IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

A Machine Learning and NLP Approach for Analyzing Eligibility Based on Resume and CV

Atharva Kulkarni[1], Yash Dewangan[2], Jayesh Padhiar[3], Sumit Kolpekwar[4], Prof. Ravindra Borhade[5]
UG Students, Department of Information Technology[1,2,3,4]
Assistant Professor, Department of Information Technology[5]
Smt. Kashibai Navale College Engineering, Pune, Maharashtra, India

Abstract: *Today Jobs are curved to consider the wishes of the organization. Ideally, job design is destined to be a tired response to maintain the operational model and hence the functions needed. Every job should serve a purpose and contribute to the interests of business outcomes. All changes within the business should consider the impact on jobs and managers should re-design those jobs to suit. Good jobs may even contribute to motivating and inspiring generations to come. Dangerous job design can fail to make that motivation. Job descriptions are in all probability the foremost necessary employment-related documents for a firm. employment description underpins the utilization contract by commencing what management wants from an employee. They are thus a kind of the premise of all employer-employee interactions leading on to appraisal, development, pay and succession discussions. The application that will be presented will try to solve the above concern by acquiring the CV/Resume and related examination scores for the job description the candidate has applied to and by using Text Extraction Techniques the information extracted will be sent in the backend of the application and use the necessary data extracted and run it on the ML model and compare it with some of the top Resumes, Personality and Technical Knowledge evaluation and provide the candidate with good insights on various aspects. The proposed system will first extract the useful terms by using OCR engines from the scanned PDF of CV/Resume, analyze and split the data into categories like Education Qualifications, Experience, Skills, Projects, Candidate Description etc. And based on the data gathered, an evaluation for a particular job description will be done.*

Keywords: Resume Classification, Resume Rating, Job Description Analysis, Candidate Employability Evaluation

I. INTRODUCTION

Analyzing candidate eligibility is the process of analyzing a candidate's CV or Resume. It is a classic and challenging problem being studied for a long time in Resume Classification and Ranking. It is a huge benefit in the field of the corporate hiring process.

The most challenging part is screening which requires domain knowledge and understanding of relevance and applicability of the candidate which tends to be inaccurate due to the pressure of finding the right candidate in less time. This makes the whole hiring process slow, inefficient and costly. Often to reduce the cost and time shortlisting through resumes is kept a priority, which increases the importance of resumes.

CVs and Resumes are two of the common forms of documents used extensively by candidates to show off their skills, Achievements and Experience. Thus, Classification and Ranking of Resumes and CV will lead to benefits in many ways. Classification deals with Classifying resumes into their respective category and Ranking Resumes deals with Ranking all resumes of a particular category and for a particular Job Description. The Classification and Ranking segment have had great success in recent years with the help of machine learning algorithms such as Multinomial Naive Bayes, Support Vector Machine and Cosine Similarity.

Classification of the resume is one of the most important features required to analyze a resume. Based on the resume and category the model has been designed to arrange the resume in the right groups based on the numerical analysis of pre-processed resumes done by the TFIDF technique.

For ranking the resumes according to the job description, There's a need certain algorithms that can take numerical values as input and provide numerical output that can be ranked in an understandable manner. As these algorithms cannot directly

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

process the natural language texts in the sentences, There's a need to convert the text information into numerical values which can help us use these algorithms to perform the ranking operation.

II. Related Work

2.1 First Generation Short-Listing Systems

In this system the recruitment team would publish their vacancies in newspapers, in TV advertisements and through the mouth. Interested candidates would apply for the job by sending their biodata. These bulk of biodatas are sorted and screened by the recruitment team then the shortlisted candidates are called for further rounds of interviews. This entire recruitment process is time-consuming and hectic to find the right candidate for the right job.

2.2 Second Generation Short-Listing System

As the industries have grown in several decades, their hiring needs have rapidly grown. To serve this demand hiring consultancies have come into existence. They offered a solution in which the candidate has to send their information/biodata in a particular format and submit it to the agency. Then these services would search the candidates based on specific keywords. These agencies were middle-level organizations between the candidate and the company. These systems are not flexible as the candidate has to send their resume in a particular format, and these formats change from agency to agency.

2.3 Third Generation Short-Listing System

This work is different from all earlier proposed systems, as in most of the existing systems a job is recommended to the candidates strictly based on their resume context; it leads to low classification accuracy. In order to improve it, the following system was proposed which works in Two phases:

- Classifying the resume in their respective classes, and
- Ranking the candidate's resume based on the job description and their resume content.

III. Gap Analysis

The existing papers show work on limited information which does not include location information, professional skills and description of requirements from both job seekers and employers resulting in this results in lesser accuracy. Some of the existing solutions only work on specific dataset formats. The model takes CVs in CSV format, but in the real world, the CVs are either in .doc, .pdf, etc format. due to the limitation of the data set, the model could not be enhanced to take .doc or .pdf as input. The summarization process is inefficient and results in loss of information. Thereis the scope of fine-tuning this summarization process to ensure minimal information loss. The accuracy of the relevance of keywords is poor, such as Python, Java, and R, in the Data Science domain, Python and Rare more relevant than Java. Some of the existing solutions have a limited dataset which hampers the performance and accuracy.

IV. Proposed System

In the Resume Ranking System, the candidate will give/upload their CV, resume, forms and online portfolios. Now using Pre-processing tools like Tokenization, Lemmatization, and Stop-word removal then filter the data extracted from the resumes and use ML models like TFIDF, Classification algorithms, and Cosine Similarity for feature extractions then after extracting features create Datasets for training the Ranking model and to increase the accuracy of the ranking algorithm this system constantly train the model with input resumes and keep the ranking model updated in the backend and finally system will provide feedback with rich insights for candidates and top k recommended candidates suitable for the job of given Job Description.

In this proposed system an open dataset of resumes from Kaggle. The dataset is labelled with 46 classes It contains a total of 3,674 resumes that were later parsed and pre-processed, divide the dataset into 70-30, 67-33, and 80-20 proportions for training purposes. As the 70-30 proportions attain a better accuracy, consider 70-30% of training and test resumes by different classifiers.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

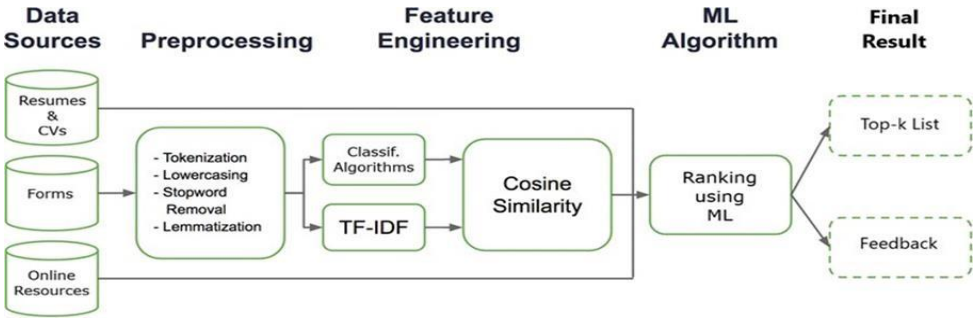

Figure 1: Proposed System Architecture

V. ALGORITHMS

5.1 Data-Preprocessing

Data-Preprocessing is the step-in which data/ raw data gets transformed, or encoded, to bring it to such a way that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted/evaluated by algorithms.

- Data Cleaning: Processes, like filling or removing missing values, smoothing noisy data or resolving inconsistencies and cleansing the data.
- Data Integration: Data consisting of various representations is/are clustered together and the clashes between thedata are taken care of while integration.
- Data Transformation: Data is distributed, assembled and theorized.
- Data Reduction: The objective of this step is to present a contracted model (an individual who models under contract for a specific agency.) in a data warehouse.
- Data Discretization: In this step, the number of values of an uninterrupted characteristic is reduced by the divisionof the range of intervals of characteristics of data.

A. Tokenization:

Tokenization is the process of removing a character sequence and a detailed document unit from a given document. It does away with certain characters and words or string-like punctuation and the chopped units are further called tokens.

B. Lemmatization

The technique of gathering together inflected versions of a word such that they can be studied as a single item is known as lemmatization, or dictionary form. The algorithmic process of identifying the lemma/lemmas of a word/string based on its intended meaning is known as lemmatization in computational linguistics.

C. Regex

A regular expression (shortened as regex or regexp also referred to as rational expression) is a sequence of characters that specifies a search pattern in the text, Such as Phone number, Email, usually such patterns are used by string-searching algorithms for "find" or "find and replace" operations on strings, or for input validation. It is a formal language theory and theoretical computer science technique.

D. Numerical Conversion of Preprocessed Data

It is very important to convert the pre-processed data into vector form. It will be used by the respective Machine Learning model. Machine Learning models can only understand numerical data hence it's important to convert the structured English into vector form. Natural language processing (NLP) is the technique that converts structured English or any other language into vector/numerical form with the proper mapping, such that words can be represented by vectors. Some of the techniques include Bag of Words, Word2Vec, TF-IDF etc. In this system usage of TF-IDF to convert pre-processed data to vectorized numerical values has been done.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

E. TFIDF

A popular natural language processing (NLP) research method that is used in the implementation of the algorithm discussed in this article. The TF-IDF method uses an inverse proportion of the word over the entire document corpus to determine the relative frequency of terms in a single document. The approach employs two elements to determine the value:

- DF - Document Frequency tests the meaning of the text, DF is the number of occurrences in the document set N of the term t. To put it another way, the number of publications in which the word appears is DF.
- TF - In document d, the frequency represents the number of instances of a given word. Therefore, it can be seen that it becomes more relevant when a word appears in the text.
- IDF - It tests how relevant the word is. The main goal of the search is to find relevant records that match the requirement.

In our research and testing of the algorithm of the framework, this method showed good results. TFIDF can be calculated as:

$$a_{ij} = tf_{ij} idf_i = tf_{ij} \times log_2 \left(\frac{N}{df_i}\right)$$

Figure 2: Formula for TFIDF

3.3 Resume Classification

Resume classification is the process of analyzing structured resumes and organizing them into categories based on the numerical analysis of preprocessed resumes done by the TFIDF technique.

A. Naive Bayes

It is a classification technique based on the Bayes Theorem with a hypothesis of independence among predictors. Insimple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to thepresence of any other feature.

B. Support Vector Machines

SVM is a supervised machine learning algorithm that helps in classification problems. It tries to find an optimal boundary known as a hyperplane between different categories. In simple words, SVM does complex data transformations depending on the selected kernel function and based on those transformations, it aims to maximize the separation boundaries between categories.

C. K-Nearest Neighbors

One of the supervised learning algorithms is K-Nearest Neighbors. Unlike classic supervised learning algorithms like the Multinomial Naive Bayes method, K-Nearest Neighbors does not have an independent training stage followed by a stage where the test data labels are predicted using the trained model. Rather, the features of each test data item are compared in real-time to the features of each training data item, and then the K closest training data items are chosen, and the test data item is assigned to the most frequent class among them.

3.4 Ranking Algorithm

For ranking the resumes according to the job description, there's need certain algorithms that can take numerical values as input and provide numerical output that can be ranked in an understandable manner. As these algorithms cannot directly process the natural language texts in the sentences, there's need to convert the text information into numerical values which can help us use these algorithms to perform the ranking operation.

A. Cosine Similarity

A Similarity measure is a metric that helps us determine the similarity of two objects. Cosine similarity is a measure that enables us to find the similarity of two documents regardless of their size. Cosine Similarity represents the position of the documents when plotted on an n-dimensional space, where each dimension represents unique features of the object. Cosine Similarity is a symmetrical algorithm which implies that the results obtained from computing the similarity of item X to

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

item Y is equal to that of the similarity of item Y to item X. Cosine similarity is a technique that considers the cosine of angle between two vectors and converts them into a similarity score between 0 to 100 percent. The Cosine Similarity Algorithm, in which the employer's Job Description can be matched to each one of the resumes in the database, and the top-k most similar resumes will be suggested to the recruiter.

IV. RESULTS

4.1 Data-Pre-processing Results

Removed all the Stops with the accuracy of 90-95%, Converted all the text into lower casing, Grouping of Similar words, classified all the primary information with 100% Accuracy using Regex and Converted Structed English intoVector form using TFIDF

4.2 Classification Results

Testing 3,674 pre-processed resumes, including 3000 well labelled and 674 Non-Labelled resumes using classifiers like SVM, MNB and KNN with variations of TF–IDF. As system achieves the best accuracy using KNN with TF–IDF features (extracted after Tokenization, Lemmatization, stop word removal techniques), it is shown in Table for individual classes. KNN achieves better accuracy in every case than SVM and MNB due to the performance loss forimbalances in positive and negative support vectors in SVM and MN.

4.3 Ranking Results

The development of this project involved a thorough research of various ranking algorithms which included Analytic Hierarchy Process, Cosine Similarity and K-Nearest Neighbors. During research the use case of Cosine Similarity and Analytic Hierarchy Process was measured and after a lot of testing and going through more than 200+ resumes forranking, it was found that Cosine Similarity suited best for the required use case. Hence the ranking functions of this project are made with cosine similarity that takes the output of TF-IDF weights as in input and ranks according to the cosine angle.

Table 1: Ranking report for Cosine Similarity & KNN

Sr.no	Feature Extraction process	Classifier	Accuracy	F1-Score
1	Pre-processing and TF-IDFscores	Cosine Similarity	83%	71%
2	Pre-processing and TF-IDFscores	K-NearestNeighbors	72%	70%

V. CONCLUSION AND FUTURE SCOPE

Develop a full-stack system that extracts applicant information from Resumes/CVs using modern text extraction techniques, then uses a Classification System and a Ranking Engine that uses Machine Learning Algorithms hence, produces a curated list of candidates along with feedback and evaluation. In recent years the number of college graduates is increasing day by day and there is a need for applications which try to solve the problem of candidates and organizations and help candidates grow by learning their weaknesses and strong abilities. In future this application can be designed such that it can also track the social behaviour of a candidate. Also, it can be integrated with the modern-day Application Tracking Services to enhance the User Experience and to increase the intelligence of the application.

REFERENCES

[1]. Text extraction using OCR: A Systematic Review, 2020
[2]. K. Hamad and M. Kaya," A Detailed Analysis of Optical Character Recognition Technology ", International Journal of Applied Mathematics, Electronics and Computers, 3rd September 2016
[3]. Text Mining and NLP
[4]. Sanyal, Satyaki & Hazra, Souvik & Ghosh, Neelanjan & Adhikary, Soumyashree. (2017). Resume Parser with Natural Language Processing
[5]. Machine Learning Approach for Determining the Relevant Skills from Job Description
[6]. Faliagka, Evanthia & Ramantas, Kostas & Tsakalidis, Athanasios & Tzimas, Giannis. (2012). Application of Machine Learning Algorithms to an online Recruitment System

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

[7]. Roy, Pradeep & Chowdhary, Sarabjeet & Bhatia, Rocky. (2020) . A Machine Learning approach for automation of Resume Recommendation s system. Procedia Computer Science. 167. 2318-2327. 10.1016/j.procs.2020.03.284

[8]. Lin, Y., Lei, H., Addo, P.C., Li, X., 2016. Machine learned resume-job matching solution. arXiv preprint arXiv:1607.07657, 1–8

[9]. An integrated e-recruitment system for CV Ranking based on AHP