International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 08 | August - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Resume Screening using Machine Learning

Mehul Patel, Savio Rodricks, Samuel Emmatty, Aaron Pereira

Department of Computer Engineering
Fr. Conceicao Rodrigues College of Engineering
Bandra West, Mumbai, Maharashtra 400050

Abstract–The rapid advancement of the technology field has ushered in a wave of employment opportunities across various sectors, necessitating organizations to efficiently sift through a multitude of resumes to identify the most suitable candidates for their specific job requirements. To address this challenge, we propose a research paper that presents a systematic approach to effectively sorting resumes based on company-specific requirements, enabling the identification of the most qualified candidates for specific job roles.

This research paper introduces a cutting-edge solution that harnesses state-of-the-art technologies, including Machine Learning and Natural Language Processing (NLP), in conjunction with cosine similarity. A model has been developed to recognize keywords and extract pertinent information from job requirements provided by recruiting companies. To train this model, a comprehensive database of diverse resumes from various fields has been employed. The model is trained to exclusively search for keywords or requirements specified by the recruiting company, thereby sorting the resumes, accordingly, accepting those that align with the specified criteria, and dismissing those that fall short. The entire process is designed to optimize efficiency and streamline the candidate selection process.

Keywords: Natural language processing, Cosine Similarity, Tfidfvectorizer.

I.INTRODUCTION

To apply for any job the most important document is a resume. A resume gives a lot of information about a particular individual's achievements and various skill sets.

The individual highlights his/her strong points and skills that are required. Various multinational companies receive a huge number of resumes for a particular job. It is very much challenging to know which resume is to be sorted and shortlisted according to the requirements. One method is to check every resume manually and sort the resume accordingly. This method is very time-consuming and can lead to a lot of human errors. Therefore there are problems of less efficiency and greater consumption of time.

Thus we have proposed a system that will quickly find the required skill set by scanning the resume and then sorting them according to the specified requirements by the company. We are going to make use of Machine learning. Machine learning will help to reduce the time-consuming activities of manually screening the resumes.

II. RELATED WORK

There are a few studies and research that are done to address the automated resume screening system. The recruitment process has seen a large amount of development with the introduction of the automated resume screening process. The following section summarizes some of the literary works that are performed in this particular domain.

The work presented as EXPERT (Kumaran. V.S. and Sankar, A., 2013) proposed the use of ontology mapping for screening candidates for the given job description. It includes three phases of operation which are to create candidate ontology, creation of job criteria ontology document , and then lastly mapping of both of these to evaluate which candidates are the most suitable for the job position.

Pradeep Kumar Roy, Sarabjet Singh Chowdhary , and Rocky Bhatia [1] The paper gives an idea about the machine learning approach for the resume recommendation system. Their system works in two phases namely classification based on the similarity index and secondly recommending the resumes based on the similarity index.

Adem Golec and Esra Kahya [2] proposed a model based on fuzzy logic for competency-based employee evaluation and selection of candidates. Their model assesses various corporate factors and guidance based on the goals provided by the organization.

Rahul, Surabhi Adhikari , and Monika [3] proposed an NLP-based machine-learning approach for text summarization. Due to the abundant amount of data, text summarization plays a very vital role in saving time. But the summarized data are not always up to the mark and there is not a specific model developed.

Gabriel Silva, Rafael Ferreira [4] proposed Automatic Text Document Summarization Based on Machine Learning in which they were able to summarize 554 sentences with 98% accuracy.

Begum Mutlu, Ebru A. Sezer , and M Ali Akcayol [5] proposed a multi-text document summarization with a comparative assessment of the features. There are two basic categories in which they have done the summarization which are extractive summarization and abstractive summarization.

Mohamed Abdel Fattah [6] proposed a hybrid machine -learning algorithm for multi-text summarization. In this paper, he has made use of support vector models and the Naive Bayes algorithm for multi-document text summarization.

III.METHODOLOGY

The main aim of this project is to find the right candidates for the specific job position from thousands of resumes. To achieve this goal we have developed a machine learningbased solution.

3.1 Flowchart

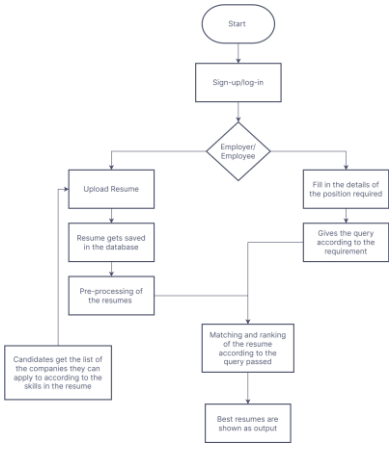

Fig. 1 Flowchart

1. Start: The process begins.

2.Sign/Login Check: The user is prompted to sign in or log in to the system. This step verifies the user's identity as either an employee or an employer.

3.Employee Check: If the user is identified as an employee, the system proceeds to the next step. If not, the process ends.

4.Resume Upload: The employee is given the option to upload their resume. If the employee chooses to upload their resume, the process continues. Otherwise, the process ends.

5.Resume Database: The uploaded resume is stored in a database specifically designed to store resumes.

6.Resume Preprocessing: The system performs preprocessing on the resumes in the database. This typically involves tasks such as the removal of stop words (stop words are the commonly occurring words with little semantic value) and tokenization (breaking the text into individual words or tokens).

7.Employer Query: The employer provides a query specifying their requirements for a potential candidate. This query can include keywords, skills, job titles, or any other relevant criteria.

8.Resume Matching: The system matches the employer's query with the preprocessed resumes in the database. It compares the keywords and criteria specified in the query with the content of each resume.

9.Resume Ranking: Based on the matching process, the system ranks the resumes according to their relevance to the employer's query. Resumes with a higher degree of match are assigned a higher rank.

10.Output to Employer: The system presents the best-ranked resumes to the employer as output. These resumes are the most relevant to the employer's query and are likely to meet their requirements.

11.End: The process concludes.

3.2 Data Collection and Data Preprocessing

It is the first phase of our proposed project.The resumes that are uploaded on the website get stored in our database.The resumes are updated regularly.

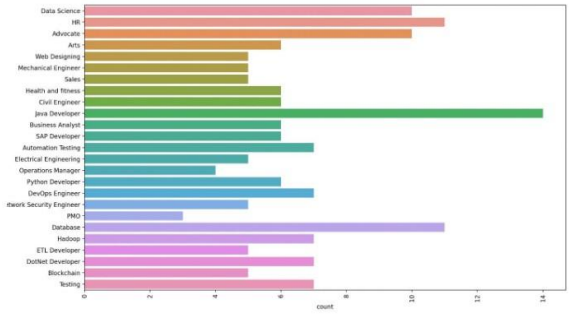

Fig 2.Data representation

The resumes that are collected are then categorized according to various fields like Data Science, HR, Web Designing , and many other fields as shown in Fig1.

In Data Preprocessing the resumes are then cleansed to remove the unwanted characters that are present in the resume. The resumes are then split into tokens with the help of the NLTK tokenizers. Furthermore, the preprocessing steps are done like stopword removal, stemming ,and lemmatization.

3.2.1 Tokenization

Tokenization plays a crucial role in our project as it involves identifying constituent terms or words within a character sequence. By breaking down the larger chunks of text into smaller units called tokens, tokenization enables us to extract valuable information such as word frequency and more. In our project, we have utilized the Natural Language Toolkit(NLTK) as the chosen tool for tokenization.

3.2.2 Lemmatization and Stemming

In the English language, it is seen that A single word manifests in a multitude of forms, each with its distinct variation in sentences and is used according to the grammar rules example: eating, eats, and eaten are different forms of the word eat. Stemming and Lemmatization share a common objective of reducing the word to their base or root forms, but they employ distinct approaches.

Stemming: It is a technique in which we remove the affixes from the words. It is similar to cutting the branches of a tree to its stem.
Lemmatization: The output that we get after doing lemmatization is called ‘lemma’, which is a root word. The output that we get after is a valid word meaning the same thing.

3.3 Candidate recommendation

It is the second phase of our proposed project that aims to give the best recommendations of candidates to employers by skill-based recommendation. It gives the results by using techniques like cosine similarity and TF-IDF vectorizer.

3.3.1 Cosine Similarity

Cosine similarity is a similarity measure that determines the similarity between 2 objects. We have used it to determine to which extent the requirements and the skills mentioned in the resume are similar. It measures the similarity regardless of the size of the documents. Cosine similarity is a symmetric algorithm, meaning that the similarity computed between item X and item Y is equal to the similarity between item Y and item X. This symmetry is represented mathematically as::

$$\cos(\theta) = a \cdot b \div ||a||\,||b||$$

$$a \cdot b = \sum_{1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + . + a_n b_n \quad \text{Here}$$

This formula allows us to calculate the cosine similarity between documents and requirements.

3.3.2 TF-IDF vectorizer

TF-IDF- “Term Frequency - Inverse Document Frequency”. The TF-IDF weight is frequently used in text mining techniques. TF-IDF weight is a quantitative measure that effectively evaluates the significance of a specific term within a given document. The importance increases proportionally to the number of times the word or term is present in the document. Words like what, whom, this ,and that appear frequently in the documents. The TF-IDF value of a term is calculated by multiplying the 2 metrics shown below.

	0	1	2	3	...	7	8	9	10
000	0.0	0.0	0.000000	0.046451	...	0.0	0.00000	0.000000	0.050067
01	0.0	0.0	0.113302	0.000000	...	0.0	0.00000	0.000000	0.000000
02	0.0	0.0	0.113302	0.000000	...	0.0	0.00000	0.000000	0.000000
03	0.0	0.0	0.226603	0.000000	...	0.0	0.00000	0.000000	0.000000
04	0.0	0.0	0.169953	0.000000	...	0.0	0.00000	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...
youremail	0.0	0.0	0.000000	0.075202	...	0.0	0.00000	0.000000	0.000000
yourprofile	0.0	0.0	0.000000	0.102928	...	0.0	0.00000	0.000000	0.000000
zend	0.0	0.0	0.000000	0.000000	...	0.0	0.00000	0.080971	0.000000
zero	0.0	0.0	0.000000	0.000000	...	0.0	0.00000	0.000000	0.000000
zones	0.0	0.0	0.000000	0.000000	...	0.0	0.04058	0.000000	0.000000

[1216 rows x 11 columns]

$$TF - IDF(t, d) = TF(t, d) * IDF(t, d)$$

Fig no. 3

Fig no. 3 shows the TF-IDF scores of the words used. More relevance of the word is reflected by the high TF-IDF score. The value calculated by the above equation is the TF-IDF score of a particular word. 3.3.3 Random Forest Classifier

model report: RandomForestClassifier(max_depth=8, max_features='auto', n_estimators=500, random_state=42):

	precision	recall	f1-score	support
ACCOUNTANT	0.54	0.83	0.66	30
ADVOCATE	0.93	0.50	0.65	26
AGRICULTURE	0.00	0.00	0.00	9
APPAREL	0.89	0.47	0.62	17
ARTS	0.00	0.00	0.00	16
AUTOMOBILE	0.00	0.00	0.00	10
AVIATION	0.68	0.75	0.71	20
BANKING	0.48	0.63	0.55	19
BPO	0.00	0.00	0.00	3
BUSINESS-DEVELOPMENT	0.65	0.42	0.51	31
CHEF	0.64	1.00	0.78	29
CONSTRUCTION	0.88	0.92	0.90	24
CONSULTANT	0.00	0.00	0.00	23
DESIGNER	0.83	0.58	0.68	26
DIGITAL-MEDIA	0.67	0.57	0.62	21
ENGINEERING	0.70	0.73	0.71	22
FINANCE	0.82	0.47	0.60	30
FITNESS	0.27	0.40	0.32	15
HEALTHCARE	0.52	0.64	0.57	22
HR	0.45	0.87	0.59	15
INFORMATION-TECHNOLOGY	0.41	0.91	0.56	22
PUBLIC-RELATIONS	0.58	0.84	0.69	25
SALES	0.45	0.56	0.50	18
TEACHER	0.73	0.92	0.81	24
accuracy			0.60	497
macro avg	0.51	0.54	0.50	497
weighted avg	0.57	0.60	0.56	497

Fig no. 4

The Random Forest Classifier is a well-known supervised machine learning algorithm that is applicable for both classification and regression tasks. It is a part of the ensemble learning approach, which involves combining multiple classifiers to enhance model performance when dealing with complex problems. In the case of the Random Forest Classifier, it utilizes an ensemble of decision trees constructed on different subsets of the given dataset, selected randomly. By leveraging this approach, the Random Forest Classifier can effectively address various challenges in predictive modeling and deliver improved results. It takes the average of all the trees to improve the accuracy of the model.

Fig no. 4 shows the accuracy of the model by random forest classifier. This average is calculated by creating decision trees and then taking the average of the trees and then giving the prediction.

IV.CONCLUSION

In this research paper, we introduce an automated resume screening system designed to address the challenges encountered by recruiters throughout the recruitment process. Our system aims to streamline the screening process and alleviate the difficulties faced by recruiters. Our system uses Natural language processing for extraction of the relevant information and removal of the unwanted data from the resumes which then becomes easier for further processing. Top n resumes of the candidates that are best suitable according to the job description provided With a Summary of the Resume.

V.RESULT

Our project Resume screening system gives the top 3 resumes as the result along with the summary of the resumes of these applicants.

Along with this, the applicant can also see the types of companies he/she can apply to based on the resume provided by them.

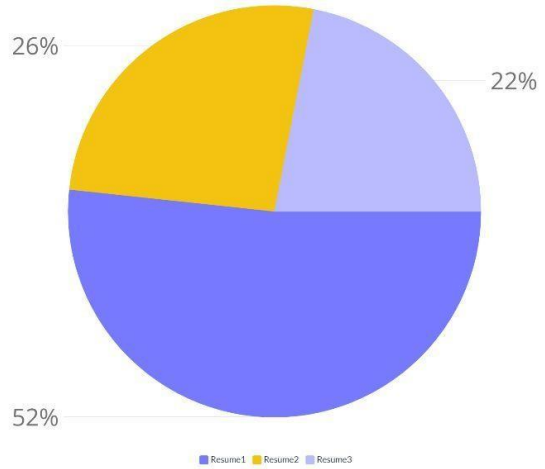

Fig no. 5

Fig no. 5 shows the percentage match of their resumes to the requirements provided by the hiring companies.

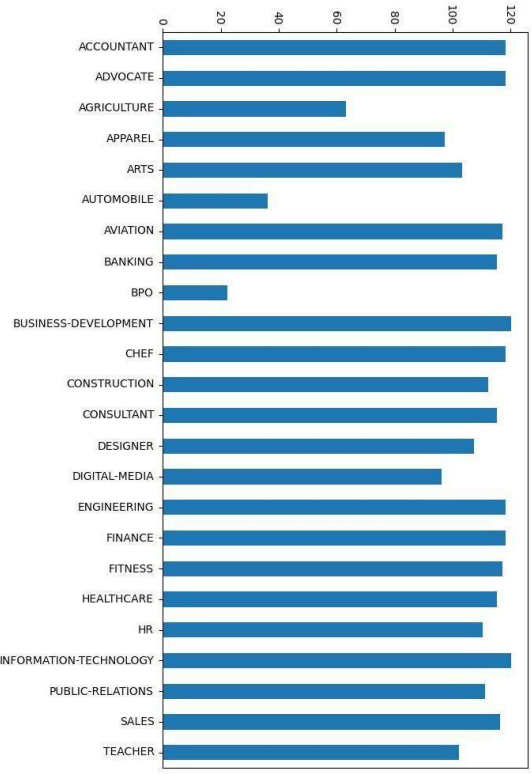

Fig no. 6

Fig no.6 shows the number of types of jobs that are available in the form of a bar graph.

VI.REFERENCES

[1] Al-Otaibi, S.T., Ykhlef, M., 2012. A survey of job recommender systems. International Journal of Physical Sciences 7, 5127–5142.

[2] Ramos, J., et al., 2003. Using TF-IDF to determine word relevance in document queries, in Proceedings of the first instructional conference on machine learning, Piscataway, NJ. pp. 133–142.

[3] Berry M. 2001. Computational information retrieval, Philadelphia: Society for Industrial and Applied Mathematics,121-144.

[4] Gelbukh A., 2014. Computational linguistics and Intelligent Text Processing Berlin, Hiedelburg: Springer Berlin Hiedelburg.

[5] Guo, X, Jerbi, H and O'Mahony, M.P.,2014, September. An analysis framework for content-based job recommendation system in 22nd International Conference on Case-Based Reasoning [ICCBR], Cork, Ireland, 29 September-01 October 2014.

[6] Jivani, A.G., 2011. A Comparative study of stemming algorithms. Int. J. Comp. Tech.Appl,
2(6),1930-1938

[7] V.V Dixit, Trisha Patel, Nidhi Deshpande, Kamini Sonawane, "Resume Sorting using Artificial Intelligence". International Journal of Research in
Engineering Science and Management Volume-2, Issue-4, April 2019

[8]Abeer Zaroor, Mohammad Maree, Muath Sabha, "JRC: A Job Post and Resume Classification System
For Online Recruitment". 2017 International Conference on Tools with Artificial Intelligence.

[9] Shweta C. Dharmadhikari, Maya Ingle, Parag Kulkarni, "Empirical studies on Machine Learning Based Text Classification Algorithms". Advanced Computing: An International Journal, Vol 2, No.6, November 2011.

[10] Gaurav S. Chavan, Sagar Manjare, Parikshit Hegde, Amruta Sankhe, "A Survey of Various Machine Learning Techniques for Text Classification" International Journal of Engineering Trends and Technology, vol. 15, no.6, September.

[11]Vandana Korde, "Text Classification and Classifiers: A Survey" International Journal Of Artificial Intelligence and Applications, Vol. 3, no. 2, March 2012.

[12]Faizan Javed, Qinlong Luo, Matt McNair, Ferosh Jacob, Meng Zhao, Tae Seung Kang, "Carotene: A Job Title Classification System for the Online Recruitment Domain" 2015 IEEE First International Conference on Big Data Computing Service and Applications.