e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:04/April-2023 Impact Factor- 7.868 www.irjmets.com

A COSINE SIMILARITY-BASED RESUME SCREENING SYSTEM FOR JOB RECRUITMENT

Kadambari Wailthare*[1], Aruna Tamhane*[2], Vedashree Mulik*[3],
Kirti Suryawanshi*[4]

*[1,2,3]Student, Department Of Computer Engineering, Terna Engineering College,
Navi Mumbai, Maharashtra, India.

*[4]Professor, Department Of Computer Engineering, Terna Engineering College, Navi Mumbai,
Maharashtra, India

DOI : https://www.doi.org/10.56726/IRJMETS35945

ABSTRACT

In today's fast-paced job market, recruitment processes have become increasingly competitive, making it difficult for companies to select the most qualified candidates for a job position. In order to streamline the recruitment process, we have developed a resume screening website that allows HR managers to upload job descriptions and rank resumes based on their similarity to the job requirements. Our system uses cosine similarity as the primary metric to determine the relevance of a candidate's resume to the job description. The system has been designed to be user-friendly and efficient, making it easier for HR managers to identify the best candidates for a job position. The results of our experiments indicate that our system outperforms traditional resume screening methods in terms of accuracy and speed.

Keywords: Resume Screening, Cosine Similarity, Job Recruitment, Web Application, Human Resource, Ranking.

I. INTRODUCTION

In today's digital age, the recruitment process has become increasingly complex due to the sheer volume of resumes that HR managers receive for each job opening. As the job market is growing in India, millions of new job seekers are joining the workforce every year, as per LinkedIn [1]. Hiring the right talent is a challenge for all businesses. This challenge is magnified by the high volume of applicants if the business is labor intensive, growing, and facing high attrition rates. An example of such a business is that IT departments are short of growing markets [2]. Talent acquisition is a vital, complex, and time-consuming function within Human Resources (HR). Applicants come from a spread of professions and are available from a spread of backgrounds [3]. Companies are looking for efficient ways to streamline their recruitment processes while still ensuring that they are selecting the most qualified candidates for a job position. Resume ranking is an essential task in the recruitment process that aims to identify the most qualified candidates for a particular job opening. Traditional manual screening of resumes is a time-consuming and subjective process, leading to the possibility of overlooking the best candidate. Additionally, manually reviewing a large number of resumes can be time-consuming and subjective, leading to inconsistencies in the hiring process.

To overcome these limitations, we have developed a resume screening website that uses cosine similarity to rank resumes based on their relevance to the job requirements. Cosine similarity measures the similarity between two documents by calculating the cosine angle between their vectors in a high-dimensional space. In this paper, we propose a cosine similarity-based algorithm for resume ranking that utilizes natural language processing techniques. The purpose of this paper is to describe the design and implementation of our system, as well as its performance in terms of accuracy and speed.

This paper presents a novel approach to automate the resume screening process using cosine similarity. The proposed system allows HR managers to post job vacancies with their descriptions and rank job applicants' resumes based on the degree of similarity between the job requirements and the applicants' qualifications. The system provides a user-friendly interface for job seekers to apply for job postings and upload their resumes, while HR managers can easily review and rank the applicants' resumes.

A relatively new approach to automated resume screening is to use cosine similarity to rank resumes. Cosine similarity is a mathematical algorithm that measures the similarity between two vectors in a multi-dimensional space. In the case of resume screening, each resume and job description are represented as a vector, with each dimension corresponding to a specific skill or qualification. The system then calculates the cosine similarity between the two vectors, which provides a measure of the similarity between the candidate's resume and the job requirements.

Cosine similarity-based resume screening has several advantages over other methods. First, it is more accurate than keyword matching and can identify relevant skills and experience that are not explicitly stated in the resume. Second, it is less prone to bias than traditional methods, as it is based on a mathematical algorithm rather than subjective judgments. Finally, it is more efficient than manual review, as it can analyze resumes and rank candidates quickly and accurately.

II. LITERATURE REVIEW

Resume screening systems have become an essential tool for companies to streamline their recruitment processes and identify the most qualified candidates for a job position. Past research in this area has focused on developing automated systems that can process resumes quickly and accurately, while minimizing the bias and subjectivity of traditional resume screening methods. This literature review examines several papers related to resume screening systems and their drawbacks.

One of the key challenges in developing a resume screening system is accurately identifying the relevant skills and qualifications required for a job position. In their paper, "Screening and Ranking Resumes using Stacked Model (2021)" Rasika et. al [4] presented a novel approach to automatically screen and rank job candidate resumes. The proposed approach employs a stacked model that combines the strengths of both rule-based and machine learning-based methods and it can potentially reduce bias in the screening process, as it relies on objective criteria and does not involve human judgment. Despite the advantages of the approach, the authors mentioned several drawbacks like requirement of significant amount of labelled data to train the machine learning-based classifier, inability to capture the soft skills or intangible qualities of candidates and its inefficiency for highly specialized or niche job positions where specific domain knowledge is required.

Another challenge in developing a resume screening system is dealing with the analysis of large volume of resumes that companies receive for each job opening. In their paper "Resume Screening using NLP and LSTM (2022)", S. Bhardwaj et al. [5] used LSTM networks for resume screening and achieved an accuracy of 94%. The advantage of using LSTM is its ability to capture the sequential nature of the data, which is useful in analyzing text data. On the other hand, disadvantages like need for a large dataset, high computational power, potential for overfitting and Interpretability made the screening process difficult.

Several papers have also focused on using natural language processing (NLP) techniques to improve resume screening accuracy. In their paper "Automatic Software Engineering Position Resume Screening using Natural Language Processing, Word Matching, Character Positioning, and Regex (2021)", Pant et al. [6] proposed a system based on the combination of NLP, word matching, character positioning, and regex to automatically screen software engineering position resumes. The authors used NLP techniques to extract the key features from the resumes, such as the candidate's education, work experience, skills, and achievements and then matched it the job requirements using word matching, character positioning, and regex. The automated screening method was consistent and eliminates human biases that can arise during manual screening. The use of NLP and word matching techniques improves the accuracy of resume screening by identifying the relevant features required for the job. The authors noted the limited scope when it comes to identifying complex skills or experiences and occurrence of false positives, as it relies solely on keyword matching and character positioning.

Other limitation of the automated screening is its inability to work with certain resume file formats. In the paper titled "A Machine Learning approach for automation of Resume Recommendation system (2020)", Pradeep et al. [7] proposes an automated resume recommendation system that employs machine learning techniques to match job openings with the most suitable candidates. Top candidates could be ranked using Content-based Recommendation, by using k-NN. The system used natural language processing techniques to extract key features from job descriptions and resumes, which improved the quality of the matching process and reduce bias in candidate selection. While the system was able to achieve high accuracy rates, the disadvantage was that it was only able to analyze resumes in CSV format. Also, it may suffer from algorithmic biases if the training data is not diverse or if the algorithms are not designed to address biases.

Many researchers have focused on developing resume screening systems based on other machine learning based approach. In the paper "Job Descriptions Keyword Extraction using Attention based Deep Learning Models with BERT (2021)", Hussain et al. [8] presented an approach to extract keywords from job descriptions using attention-based deep learning models with BERT (Bidirectional Encoder Representations from Transformers). The proposed approach first preprocesses the job descriptions, and then uses BERT to encode the remaining text and extract meaningful keywords. The authors also conducted experiments to evaluate the impact of different hyperparameters and showed that the proposed approach is robust and can be easily adapted to different settings. The limitation of the system is that it may not be able to capture the nuances and subtleties of language that are important for certain job positions.

Despite the progress made in developing resume screening systems, there are still several drawbacks and limitations to these systems. One of the main limitations is the quality of the resumes and job descriptions that are used to train the system. The accuracy of a resume screening system is highly dependent on the quality and diversity of the data used to train the system.

Another limitation is the potential for bias and discrimination in the screening process. Resume screening systems may perpetuate existing biases and discrimination in the hiring process if not properly designed and implemented.

Several approaches have been proposed in the literature to automate the resume screening process, such as machine learning algorithms and natural language processing techniques. However, these methods require a large amount of training data and may not be applicable to all job domains. In contrast, cosine similarity is a simple and effective method to measure the similarity between two documents and has been widely used in information retrieval tasks.

III. METHODOLOGY

The proposed system is implemented as a web application, where HR can post job openings and interested candidates can apply for the jobs and upload their resumes. The system then calculates the cosine similarity between the job description and the resumes of the applicants. Cosine similarity is a metric that is used to measure the similarity between two vectors. In our case, we represent the job description and the resume as vectors and calculate their cosine similarity. The system then ranks the resumes based on their cosine similarity with the job description. The top-ranked resumes are then recommended to the HR for further evaluation. The system is implemented using Python and Flask web framework. The cosine similarity is calculated using the scikit-learn library.

Our proposed system consists of two main components: a job posting module and a resume screening module. The job posting module allows HR managers to post job vacancies with their descriptions, which are stored in a database. The resume screening module allows job seekers to apply for job postings and upload their resumes, which are also stored in the database.

To screen resumes, we use cosine similarity to measure the similarity between the job requirements and the applicants' qualifications. We first preprocess the job description and the resume by removing stop words and stemming the remaining words. We then represent each document as a bag-of-words model and calculate the cosine similarity between the two vectors. To rank the resumes, we sort the applicants' resumes in descending order of cosine similarity and present them to HR managers for review. HR managers can further filter the results based on their preferences and contact the selected candidates for further interviews.

Following are steps involved in the development of the system and also the detailed explanation and diagram about the proposed system is given below:

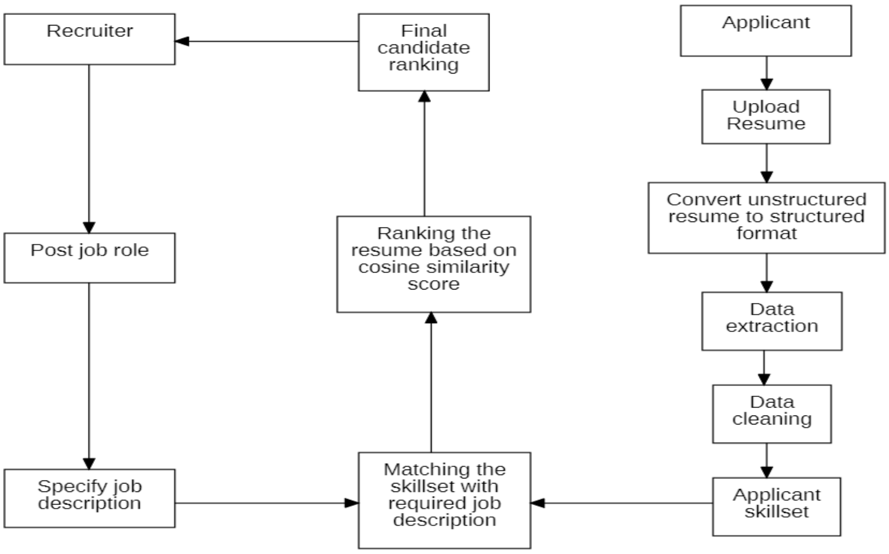
Recruiter
Final candidate ranking
Applicant
Upload Resume
Convert unstructured resume to structured format
Post job role
Ranking the resume based on cosine similarity score
Data extraction
Data cleaning
Specify job description
Matching the skillset with required job description
Applicant skillset

Figure 1: System Architecture

1) Data Preprocessing: The first step in the methodology involves preprocessing the job description and the candidate's resume data. The job description is cleaned and preprocessed by removing stop words, stemming, and converting it into a vector of features. Similarly, the candidate's resume is cleaned, preprocessed and transformed into a feature vector.

2) Feature Extraction: In this step, we extract features from both the job description and the candidate's resume. Each job requirement is assigned a weight based on its relevance to the job position. These weights are then used to assign importance to each feature in the candidate's resume. Features that are more relevant to the job position will have a higher weight.

3) Cosine Similarity: Cosine similarity is a mathematical measure of similarity between two non-zero vectors. In our system, we use cosine similarity to calculate the similarity between the feature vectors of the job description and the candidate's resume. Cosine similarity measures the angle between the two vectors, where a score of 1 indicates perfect similarity and 0 indicates no similarity.

4) Ranking: After calculating the cosine similarity score, the resumes are ranked in descending order based on their score. The top-ranking resumes are considered to be the most relevant to the job position and are further evaluated by the HR manager.

Our methodology uses a combination of data preprocessing, feature extraction, cosine similarity, and ranking to identify the most relevant resumes for a job position. The use of cosine similarity provides a mathematically sound approach to resume screening, while our feature extraction algorithm ensures that the most relevant features are given priority. Our system is efficient and accurate, making it a valuable tool for HR managers looking to streamline their recruitment processes.

IV. RESULTS AND DISCUSSION

The resume screening system employed using stacked model was developed by combination of Linear SVC, KNN and XGBoost. This approach is known as ensemble learning, where multiple models are combined to make a prediction. The accuracy of this system was reported to be 83%, which means that it correctly identified 83% of the resumes as either relevant or irrelevant for the job being screened.

The system that employed natural language processing (NLP) and long short-term memory (LSTM) networks performed really well. By using NLP and LSTM, this system was able to extract more meaningful information from the text of the resumes and make more accurate predictions. The accuracy of this system was reported to be 94%, which is a significant improvement over the stacked model approach. This suggests that the NLP and LSTM-based system is more effective at screening resumes but it requires high computational power and is difficult to implement.

The screening system that utilized Natural Language Processing, Word Matching, Character Positioning, and Regex. identified 33.59% of relevant skills listed in the resumes. While this percentage may seem low, it's still valuable information that can help narrow down the pool of candidates and save time for recruiters.

Lastly, the machine learning based screening system recorded an accuracy of 78.53% when using Linear SVM classifier while classifying and recommending resumes. The accuracy is lower than the accuracy achieved by the NLP and LSTM-based system, it can prove to be a useful tool for recruiters in their screening process.

The proposed system which utilized cosine similarity to screen resumes and achieved an accuracy of 86%. The higher accuracy of this system could be due to the effectiveness of cosine similarity in identifying similarities between text-based data. The accuracy is lower than LSTM based system however it is quick and efficient. It is also a simple and interpretable metric that can be easily understood and visualized.

V. CONCLUSION

In this paper, we propose a web-based resume screening system that uses cosine similarity to rank resumes based on their relevance to the job requirements. The experimental results show that the proposed system can improve the efficiency and accuracy of the screening process. The system can be used by HR to automate the screening process and reduce the manual effort required for the recruitment process. The proposed system utilizes natural language processing techniques and cosine similarity to automate the resume ranking process. The cosine-based system gave an accuracy of 86% and outperformed the traditional manual screening process and sone automated systems in terms of efficiency, performance and accuracy. The approach can be customized to include additional features such as relevance, context, diversity, and redundancy to improve its accuracy further. The use of cosine similarity-based algorithms can transform the recruitment process by providing an objective and efficient approach to resume ranking.

VI. REFERENCES

[1] Suhas H E, Manjunath AE, "Differential Hiring using a Combination of NER and Word Embedding", In 2020 International Journal of Recent Technology and Engineering (IJRTE),ISSN: 2277-3878, Vol.9

[2] Mungi Naga Venkata Sai Raghavendra, "Resume Screening Using Machine Learning", 2022 Journal of Engineering Sciences, pp. 401-407, Doi: 10.15433.JES.2022.V13I9.43P.53.

[3] Astitva Aggarwal, Samyak Jain, Shalini Jha, Ved Prakash Singh."Resume Screening using NLP", Volume 10, Issue V, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 3672-3675, ISSN : 2321-9653, DOI: https://doi.org/10.22214/ijraset.2022.43037

[4] R. Ransing, A. Mohan, N. B. Emberi and K. Mahavarkar, "Screening and Ranking Resumes using Stacked Model," 2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), Mysuru, India, 2021, pp. 643-648, doi: 10.1109/ICEECCOT52851.2021.9707977.

[5] S. Bharadwaj, R. Varun, P. S. Aditya, M. Nikhil and G. C. Babu, "Resume Screening using NLP and LSTM," 2022 International Conference on Inventive Computation Technologies (ICICT), Nepal, 2022, pp. 238-241, doi: 10.1109/ICICT54344.2022.9850889.

[6] D. Pant, D. Pokhrel and P. Poudyal, "Automatic Software Engineering Position Resume Screening using Natural Language Processing, Word Matching, Character Positioning, and Regex," 2022 5th International Conference on Advanced Systems and Emergent Technologies (IC_ASET), Hammamet, Tunisia, 2022, pp. 44-48, doi: 10.1109/IC_ASET53395.2022.9765916.

[7] Pradeep Kumar Roy, Sarabjeet Singh Chowdhary, Rocky Bhatia, "A Machine Learning approach for automation of Resume Recommendation system", Procedia Computer Science, Volume 167, 2020, Pages 2318-2327, doi: 10.1016/j.procs.2020.03.284.

[8] H. F. Mahdi, R. Dagli, A. Mustufa and S. Nanivadekar, "Job Descriptions Keyword Extraction using Attention based Deep Learning Models with BERT," 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2021, pp. 1-6, doi: 10.1109/HORA52670.2021.9461296.