in sequence classification. The similarity in job description and the profile of the candidate was evaluated using BERT. The classification was done with an accuracy of 72.77

In [5], the researchers used 400 randomly selected resumes in various formats as their dataset. The text was extracted from the resumes using libraries like pypdf, docx2txt etc. Each extracted row was labeled manually, Label 1 for heading of a section and Label 0 for not a section heading in 333 resumes from the dataset to be used as training data for the parser. This training data was used for training multiple classifiers which classified the segments of text as headings or not headings. A total of seven algorithms were used for the experimentation namely, (1) KNN. (2) Bagging (BAG), (3) Random Forest (RF), (4) Gradient Boosting (GB), (5) Adaboost (AB), (6) SVM and (7) XGB. The results of the experiments are summarized in the following table. XGB was found out to be the most effective algorithm out of all the algorithms used.

IV. Open Challenges and Future Direction

Despite the tremendous progress made in applying machine learning to resume analysis and job description matching, several compelling challenges remain. Addressing these challenges and exploring new directions will be crucial for maximizing the efficacy and ethical development of AI-powered recruitment. We have summarized the open challenges as follows:

- Mitigating Bias and Promoting Fairness: Machine learning models can perpetuate existing biases in hiring practices, potentially putting candidates from specific demographics or backgrounds at a disadvantage
- Enhancing Explainability and Interpretability: The ”black box” nature of certain algorithms can make it difficult to understand how they arrive at candidate-job matches, hindering trust and transparency.
- Handling Dynamic and Unstructured Data: Resumes and job descriptions come in diverse formats and lengths, with varying degrees of structure and ambiguity, posing challenges for accurate data extraction and analysis
- Assessing Soft Skills and Cultural Fit: Machine learning models often struggle to accurately assess a candidate’s soft skills, adaptability, and cultural fit, which are crucial for job success.
- Balancing Automation and Human Judgment: While automation can improve efficiency, overreliance on algorithms can devalue human judgment and overlook intangible qualities.
- Ensuring Privacy and Security: Collecting and analyzing sensitive personal data in resumes raises privacy concerns, requiring robust security measures and ethical considerations
- Adapting to Evolving Job Market Needs: Machine learning models need to keep pace with rapid changes in job requirements, emerging skills, and industry trends.

The future direction of work as envisaged by the various authors of the cited works are summarized below:

- Social aspects like job transition patterns of people in the social circles of the candidates can be exploited to understand more about how candidates make decisions in switching jobs. [13]
- CNN can be used to analyze and extract information from resumes which are sent by the candidates in image formats. [14]
- Algorithms like Generative Pre-Trained Transformer (GPT) can be used to increase the size of the dataset synthetically. [14]
- The feature set can be enlarged to include many things like hobbies, publications etc. to get more holistic information about the candidate. [5]

References

[1] T. Zimmermann, L. Kotschenreuther, K. Schmidt, Data-driven hr-r\’esum\’e analysis based on natural language processing and machine learning, arXiv preprint arXiv:1606.05611 (2016). J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] R. Pal, S. Shaikh, S. Satpute, S. Bhagwat, Resume classification using various machine learning algorithms, in: ITM Web of Conferences, Vol. 44, EDP Sciences, 2022, p. 03011.

[3] P. K. Roy, S. S. Chowdhary, R. Bhatia, A machine learning approach for automation of resume recommendation system, Procedia Computer Science 167 (2020) 2318–2327

[4] C. Ayishathahira, C. Sreejith, C. Raseek, Combination of neural networks and conditional random fields for efficient resume parsing, in: 2018 International CET Conference on Control, Communication, and Computing (IC4), IEEE, 2018, pp. 388–393

[5] V. Bhatia, P. Rawat, A. Kumar, R. R. Shah, End-to-end resume parsing and finding candidates for a job description using bert, arXiv preprint arXiv:1910.03089 (2019).

[6] S. Maheshwary, H. Misra, Matching resumes to jobs via deep siamese network, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 87–88.

[7] X. Yi, J. Allan, W. B. Croft, Matching resumes and jobs based on relevance models, in: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 809–810.

[8] M. Anujna, A. Ushadevi, Converting and deploying an unstructured data using pattern matching, American Journal of Intelligent Systems 7 (3) (2017) 54–59.

[9] D. C¸ elik, A. El¸ci, An ontology-based information extraction approach for r´esum´es, in: Pervasive Computing and the Networked World: Joint International Conference, ICPCA/SWS 2012, Istanbul, Turkey, November 28-30, 2012, Revised Selected Papers, Springer, 2013, pp. 165–179.

[10] B. Gunaseelan, S. Mandal, V. Rajagopalan, Automatic extraction of segments from resumes using machine learning, in: 2020 IEEE 17th India Council International Conference (INDICON), IEEE, 2020, pp. 1–6.

[11] K. Yu, G. Guan, M. Zhou, Resume information extraction with cascaded hybrid model, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05), 2005, pp. 499–506.

[12] V. Mittal, P. Mehta, D. Relan, G. Gabrani, Methodology for resume parsing and job domain prediction, Journal of Statistics and Management Systems 23 (7) (2020) 1265–1274.

[13] I. Paparrizos, B. B. Cambazoglu, A. Gionis, Machine learned job recommendation, in: Proceedings of the fifth ACM Conference on Recommender Systems, 2011, pp. 325–328.

[14] S. Ramraj, V. Sivakumar, et al., Real-time resume classification system using linkedin profile descriptions, in: 2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE), IEEE, 2020, pp. 1–4.