RESUME SCREENING AND RECOMMENDATION SYSTEM USING MACHINE LEARNING APPROACHES

Lokesh. S, Mano Balaje. S, Prathish. E and B. Bharathi

Department of Computer Science and Engineering, SSN College of Engineering,
Rajiv Gandhi Salai, Kalavakkam - 603110

ABSTRACT

Candidates apply in large numbers for jobs on web portals by uploading their resumes, due to the rapid growth of online-based recruitment systems. On the other hand, the resume has its formatting style, data blocks, and segments, as well as a variety of data formatting options such as text alignment, color, font type, and font size, making it an excellent example of unstructured data. As a result, filtering applicants for the appropriate position in an organization becomes a difficult task for recruiters. We can use Natural Language Processing (NLP) techniques to extract the relevant information from the resume to save time and effort. Also, a Machine Learning (ML) model is trained to check whether a candidate's skills, experiences, and other aspects are suitable for that particular role. In addition to that, our system will also recommend the other available job roles based on the candidate's skillset.

KEYWORDS

Natural Language Processing, Machine Learning, Logistic Regression, Support Vector Machine & Decision Tree.

1. INTRODUCTION

Although many improvements have been made to ensure the smooth progression of the recruitment process, the process of shortlisting a candidate has not been completely automated or open-sourced. To deal with such situations, an approach combining Natural Language Processing and Machine Learning appears to be a viable option for breaking down the barrier.

Section 2 describes different formats of resumes available and the problems in analysing it. It also portrays the characteristics of a resume screener. Section 3 depicts the existing solutions and the flaws present in a resume screener. The proposed solution of the resume screening system is described in Section 4. Section 5 will give an insight into the data and algorithm used. Section 6 compares the performance of different machine learning algorithms which were used in the proposed system. Section 7 talks about the final result and improvements which can be done to the resume screener.

2. BACKGROUND STUDY

Different Formats of CV/Resume: Although the data formats used in CV/Resumes are not entirely unstructured, it is still difficult to accept them in a standardized format since there is no set of rules for writing a CV/Resume.

Natural Language Processing Approach: With all of the advantages and disadvantages, there has always been a search for an automated process in which employers can quickly select eligible

DOI:10.5121/cseij.2022.12101

candidates and applicants can show their ingenuity by using a single application format to apply to several organizations. To analyze any written documents such as resumes, the potential to interpret unstructured data and extract relevant information from it, as well as the ability to teach the computer, is needed.

Machine Learning Approach: Researchers used Machine Learning in addition to Natural Language Processing to improve the accuracy and correctness of their models. Since there are numerous Machine Learning methods, there are various approaches to train a model and solve problems. Machine learning-based techniques such as Naive Bayes classifiers, Decision trees, and Logistic regression are widely used to decide if something is right or wrong, good or bad.

2.1. Problems in Manual Screening

– Manually screening a large number of resumes takes at least one day.
– If a recruiter considers 4-6 appropriate resumes when going through the initial resumes, chances are that they will not consider the other submitted resumes. This decreases the likelihood of a successful resume being shortlisted.
– Going through each resume is time-consuming, and manually organizing and managing a large number of resumes is challenging.
– It's normal to have some prejudice, wherever there's been human involvement.

2.2. Problem Statement

To design a model that can parse information from resumes of any format and forms and do a much more efficient and effective analysis on those resumes, along with a prediction of the suitability of a candidate for a job role.

Input: Resume of the candidate and the role he/she applies.
Output: The role the candidate is suitable for if any.

2.3. Characteristics of Resume Screener

– Saves Time
– Multiple Format Support
– Smarter Hiring
– Eliminate Bias
– Easy Integration

3. RELATED WORK

3.1. Gen-1 Hiring Systems

The recruiting team will advertise their openings in the media, on television, and in classified ads [1]. Interested candidates will apply for the position by mailing their resumes. After that, the recruiting team will sort and screen the bulk resumes manually, and later the shortlisted applicants were contacted for additional rounds of interviews. Here, finding the right candidate for the right opening is a time-consuming and stressful operation.

3.2. Gen-2 Hiring Systems

When the number of industries grew in number, the need for people to work in them also grew. Certain consulting units stepped in to help with the recruiting needs. They proposed a solution in which the candidate uploads his or her details in a specific format and submits it to the department. The candidates would then be scrutinized based on a set of criteria, and they would be ranked accordingly. These agencies served as a link between the applicant and the business.
Many organizations began to provide specific formats or forms, which the applicant must fill out with necessary information before the CV/Resume is evaluated by computer using simple pattern recognition and keyword searching. Although this approach decreased the workload for employers, it greatly increased the workload for candidates, who must maintain various formats for each job they apply for. Additionally, it appears to limit the ingenuity and versatility with which skills and qualifications are written in a CV/Resume.

3.3. Gen-3 Hiring Systems

Candidates will upload their resumes in a specific format (either .pdf or .docx) using a framework. The framework then evaluates these resumes. They are however saved in a particular format to make searching easier. The evaluating method employs a Natural Language Processing algorithm. It reads resumes and converts them into a particular format by understanding the natural language/format used by the candidate. The knowledge base is where all of this new information is saved.

3.4. Related Work

– Turney and Littman suggested a technique in 2002 and 2003 to infer the semantic orientation or evaluative character of a word from its massive 100 billion-word corpus, taking into account the semantic connections with other words, which he referred to as paradigms [2, 3].
– The authors of [4] devised a method that transformed a resume into an ontological structural model, making resume analysis easier in both Turkish and English.
– The writers of [5] & [6] did not extract any of the information from the resume, only focusing on such sections as personal information and education sections, without taking into account the candidate’s experience, skillset, and other aspects.
– In this work, the authors of [7] extracted manual and cluster features to train a Chinese word2vec, and concluded that learning-based methods outperform manual rule-based methods for this work.
– To align resumes and jobs, structured relevance models were used [8]. But the outcomes were not what they anticipated from the analysis. Only 1 in 35 related resumes were being put in the top five projected applicants for a job role.
– To fit resumes to job requirements, the authors of [9] used a deep Siamese network. There are almost three times as many job descriptions as resumes in the data collected, and the job description has no domain limitation.

4. PROPOSED SYSTEM

Once the resumes are uploaded by the company in our web interface, they will be stored in our database. Then, from the database, our system retrieves the uploaded resume and starts with the text extraction process. And later the extracted text is tokenized into individual keywords. Our system extracts the necessary information such as experience, skills, education present in the resume by using certain Natural Language Processing (NLP) techniques. The extracted information is used to train the machine learning models. The trained model is used to predict whether the particular candidate is suitable for the role or not. If he/she is not qualified, then the system will recommend other suitable roles for that candidate.

Our model will overcome the issue of submitting a specific file format of resume, which exists in the third generation hiring systems, and will also recommend other vacant roles in the company, for which the candidate might be worthy.

4.1. Steps for the Proposed System

The process of developing the resume screening and recommendation system is depicted in Figure 1.

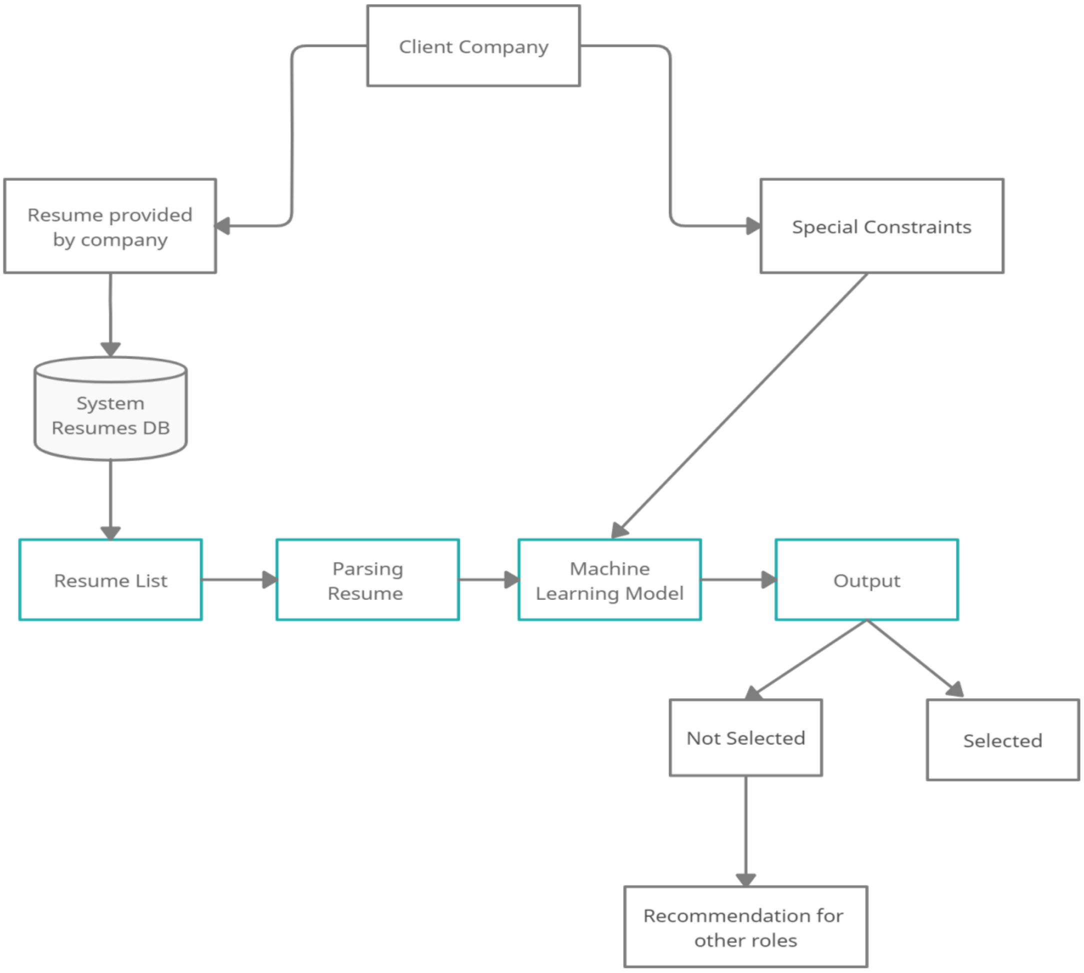

Figure 1. Steps Involved in the Proposed System

5. Experimental Setup

5.1. Dataset

Our dataset consists of resumes taken from kaggle which was available in .docx format.

Table 1. Dataset Classification.

Job Role	Number of Training Data	Number of Testing Data
Java Developer	95	24
Business Analyst	65	17
Project Manager	65	17

The dataset trained upon is present in .docx format. However, it will give the same result when trained with .pdf or .txt or any other format. The number of training and testing data is given in Table 1.

5.2. Machine Learning Algorithm

We tried our resume screener with three different Machine Learning classifiers. They are Logistic Classifier, Support Vector Classifier, and Decision Tree.

6. Performance Analysis

The proposed resume screening system is assessed about Precision, Recall, and F1 score. Precision (p) is the proportion of arguments anticipated by a system that is correct. Recall (r) is the proportion of correct arguments which are anticipated by the system. At last, the F1 score computes the harmonic mean of precision and recall.

6.1. Java Developer

Table 2. Performance of the Resume Screener for Java Developer role, using different Machine Learning algorithms.

Machine Learning Algorithm	Class Label	Precision	Recall	F1-Score
Logistic Regression	0	1.00	1.00	1.00
	1	1.00	1.00	1.00
Support Vector Machine	0	1.00	1.00	1.00
	1	1.00	1.00	1.00
Decision Tree	0	1.00	1.00	1.00
	1	1.00	1.00	1.00

6.2. Business Analyst

Table 3. Performance of the Resume Screener for Business Analyst role, using different Machine Learning algorithms.

Machine Learning Algorithm	Class Label	Precision	Recall	F1-Score
Logistic Regression	0	0.80	0.67	0.73
	1	0.83	0.091	0.87
Support Vector Machine	0	0.50	0.33	0.40
	1	0.69	0.82	0.75
Decision Tree	0	0.50	0.50	0.50
	1	0.73	0.73	0.73

6.3. Project Manager

Table 4. Performance of the Resume Screener for Project Manager role, using different Machine Learning algorithms.

Machine Learning Algorithm	Class Label	Precision	Recall	F1-Score
Logistic Regression	0	0.69	1.00	0.82
	1	1.00	0.50	0.67
Support Vector Machine	0	0.69	0.90	0.78
	1	0.75	0.43	0.55
Decision Tree	0	0.64	0.78	0.70
	1	0.67	0.50	0.57

7. CONCLUSIONS

In this report, we have discussed the detailed design and related algorithms for a resume screener, to decide whether a particular candidate is suitable for the applied role or not. On analyzing the performance of the models, we found that Logistic Regression performs the best for this problem statement. We also found that more dataset is required for making this model work even more efficiently. More attributes can be added to find much better performance. Overall, the system performs pretty well with the current resources. As a part of our future work, we intend to improve the accuracy of our system by collecting more resumes from organizations and training our model for all the available roles. In addition to that, we could also analyze the candidate's information from social networking sites like Facebook, Twitter, Linkedin, so that we can decide more accurately and authentically whether to offer the job or not. Additionally, algorithms such as Naive Bayes, K-Nearest Neighbor, and C4.5 Analysis can be performed, to check if it improves the result.

REFERENCES

[1] Juneja Afzal Ayub Zubeda, Momin Adnan Ayyas Shaheen, Gunduka Rakesh Narsayya Godavari, and Sayed ZainulAbideen Mohd Sadiq Naseem, (2015) "Resume Ranking using NLP and Machine Learning", pp. 1–6.

[2] Turney, P.D, Littman, M.L, (2002) “Unsupervised learning of semantic orientation from a hundred billion word corpus”, In: Technical Report ERC-1094 (NRC 44929), National Research Council of Canada.

[3] Turney, P.D, Littman, M.L, (2003) “Measuring praise and criticism: Inference of semantic orientation from association”, In: ACM Transactions on Information Systems (TOIS), Vol. 21, No. 4, pp.315–346.

[4] Duygu Celik, Askyn Karakas, Gulsen Bal, Cem Gultunca, Atilla Elci, Basak Buluz, and Murat Can Alevli, (2013) “Towards an Information Extraction System based on Ontology to Match Resumes and Jobs”, In: Proceedings of the IEEE 37th Annual Computer Software and Applications Conference Workshops, Japan, pp. 333–338.

[5] Jiaze Chen, Liangcai Gao, and Zhi Tang, (2016) “Information extraction from resume documents in pdf format”, pp. 1–8.

[6] Jie Chen, Chunxia Zhang and Zhendong Niu, (2018) “A two-step resume information extraction algorithm”, pp. 1–10.

[7] Yiou Lin, Hang Lei, Prince Clement Addo, and Xiaoyu Li, (2016) “Machine learned resume - job matching solution”, https://arxiv.org/pdf/1607.07657.pdf

[8] Kun Yu, Gang Guan, and Ming Zhou, (2005) “Resume information extraction with cascaded hybrid model”, In: Proceedings of the 43rd Annual meeting on Association for Computational Linguistics, pp. 499–506.

[9] Saket Maheshwary and Hemant Misra, (2018) “Matching resumes to jobs via deep siamese network”, In: Proceedings of The Web Conference, International World Wide Web Conferences Steering Committee, pp. 87–88.