IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Resume Screener Using Machine Learning

Ashwin Praveen Khairnar[1], Shubham Ganpat Khupase[1], Prashik Vikas Agale[1],

Yash Shankarrao Veer[1], Sonali Lunawat[1]

[1] Pimpri Chinchwad College of Engineering and Research, Ravet, Pune

Abstract. Resume screening is an important task in the recruitment process, and machine learning techniques have shown great promise in automating this task. However, most of the existing research in this area has focused on supervised learning algorithms, which rely on labeled data for training. While supervised learning can be effective, it has several limitations in the context of resume screening. First, obtaining labeled data can be time- consuming and expensive, especially for smaller companies with limited resources. Second, labeled data may not always be available or may not be representative of the entire pool of resumes. Finally, supervised learning algorithms may be biased towards the labels in the training data, which can lead to inaccurate or unfair results. To address these limitations, this paper proposes the use of unsupervised learning algorithms for resume screening. Unsupervised learning algorithms do not require labeled data for training, and can identify patterns and structures in the data without external guidance. In the context of resume screening, unsupervised learning algorithms can be used to cluster resumes based on their similarities, and identify important keywords and phrases that can be used to rank and filter

1 Introduction

Nowadays, hiring new talent is a time-consuming and complex procedure for the recruiting or hiring companies. Several resume applications appear for acquisition of the job, and with the exponential rise in the number of new students entering the job market with a wide variety of skill sets, the applications received by hiring teams have increased significantly. The system for shortlisting resumes using supervised learning [1] was being used to simplify the process of structuring the shortlisting process. However, with the recent rapid increase in internetconnectivity and networking, the recruitment process has undergone modifications over time. Hiring managers attract a wide variety of resume applications for the opening, and diversified resumes make it difficult for the supervised machine learning model to work efficiently, leading to inaccurate solutions and suggestionsand ultimately resulting in the failure of the recruitment drive. While supervised learning methods have been widely used for resume screening, unsupervised learning methods [2] have also demonstrated promise. Unsupervised learning may aid in the identification of patterns and correlations in data without the requirement for labelled samples, making it a handy tool for analysing big datasetsof resumes [3]. In this study, we investigate the application of unsupervised learning methods for resume screening, especially clustering [4] and topic modelling approaches. We show how these algorithms may be used to sort resumes based on similarities and highlight relevant themes and abilities. We also examine the possible benefits of utilising unsupervised learning for resume screening and compare the performance of our technique to established methods.

2. Methodology

1) **Dataset:** The "Category" and "Resume" columns make up the Kaggle dataset. A breakdown of each column is shown below:

Category: The category or domain to which the resume belongs is shown in this column. It identifies the industry or subject that the resume is related to, such as data science, testing, finance, etc. Each resume is grouped according to its area of expertise.

Resume: The text taken from the resumes is contained in this column. It comprises each resume's textual content, which may include personal information, educational background, work experience, talents, and other pertinent details.

2) **Data Preprocessing:** Data preprocessing is the process of transforming, cleaning, and preparing raw data for analysis in order to make it more suitable for machine learning and other data-driven processes.

i) Term Frequency-Inverse Document Frequency (TF-IDF) Vectorization [5]: Describe how the textual data from the "Resume" column is converted into numerical features using TF-IDF vectorization. Words are given weights depending on their frequency throughout the corpus as well as their frequency in each resume.

ii) Dimensionality Reduction [6]: Describe how the TF-IDF vectors' dimensionality was reduced using the principle component analysis (PCA) [7]. It preserves the most crucial details while projecting the high-dimensional feature space onto a lower-dimensional.

3. Proposed System

1. **Data Collection**: In digital format resumes are gathered through employment portals, business websites, and other sources. These resumes are available inWord, PDF, or plain text formats, among others.

2. **Data Pre-Processing:** Pre-processing, where the resumes are cleaned, normalized, and converted to a suitable format for the machine learning algorithm to process, is a crucial stage in the proposed system. Stop words, punctuation, special characters, and digits are removed of the resume text as part of the pre- processing. The frequency of each word or phrase is then computed after tokenizing the remaining text**.**

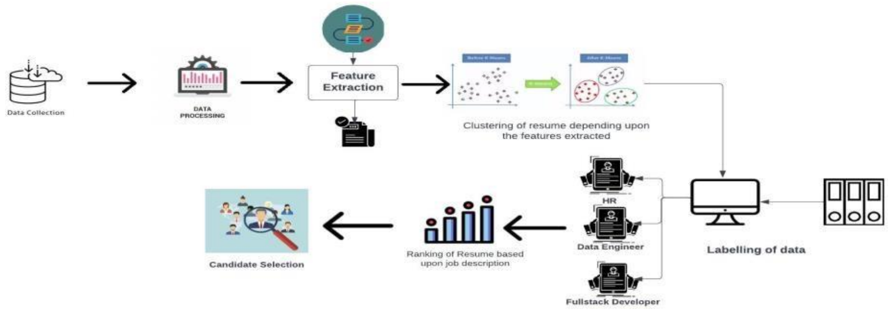


Fig. 1. Proposed System Diagram

3. Feature Extraction: The process of feature extraction consists of converting the content of the resume into a clustering-capable numerical feature vector. The bag-of-words model, which represents the frequency of each word in the resumeas a numerical feature vector, is the most used method of feature extraction. Word embeddings, which express each word's meaning as a high-dimensional vector,are a different method of feature extraction.

4. K-Mean Clustering: Unsupervised machine learning algorithm K-means divides the data into K clusters, where K is an adjustable value. The feature vectors in the proposed system are subjected to the k-means algorithm in order to group the resumes according to how similar they are. The feature vectors and cluster centroids are separated from one another by a sum of squared distances that the algorithm seeks to minimize. The Elbow approach, which detects the value of K where the decrease in within-cluster sum of squares begins to level off, is used to identify the best value of K.

5. Labelling of Data: The k-means algorithm is used to cluster the resumes, and the clusters are then labelled according to the job specifications. The cosinesimilarity between the job requirements and each cluster centroid is computed, and the job requirements are represented as a feature vector. The label "job requirements" is given to the cluster with the highest cosine similarity. This step makes sure that each cluster contains a collection of resumes that are appropriate for the position.

6. Ranking of Resume: The resumes in each cluster are graded according to how closely they match the job requirements once the clusters have been labelled. Cosine similarity is used to determine how closely each CV matches the job requirements. The resumes that meet the job requirements the best are given a better ranking, depending on the similarity score.

7. Candidate Selection: The candidates are chosen for further review as the last step depending on their rating. The top applicants may be chosen for additional screening, includes interviews, tests, or assessments.

4. Classification algorithms

The resume data is classified on the basis of two clustering models, and their respective accuracies are monitored.

1. **Kmeans Clustering [8]:** A machine learning approach called KMeans clustering is used to group or cluster comparable data points in a dataset. A given dataset is divided into K clusters using this unsupervised learning technique,where K is the predetermined number of clusters. Each data point in the procedureis repeatedly assigned to the closest centroid or centre of a cluster, and the centroids are then recalculated based on the updated cluster assignments. Until convergence, which happens when the cluster assignments stop changing or a specific number of iterations has been reached, the process is repeated. The algorithm's goal is to reduce the sum of squared distances between the data points and the cluster centroids that are allocated to them. Data mining, image processing, and other industries frequently use KMeans clustering.

2. **Heirarchial Clustering [9]:** A clustering process called hierarchical clustering divides data points into a hierarchy of clusters, with the number of clusters varyingaccording to how similar the data points are. Agglomerative is the algorithm used in heirarchial. Each data point is first taken into account as a separate cluster in the agglomerative technique, and the algorithm gradually combines the most comparable clusters until all the data points are in one cluster.

5. Evaluation

For evaluating the models, we have used silhouette score and inertia.

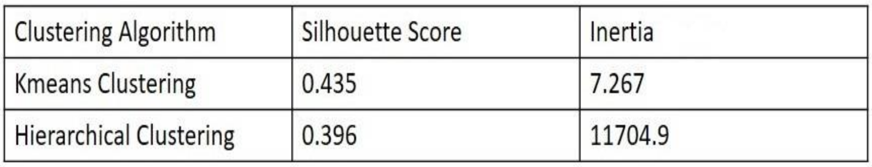
Silhouette score [10]: A criterion used to assess the calibre of clustering findings is the silhouette score. In comparison to other clusters, it gauges how similar a data point is to its own cluster. A higher score indicates a better clustering result. The value goes from -1 to 1.

Inertia metric: The inertia metric computes the sum of squared distances between the centroid of each assigned cluster and each sample in a dataset. It measures how tightly the data points are clustered together within each cluster and quantifies how compact the clusters are.

The dataset was subjected to the k-means method, which produced the development of eight different groups. A silhouette score of 0.435, which indicates acceptable within-cluster similarity and discernible differences between clusters, was obtained during the evaluation of the clustering solution, showing good clustering performance. The clusters' compactness was mirrored in the inertia metric value of 7.267, which showed less within-cluster fluctuation and tighter proximity between data points and the cluster centroids for each cluster.

The silhouette score for hierarchical clustering, however, was 0.396, indicating some overlap or ambiguity as well as meaningful internal cohesiveness and distinctiveness between clusters.

In comparison to k-means, hierarchical clustering's inertia metric value of 11704 suggested greater within-cluster heterogeneity and possibly looser clustering. Hierarchical clustering showed higher spread and dispersion within the created clusters than the k-means approach, despite the fact that it also captured structure in the data.



Clustering Algorithm	Silhouette Score	Inertia
Kmeans Clustering	0.435	7.267
Hierarchical Clustering	0.396	11704.9

Table1. Silhouette score for clustering algorithms

The visualization in Figure 2 displays the results of the clustering algorithm applied to the dataset. The dataset was divided into eight distinct clusters labeled as 0 to 7, each represented by different colors or markers for clarity.

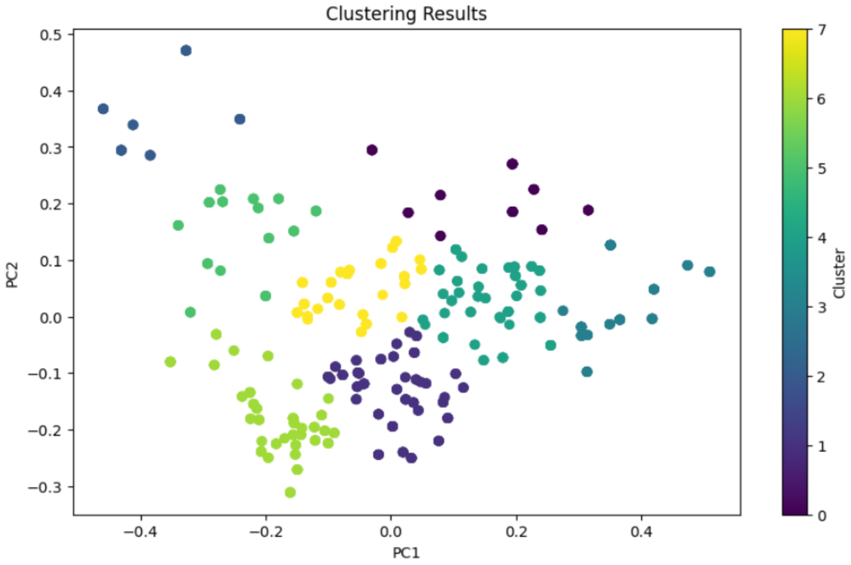


Fig. 2. Results of the clustering algorithm

6. Limitations and Challenges

1. **Subjectivity in Labeling**: The process of labeling clusters based on job specifications introduces subjectivity. Determining the appropriate label for each cluster relies on the cosine similarity between the job requirements and the cluster centroid. However, the selection of the job requirements and the threshold for similarity can vary depending on the hiring manager's perspective, potentially leading to inconsistencies in the labeling process.

2. **Feature Extraction**: Extracting meaningful features from resumes can be challenging. While methods like the bag-of-words model and word embeddings are commonly used, they may not capture the full complexity of a candidate's qualifications. The choice of features and their representation can impact the clustering results and ultimately the accuracy of the system.

3. **Handling Noisy or Irrelevant Data**: Resumes may contain noisy or irrelevant information that can affect the clustering process. Pre-processing techniques, such as stop-word removal and normalization, can help mitigate this issue. However, there is still a possibility of losing valuable information or introducing biases during the pre-processing stage.

7. Future Scope

Some of the future scopes are as follows:

1. **Integration of Supervised and Unsupervised Approaches**: The combination of supervised and unsupervised learning methods holds potential for improving the accuracy and effectiveness of resume screening. Supervised techniques can be used to provide initial labeled data for training unsupervised models, enabling the system to benefit from both labeled and unlabeled samples. This hybrid approach could enhance the system's ability to handle diversified resumes and improve the clustering and labeling process.

2. **Incorporation of NLP Techniques**: Natural Language Processing (NLP) techniques can enhance the feature extraction process and enable a deeper understanding of resume.

3. **User Feedback and Iterative Improvements**: Gathering feedback from hiring managers, recruiters, and candidates is essential for improving the effectiveness of the system. Conducting user studies and collecting feedback on the clustering results, labeling accuracy, and overall usability can help identify areas for improvement. This iterative feedback loop can lead to a more refined and user- centric resume screening system.

8. Conclusion

In this work, we looked into the use of the k-means and hierarchical clustering methods to categorise resumes. The goal was to create a system that can effectively classify resumes into groups based on their similarities, which can be useful for tasks like screening job applications and other duties linked to recruitment. We utilised the silhouette score metric, which gauges cluster quality based on the separations between data points and the clusters to which they are assigned, to assess the performance of clustering. Our findings demonstrate that both algorithms are capable of clustering resumes, but in terms of silhouette score, k-means clustering outperformed hierarchical clustering. This shows that activities requiring a greater degree of cluster separation may be better suited for k- means clustering. Overall, the use of clustering algorithms for resume classification has the potential to improve the effectiveness of job application screening and speed the hiring process. To find the best clustering approach for various resume kinds and recruitment contexts, additional study is required. The silhouette score can be supplemented with additional metrics and evaluation techniques to provide a morethorough evaluation of the clustering performance

References

[1] Liu, Qiong & Wu, Ying. (2012). Supervised Learning. 10.1007/978-1-4419-1428-6_451.

[2] Gong, W., & Guo, Z. (2021). An unsupervised machine learning approach to resume screening. Journal of Intelligent & Fuzzy Systems, 40(4), 6725-6735.

[3] Dey, A., & Ghosh, S. K. (2020). An unsupervised machine learning approach for automated resume screening. In 2020 International Conference onArtificial Intelligence and Sustainable Computing for Smart Cities (AISC2)(pp. 264-269). IEEE

[4] Malik, P., & Yadav, R. (2020). Resume Screening Using Clustering Technique. In Proceedings of the 5th International Conference on Computing Methodologies and Communication (pp. 525-533). Springer

[5] Arroyo-Fernández, I., Méndez-Cruz, C., Sierra, G., Torres-Moreno, J.M., & Sidorov, G. (2019). Unsupervised sentence representations as word information series: Revisiting TF-IDF. Computer Speech & Language, 55, 273-291.

[6] Huang, X., Wu, L., & Ye, Y. (2019). A Review on Dimensionality Reduction Techniques. International Journal of Pattern Recognition and Artificial Intelligence, 33(9), 1950017. doi: 10.1142/S0218001419500174

[7] Bro, R., & Smilde, A. K. (2014). Principal component analysis. Analytical Methods, 6, 2812. doi: 10.1039/c3ay41907j

[8] Aristidis Likas, Nikos Vlassis, Jakob J. Verbeek, 2003. The global k- means clustering algorithm, Elsevier S0031-3203(02)00060-2

[9] Fionn Murtagh and Pedro Contreras, 2012. Algorithms for hierarchical clustering. John Wiley & Sons, Inc. 10.1002/widm.53

[10] Ketan R. Shahapure, Charles Nicholas, 2020. Cluster Quality Analysis Using Silhouette Score. IEEE 10.1109/DSAA49011.2020.00096