AI based suitability measurement and prediction between job description and job seeker profiles

Sridevi G.M.* , S. Kamala Suganthi

Atria Centre for Management and Entrepreneurship Bangalore 560024, India

A R T I C L E   I N F O

Keywords:
Suitability measurement
Talent acquisition
Profiles
Artificial intelligence, Quality

A B S T R A C T

Hiring a suitable candidate for a certain job is highly demanding and requires several intense processes. Many organizations face challenges to hire a suitable candidate as they seek specific requirements mentioned in the Job Description (JD). An Artificial Intelligence (AI) based system is developed to measure and predict a suitable candidate from an available Candidate Resume (CR) database. Four clusters are prepared from JD and CR corresponding to primary skills, secondary skills, adjectives, and adverbs. The Jaccard similarity is measured between these clusters and a suitability measure is proposed based on the cluster parameters. Using the three classifiers linear regression, decision tree, Adaboost, and XGBoost the prediction of candidate suitability is performed. To carry out the classification tasks various features are formed by employing the bag of words technique. The maximum average accuracy of 95.14% is achieved for the XGBoost classifier.

1. Introduction

The latest technologies have brought radical change to Human Resource (HR) management practices. Internet connectivity has brought many opportunities for job seekers as well as employers (Zeebareae et al., 2019). The job posting done on various platforms like job portals, social media, and own company websites will attract many job seekers. In finding job opportunities many candidates apply for jobs without geographical constraints. The talent acquisition specialist faces huge challenges to scrutinize relevant profiles among many applicants. This process incurs more manpower cost, time, and hard-to-fill vacancies in-time in the organization. In recent years AI has found many applications. Authors have explored the application of AI in smart cities (Herath & Mittal, 2022). In major fields of smart cities such as risk management, security, education, etc. AI has been applied. AI is used for predictive analytics, which involves making an intelligent decision based on the existing data (Mohbey & Kumar, 2022). The AI techniques such as Genetic Algorithm (GA), and Artificial Neural Network (ANN) were utilized to develop a combined model for prediction in Yadav et al. (2022). AI is extensively applied to the security architecture of IoT with blockchain (Latif et al., 2022). AI also has been very effective in HRM. A study on the challenges related to advancements in HR for industry 4.0-led organizations is presented in Malik et al. (2021). The main focus of this study was on positive and negative experiences because of the utilization of AI in their industry. Organizations aim to utilize AI-based methods to select suitable candidates with respect to job requirements. The recruitment process efficiency has significantly increased due to the application of AI-based techniques (Hemalatha et al., 2021; Jha et al., 2020). Many research focuses on techniques using Machine Learning (ML) and text classification. A survey on the application of ML in HRM is carried out in Garg et al. (2021b). It is observed in Garg et al. (2021b) that ML has been applied in areas such as recruitment and performance management and which has considerably increased the efficiency of the HRM.

AI methods are also used in literature to improve employee work efficiency (Chowdhury et al., 2022). The use of AI in various HR functionalities such as recruitment, training, talent management, and retention was addressed in Ruby and Merlin (2018). A systematic literature survey is carried out in Votto et al. (2021) on the development of AI in HRM and HR Information Systems (HRIS). The authors have summarized the tactical HRM components and their applications in HRM. While the talent acquisition process Natural Language Processing (NLP) based parser was developed in Ponnaboyina et al. (2022) and Sanyal et al. (2017) to perform the auto-filling of the applicants’ forms. The auto-filled profiles were stored in a database after the candidates’ approval. The time-consuming task of arranging and sorting resumes based on their skill was addressed in Dixit et al. (2019). In their research, resume sorting was carried out using AI-based methods. The effective utilization of AI for HR was evaluated in Vedapradha et al. (2019). In their work, various variables were assessed, and then the employee performance was predicted using the multiple linear regression methods. A systematic survey on the utilization of text mining for services management was conducted in Kumar et al. (2021). Researchers have identified the dominant themes and their relationships using big data analytical tools. The impact of AI on HR processes mainly in the region of UAE has been carried

* Corresponding author.

https://doi.org/10.1016/j.jime.2022.100109
Received 5 May 2022; Received in revised form 11 August 2022; Accepted 12 August 2022
Available online xxx
2667-0968/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
out (Singh & Shaurya, 2021). AI-based software for the hiring process for Indian software companies was elaborated on in Nawaz (2019).

Koch et al. have carried out a study on the impact of COVID-19 on the public sector labor market in Germany (Koch et al., 2021). The results were analyzed on the database of e-recruitment providers using quantitative text mining and descriptive statistics. The analysis of job advertisements in Industry 4.0 was carried out using text mining in Pejic-Bach et al. (2020) and extracted knowledge from these advertisements. Zhang et al. (2015) have used various attributes of the knowledge workers and developed a position matching system. Researchers (Lin et al., 2016) have proposed job searching which utilizes ML-based techniques to semantically search the job positions. Three different ML techniques such as features using unsupervised, base classifiers, and combined methods were employed. An AI-based intelligent pulse application was developed in Garg et al. (2021a) by researchers to provide an efficient way to evaluate a large number of survey comments and detect actionable insights from them. The GA has been used in Wang et al. (2016) to carry out the resume recommendation. The vector form of the resume was performed using the entities such as education, age, and then JD. Using GA the resume matching model was developed which takes the user's demand into account while resume matching. A recommendation system was developed using ML and text mining in Chou and Yu (2020) for job vacancies based on various traits such as competitiveness, personality, etc. A two-level approach was developed in Roy et al. (2020) for resume recommendation. At the first level, the resume classification was carried out. Once the resumes were classified, the candidates ranked based on the content-based recommendation. A recommender system was developed in Mhamdi et al. (2020) to assist job applicants to select suitable jobs. The clusters of jobs were prepared based on the features of the JDs. Then suitable jobs for a job seeker were recommended using pre-created clusters. A job resume classification system was developed (Zaroor et al., 2017) to organize the available resumes. One of the main challenges faced by job portals is existing of differently structured resumes. This job classification system helps to route the resumes to their job categories. NLP was used by Daryani et al. (2020) and Alamelu et al. (2021) for job resume screening and to suggest the most appropriate resumes with respect to the JD. The survey on resume screening using NLP and ML is carried out in Sinha et al. (2021). The resumes were ranked, and they are used to find the best candidate. The candidate ranking using predictive analysis was given in Koyande et al. (2020). Various candidate traits such as skills, hobbies, strengths, and weaknesses were considered while resume ranking. The CR ranking can help HR managers to identify suitable candidates from a pool of CRs in a short time. A resume ranking system to aid in resume screening was developed in Kadiwal and Revanna (2021). This system matches the content between resume and JD. Then K-Nearest Neighbour an ML technique was applied to pick the top resume. In the online recruitment system, ML-based methods were used to evaluate and rank the candidates’ resumes (Lai et al., 2016). A collection of descriptive features were extracted from the LinkedIn of candidates’ data and measured the personality traits. An ontology-based resume matching with JDs was developed in Phan et al. (2021). The content of the resume and JD ontology graphs were represented, then they are matched to select an appropriate candidate. The systematic literature review was covered in Pandita et al. (2019) on the use of technology for talent acquisition. Also, the hiring process using digitalization was explained along with its impact on organizations. A review of the ethics of applying AI in recruiting and selection is covered in Hunkschroer and Luetge (2022). They selected 51 research articles and various aspects such as ethics, risks, ambiguities, and opportunities were discussed. Big data analytics along with text mining is utilized to identify new sub-management areas in Kushwaha et al. (2021). Various emerging fields in management are identified by using network analysis and NLP. This work also provides excellent future directions for the emerging management areas.

So far AI-based techniques have been developed for resume screening to assist HR managers during the recruitment process. There is still a considerable research gap to utilize the information content of CRs and JDs to decide the most appropriate matching between them. In addition to the primary information of the CRs and JDs, they also include qualitative descriptions about the nature of work and functional abilities or requirements. During the resume screening process, a quantitative measure indicating the match between a CR and JD is highly useful for ranking and selecting suitable candidates. One of the main objectives of our research is to develop a quantitative measurement based on the information content and qualitative description of CRs and JDs. The AI-based techniques are used to extract information content and qualitative descriptions from CRs and JDs and then compute the suitability measurement between a CR and JD. Also, the ML techniques are used to perform the prediction of CRs into suitable classes. The research objectives are:

1. Extract information content and qualitative descriptors of CRs and JDs and form the clusters
2. Develop a suitability measurement between the CRs and the JDs based on clusters
3. To perform the prediction of suitability measurement into significant classes.

In this research, the AI techniques such as text mining, NLP, and ML are utilized to identify and categorize the candidates based on their suitability. The contents from the CRs and JDs are grouped into four clusters of words. A suitability measurement is proposed based on the Jaccard similarity between clusters of CRs and JDs. The ML-based techniques are used to predict the candidate’s suitability such as Most Suitable (MOS), Moderately Suitable (MDS), or Not Suitable (NTS) class. The suitability measurement and its prediction assist HR managers in quickly identifying the MOS candidates for the JDs. The research paper is organized into the following sections. Both the CRs and JDs collections as the dataset is analyzed in Section 2, and the suitability measure and prediction are presented in Section 3. The results conducted on various experiments are given in Section 4, discussion in Section 5 followed by the conclusion covered in Section 6.

2. CRs and JDs dataset analysis

The analysis of the dataset containing CRs and JDs is carried out in this section. There are 14,906 job seeker resumes and eight JD existing in the dataset. This dataset is taken originally from Kaggle (database) and further refinement step is carried out. In Table 1 a few examples of JDs are presented, such as Web Developer, Linux System Admin, C Developer, and AWS Cloud Eng. For a JD various details are provided in the dataset such as job title, details of the company, and city. The description and responsibilities columns elaborate on various tasks to be carried under this job title, and the required skills are given in preference skills. Corresponding to a job title the required qualification is mentioned in the education column. The responsibilities of a web developer consist of developing and designing websites using Javascript. The minimum education requirement and preference job skills are BSc or BE and Javascript, Python as presented table.

In Table 2 a few examples of CRs for various job titles in the dataset are shown. This dataset includes various candidate information like resume title, city, description of work, job experience in that role, education details, the skills of candidates, and certification done by them. Resume title of a computer engineer, AWS solution architect, cognitive automation, and machine learning profiles are shown in the table. Fig. 1 shows the frequencies of job titles present in the CR dataset. In Fig. 1, there are 2092 candidates with a resume title of a software developer. The second-highest job applicants are for the web developer 1302 followed by machine learning with a count of 1143 is part of the CR dataset. There are 93 data scientist candidates in the database followed
Table 1
JD examples.

<table>
<thead>
<tr>
<th>Job Title</th>
<th>Company</th>
<th>City, Maharashtra</th>
<th>Description</th>
<th>Responsibilities</th>
<th>Education</th>
<th>Preference skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web Developer</td>
<td>Brainizen Technologies</td>
<td>Pune,</td>
<td>NONE</td>
<td>Developing and developing Web scripting Java, Handling issues related to the</td>
<td>BSc BE; Master’s degree is a plus Degree</td>
<td>Javascript, Python,</td>
</tr>
<tr>
<td></td>
<td>WIPRO</td>
<td>Pune,</td>
<td>NONE</td>
<td>Linux system</td>
<td></td>
<td>Linuc, Veritas Volume</td>
</tr>
<tr>
<td>C Developer</td>
<td>Cybage Software Private Limited</td>
<td>Pune,</td>
<td>Highly proficient in Microsoft</td>
<td>Design and build C++ code</td>
<td>Bachelor’s degree in Comp</td>
<td>APIs, XSLT, C++</td>
</tr>
<tr>
<td>AWS Cloud Eng</td>
<td>Loylogic</td>
<td>Pune,</td>
<td>Well Expert in AWS</td>
<td>Planning &amp; implement the AWS cloud infrastructure</td>
<td>BE / B. Tech or ME/ M.</td>
<td>AWS, DevOps</td>
</tr>
</tbody>
</table>

Table 2
CRs with different job titles.

<table>
<thead>
<tr>
<th>Resume title</th>
<th>City</th>
<th>Description</th>
<th>Work experiences</th>
<th>Edutions</th>
<th>Skills</th>
<th>Certificates</th>
<th>Additional Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Engineer</td>
<td>Anand</td>
<td>Organizational goals.</td>
<td>Developing</td>
<td>B.E. C.E. in Computer</td>
<td>[HTML], ’CSS’, ’PHP’</td>
<td>()</td>
<td>CRUD API in PHP</td>
</tr>
<tr>
<td></td>
<td>Chennai</td>
<td>AWS certified</td>
<td>websites using</td>
<td>’B.E in EEE’</td>
<td>C++, HTML</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Work IT sector.</td>
<td>HTML</td>
<td>Python 2years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cloud</td>
<td>Integrate with</td>
<td>AWS Solution</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Enggigion</td>
<td></td>
<td>Worked on</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Haryana</td>
<td>NONE</td>
<td>’Web Developer’</td>
<td>Language: C++, web</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chennai</td>
<td>Pressure environment</td>
<td>Cognitive</td>
<td>Diploma in Computer Engg.</td>
<td>’Mahindra pride class’</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Good Python programming</td>
<td>Automation</td>
<td>C, C++, ’PHP’, ’Mahindra pride class’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive</td>
<td>Chennai</td>
<td>’MCA’</td>
<td>B-Tech in Tech</td>
<td>Languages: C++, web development,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automation Machine</td>
<td>Chenni</td>
<td>Environment</td>
<td>Engineer</td>
<td>PHP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning</td>
<td></td>
<td>Good Python</td>
<td></td>
<td>Diploma in Computer Engg.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>programming</td>
<td></td>
<td>C, C++, ’PHP’, ’Mahindra pride class’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>’Mahindra pride class’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Diploma in Computer Engg.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Diploma in Computer Engg.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Suitability measurement and prediction

An AI-based model has been developed to find the most suitable candidate with respect to a JD. Various AI techniques such as the NL, clusters, and distance measurement are used to compute the CR suitability measurement. Further, the prediction of the suitability of CR for given JDs is carried out in this research work. The architecture for resume suitability measurement which is developed in this work is shown in Fig. 1. The CRs and JDs present in the datasets are read in the first step. Various pre-processing steps are followed on CRs and JDs. The tokenization process is carried out on the collection of CRs and obtained the list of words from the resumes. Noise removal is then applied to each CR and JD. The stop word elimination is performed and collected important words and thereafter lemmatization is applied. The lemmatization identifies its base root word from the dictionary. Parts of speech of sentence of the token of CR are determined by the Natural Language Tool Kit NLTK. The processing of the resume and suitability measurement is presented in Fig. 3.

By taking the reference of parts of speech and details given in the dataset, four clusters of words are formed in both CR and JD. The CR and JD have four significant information which is important for resume screening. The primary skills and secondary skills express the skill set required for the job while their functional properties are described by adjectives and adverbs in profiles. Thus we form four clusters of primary skills, secondary skills, adjectives, and adverbs. The clusters formed from JDs and CRs are depicted in Fig. 4. Clusters of primary skills and secondary skills are created from the details available in CRs and JDs. A cluster of primary skills in a candidate’s profile shows his skill set. Whereas primary skills in JD provide the details of essential skills required to perform the task. Also, each JD is supplied with non-preferred skills which form the secondary skills cluster. Correspondingly each CR, the skills not part of primary skills in the JD are obtained to create the secondary skills cluster. The cluster of secondary skills in both represents the auxiliary skills presented by CR and JD. The clusters of adjectives and adjectives are formed using the details in CR such as education, skills, description of the role, and additional information. A cluster of adjectives from the candidate profile reflects the quality of candidates while the quality of work is reflected in the cluster of adverbs. The cluster of adjectives from a JD signifies the quality of the candidate required to perform the job. The functional requirements of the candidates are indicated in the cluster of adverbs of a JD.

The suitability is measured between a JD and a CR using the Jaccard similarity among the four clusters. In general, the Jaccard similarity for two documents $D_{JA}$ and $D_{JB}$ containing sentences and words is defined as in (1),

$$J(D_{JA}, D_{JB}) = \frac{D_{JA} \cap D_{JB}}{D_{JA} \cup D_{JB}}$$ (1)

We define the Jaccard similarity measure between a cluster of CR and JD as given in (2). For a cluster $C_{RC}$ and $J_{DC}$, the Jaccard similarity is given as,

$$J(C_{RC}, J_{DC}) = \frac{C_{RC} \cap J_{DC}}{C_{RC} \cup J_{DC}}$$ (2)

The Jaccard similarity between four clusters of CR and JD are computed using (2). The Jaccard similarity of clusters is the ratio of the number of common words to total words in those clusters. It is also affected by the size of the data. In our research, the preprocessing of text is performed to obtain important keywords and only these keywords are used to compute Jaccard similarity. The Jaccard similarity between the cluster of primary skills and cluster of secondary skills are $J(C_{RP2}, J_{DP3})$ and $J(C_{RP3}, J_{DP3})$ respectively. The Jaccard similarity between the
cluster of adjectives $J(CR_{Adj}, JD_{Adj})$. Then Jaccard is utilized to compute the suitability measure between the CR and JD. We propose the following equation to compute the suitability measure,

$$\text{Suitability} = J(CR_{PS}, JD_{PS}) + J(CR_{SS}, JD_{SS}) + J(CR_{Adj}, JD_{Adj}) \times |CR_{Adj}|$$

(3)

Here $CR_{PS}$ and $CR_{SS}$ are the cluster of primary and secondary skills for CRs. $JD_{PS}$ and $JD_{SS}$ are clusters of primary and secondary skills for JD. $CR_{Adj}$ is the cluster of adjectives in the candidate’s resume. The $JD_{Adj}$ represents the cluster of adjectives in the JD. $|CR_{Adj}|$ represents the number of words present in a cluster of adjectives in a CR. The third term in (3) is multiplied with $|CR_{Adj}|$ to proportionately increase its weightage by the number of adjectives in a CR.

The suitability measurement is computed between JD and CR using Eq. (2) for the entire dataset. We developed suitability measurement prediction using AI-based techniques as shown in Fig. 5. Initially, the CR and JD are obtained from respective databases, and then the prepossessing such as elimination of hashtags, URL, and extra tabs is performed. The stop words removal is then carried out on both profiles followed by tokenization and lemmatization of the content present in each profile. The significant textual features are formed using the bag of words from keywords present in both the profiles. Then prediction of suitability measured into three classes MOS, MDS, and NTS is carried out using the AI-based classifiers.

4. Experimental results

The experiments in this research are carried out on the resume dataset originally collected from Kaggle (database). This dataset includes 14,806 CRs. Each profile contains fields such as the title of the resume, location, description of roles, education, technical skills, certification, and additional information. There are eight JDs obtained from LinkedIn on the title of Machine Learning, Data Scientist, Data Analyst, Embedded Developer, Full Stack Developer, Java Developer, Php developer, and Python Developer. Various details are present in each JD such as Title, Company, City, state, Description, Responsibilities, Education, Preference_skills, Non_Preference skills, Links, Certificate_required, Additional_information.

As described in Section 3, four clusters are formed from primary skills, secondary skills, adjectives, and adverbs from CRs and JDs. Table 3 shows the clusters of adjectives prepared using several JDs and CRs. In row 1, the cluster of adjectives for JD:1 is shown. In rows 2, 3, and 4 clusters of adjectives of three resumes CR:1604, CR:1667, and CR:1721 are given. Similarly, the cluster of adjectives for JD:2 and other resumes are shown in Table 3. The Jaccard similarities between JD:1 and CRs are depicted in Fig. 6(a). The Jaccard similarity between JD:1 and CR:1604 is 0.2857. Likewise, the Jaccard similarity between JD:1 and CR:1667 and CR:1721 are also shown in Fig. 6(b). Jaccard similarities from JD:2 to a corresponding cluster of adjectives are shown. The Jac-
card similarity between JD:2 and CR:1603 is 0.4375 and in the same way, Jaccard similarities for other CRs are depicted. The highest Jaccard similarity of 0.5608 is obtained between JD:2 and CR:1609.

The clusters of primary skills formed from JDs and CRs are shown in Table 4. The JD:1 is shown in row 1 along with its cluster of primary skills. Rows 2, 3, and 4 show the clusters of primary skills of three resumes CR:1604, CR:1667, and CR:1721. Similarly, the clusters of primary skills for JD:2, CR:1603, CR:1609, and CR:1820 are given.

In Fig. 7(a), the highest Jaccard similarity between JD:1 and CR:1604 is 0.6307 while the lowest Jaccard similarity of 0.2743 is observed between JD:1 and CR:1667. The Jaccard similarity between JD:2 and resumes CR:1603, CR:1609 and CR:1820 are shown in Fig. 7(b).

The clusters of secondary skills for a few JDs and CRs are shown in Table 5. Row 1 and row 5 show the cluster of secondary skills of JD:1 and JD:2 and a few CRs are given in Table 5.

The Jaccard similarities between JD:1 and JD:2 with a few CRs are shown in Fig. 8. The highest similarity is 0.172 between JD:1 and CR:1721 while the lowest Jaccard similarity of 0.012 is obtained between JD:1 and CR:1667.

The proposed suitability measure is computed using the Jaccard similarities between four clusters of JDs and CRs from Eq. (3). Table 6 gives the computed suitability measures in the fourth column between JD:1, JD:2, JD:3, and JD:4 with several CRs. The highest suitability of 1.881 between JD:1 and CR:1721 and the lowest of 0.021 between JD:4 and CR:2907 are obtained. The CRs are classified into three classes MOS, MDS, and NTS based on suitability measurement to facilitate the managers to make a quick decision while the resume screening process. The suitability value above 0.6 is considered a MOS class. The CR:1721 is MOS for JD:1 with a suitability value of 1.881 similarities the CR:1867 is MOS for JD:2. The suitability value between 0.6 to 0.1 is considered
Fig. 4. (a) Four clusters of a JD (b) four clusters of a CR.

Fig. 5. Suitability measurement and prediction.

Fig. 6. Jaccard similarity between clusters of adjectives of CRs and (a)JD:1 (b)JD:2.
The CR:1667 is MDS for JD:1 and the suitability value is given in the below table. A suitability value less than 0.1 is considered as NTS for the respective JD. The CR:2907 is NTS for JD:4.

The class distribution for the three classes of CRs with respect to eight JDs is shown in Fig. 9. The percentage of CRs having MOS class in our dataset is 23.5% while the percentage of MDS profiles is 23.4%. There are 53.2% of CRs are classified as NTS in our dataset.

There are eight JDs such as Machine Learning, Data Scientist, Data Analyst, Embedded Developer, Full Stack Developer, Java Developer, Php developer, and Python Developer. We collected CRs from the main dataset which are matching to these titles. Thus we obtained 1049 CRs closer to eight JDs. The suitability is measured between eight JDs and 1049 CRs and they are categorized into three classes such as MOS, MDS, and NTS. In this research, the prediction of CRs into three suitable classes is carried out using AI-based classifiers namely linear regression, decision tree, Adaboost, and XGBoost classifiers. These classifiers are trained on the bag of words feature collected from each CR to perform three-class classification. The classifier performance is tested on 5-fold cross-validation. The average accuracy rates for 5-fold cross-validation are shown in Table 7. A minimum average classification rate of 85.60% is observed for the linear regression. The improvement in the average classification rate is observed for the classifiers such as decision tree, Ad-
abooest, and XGBoost methods. A maximum average classification rate of 95.14% is obtained for the XGBoost classifier.

5. Discussion

Artificial Intelligence is becoming more popular and discovering many day-to-day applications. The efficiency and utilization of AI will be enhanced as more data is gathered from various streams. Text mining and NLP are important fields of AI which addresses the processing, extracting, and making a suitable decision based on textual data. The NLP is widely used in various applications such as chatbots, language translation, auto-correction, text summarization, etc.

5.1. Contribution to theory

One of the major factors which define the performance and growth of an organization is its human resource. Talent acquisition is the phase using which the appropriate and competent employees are brought on board. Many organizations recognized the importance of talent acquisition and invested both time and effort to update their technology. The current advancement in both computation and communication technologies have opened new opportunities which will enhance the talent acquisition process. AI is one such advancement in recent years to support data-driven decisions for the organization.

Our contribution to this research is to customize and utilize the capabilities of AI for the talent acquisition process. Identifying a suitable candidate for the job profile requires many intense processes involving a considerable amount of human effort. In recent years, many researchers have contributed to reducing human efforts by proposing various solutions using AI in talent acquisition, training, and retention (Ruby & Merlin, 2018). A few researchers have focused on improving resume screening and sorting methods. The various characteristics such as skills,
hobbies, strengths, and weaknesses of a candidate were analyzed and the resume ranking was performed (Koyande et al., 2020). Tejaswini et al. have developed a resume sorting method by matching the content of the resume and job description (Kadiwal & Revanna, 2021). In their method text summarization and then matching were carried out using the K-Nearest Neighbor classifier on the textual features. However, the text summarization has introduced information loss. Moreover, both these recent developments do not consider the quality features which are part of a candidate’s resume. Our developed method not only utilizes the various textual features but also qualitative features to perform the matching task.

In our proposed system, four clusters are formed from the candidate’s resume and job profile. Four clusters formed from primary skills, secondary skills, adjectives, and adverbs of profiles. The first two clusters identify the most relevant functional requirements in a candidate’s resume for a given job description. The quality aspects present in the candidate’s resume and job description are represented by creating clusters of adjectives and adverbs. The Jaccard similarity is then computed between each cluster to measure the closeness of a resume with the job description. A suitability measurement is proposed by considering using the Jaccard similarities between the clusters. Thus, suitability measurement reflects the functional requirements and the qualitative traits from the profiles.

5.2. Real-time application

In recent years AI-based technics have found real-time applications. AI has been widely used for language services such as natural language processing, text-to-speech conversion, and language translations. Popular services include Amazon Alexa and Apple Siri are good examples. Many companies are using AI-based interactive chatbots to have engagement with customers. Understanding customer behavior using sentiment analysis is another example of a real-time application. Moreover, many recommendation engines used in real-time E-commerce web portals are developed based on AI. Content organizing and providing the classification of documents is another extensively used real-time application of AI. The documents and journals in a library are automatically categorized using an AI application and spam filtering is another example of a real-time AI application. Similarly, our developed AI-based resume screening has real-time applications. It provides a practical way of identifying the most suitable candidate for the JD. Based on features collected from a resume, our system predicts the most appropriate candidate. Another important aspect of our system is that it can be integrated with the existing HR process and thus they can make their decision regarding candidate shortlisting based on our AI-based method.

5.3. Cost efficiency

There are two main issues while implementing and deploying AI solutions for the HR processes. One is the acceptance of AI for HR solutions and the second is cost efficiency. The acceptance among HR practitioners is mainly due to changes in their working environment as they need to shift from traditional systems to AI-based systems. The second main issue while deploying is cost efficiency, which included various costs such as design and development of AI, training of AI to the HR practitioners, and the maintenance of the AI-based solution. Our AI-based solution is a light weighted solution including the main processing of the document and text from the given profiles. The implementation is carried out in the Python libraries and data required for the training of AI solutions can easily be obtained from the existing CRs and JDs. Our AI solution to resume screening has the most interactive approach and hence it is easier for HR practitioners. The maintenance of our AI system includes updating and training the AI system which is similar to any other AI solution. On the other hand, there is a huge benefit for HR practitioners in terms of saving time while resume screening which is usually a time-consuming task.

5.4. Implication to practice

The AI-based resume matching system is developed using Python 3.7 and the packages such as Pandas, Numpy, and Matplotlib are utilized. The results are evaluated on the 14,905 resumes. The suitability measurement indicates the better quality and functional appropriateness of a resume for a given job profile. The proposed suitability measurement can be used to perform the prediction of a suitable candidate. Various textual features are gathered using the bag of words technique and prediction is carried out in MOS, MDS, and NTS classes. AI-based classifiers are employed to perform this classification. These predicted classes effectively segregate the candidates’ resumes into three categories thus helping HR managers during the selection process. The ranking of the candidates can be performed within each category by using the suitability measurement. Moreover, our developed system can be integrated with a web server to provide resume screening and matching functionalities to HR managers. The proposed system is easy to implement and integrate with the web server making it a cost-effective solution. It also fastens the time-consuming task of resume screening by classifying them into various categories. The HR manager can purportedly identify suitable resumes for the recruitment process.

6. Conclusion and future work

In the talent acquisition process identifying a suitable candidate is a very challenging task. Shortlisting suitable resume from a huge applicant dataset is a tedious process. To overcome these challenges an AI-based system is developed in this research. The JDs and CRs are analyzed and formed four clusters of primary skills, secondary skills, adjectives, and adverbs. Jaccard similarity between the different clusters is computed. The suitability measurement is developed based on Jaccard similarity to evaluate the suitability of a CR with a given JD. This suitability measurement reflects both the closeness of skills of a candidate with skills required to carry out the job and the qualities essential to perform the tasks. The suitability measurement is computed for the entire database. AI-based prediction techniques such as linear regression, decision tree, Adaboost, and XGBoost are used to predict the suitability of the candidates into three classes. The bag of word features is utilized for the classification process. Prediction experiments are conducted and evaluated on 5-fold cross-validation. The linear regression gave an average classification rate of 85.60% and the XGBoost classifier has given an average classification rate of 95.14%. The future work will focus on using social media features related to candidates to form the additional clusters and the utilization of effective textual features for classification.

---

### Table 7

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Classification Rate in%</th>
<th>Misclassification Rate in%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>85.60</td>
<td>14.4</td>
</tr>
<tr>
<td>Regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision Tree</td>
<td>94.47</td>
<td>5.53</td>
</tr>
<tr>
<td>Adaboost</td>
<td>94.78</td>
<td>5.22</td>
</tr>
<tr>
<td>XGBoost</td>
<td>95.14</td>
<td>4.86</td>
</tr>
</tbody>
</table>
Funding

No funding is received to carry out this research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References


