4th IEEE International Conference on Data Engineering and Communication System (ICDECS)

Resume Shortlisting Using NLP

Dr. Ambareesh S
Department of CSE
JAIN(Deemed-to-be University)
Bengaluru, India
ambareesh.s@jainuniversity.ac.in

Nikhil Kumar Thakur
Department of CSE
JAIN(Deemed-to-be University)
Bengaluru, India
20btrcs167@jainuniversity.ac.in

Ujjwal Bhattarai
Department of CSE
JAIN(Deemed-to-be University)
Bengaluru, India
20btrcs198@jainuniversity.ac.in

Saurav Kumar Yadav
Department of CSE
JAIN(Deemed-to-be University)
Bengaluru, India
20btrcs178@jainuniversity.ac.in

Jay Nath Thakur
Department of CSE
JAIN(Deemed-to-be University)
Bengaluru, India
20btrcs160@jainuniversity.ac.in

Amrit Kumar Mahato
Department of CSE
JAIN(Deemed-to-be University)
Bengaluru, India
20btrcs184@jainuniversity.ac.in

2024 4th International Conference on Data Engineering and Communication Systems (ICDECS) | 979-8-3503-9335-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDECS59733.2023.10502580

Abstract - Recruitment processes have evolved from manual sorting of the resumes to automation with the integration of Natural Language Processing. The benefits of NLP in hiring new employees are examined in this paper, with a focus on how it might increase candidate selection's effectiveness, objectivity, and fairness. This automated approach streamlines the candidates resume evaluation process. The core steps in the system include parsing resume to text, performing NLP of parsed text, creating a NER (Named Entity Recognition) model and using NER model to calculate P, R and F score to generate final score to sort resume based on final score.

Keywords - Natural Language Processing, Resume Screening, NER

I. INTRODUCTION

A. Overview

The resume is the most crucial document an applicant may use to portray themselves when applying for jobs. Technology now drives all of the major sectors. The process of hiring and resume screening is a hectic process as it involves a lot of challenges in managing large volumes of resume, screening of those resumes and extracting the most suitable profiles out of them. There are various methods that can be used to automate the process. Many produced inconsistent results and many require a certain level of manual processing. This paper used an NLP based system. When the resume is uploaded to the system, the document is parsed to texts, natural language processing is done on the parsed text (including segmentation, tokenization, stemming, Lemmatization, part of speech tagging, named entity recognition, parsed text generation), NER model is created and drained based on job description and that same model is used for the resume shortlisting.

B. Problem Definition

The costly and subject to error screening process that results from the traditional hiring process's inability to handle the deluge of resumes. Biases and inconsistencies in manual screening may cause suitable candidates to be chosen or qualified candidates to be overlooked. The effectiveness of automation techniques varies; frequently, they produce inconsistent results or necessitate a large amount of manual intervention. This paper proposes a resume shortlisting system based on Natural Language Processing (NLP) to address these issues. The difficulties of extracting relevant profiles from a large number of resumes, the need for a more effective and uniform screening procedure, and the need for a solution that reduces manual processing while guaranteeing accurate candidate selection based on job descriptions are some of the issues that have been identified.

C. Objectives

By utilizing modern language analysis techniques, resume shortlisting through NLP aims to transform the conventional and frequently laborious hiring process. A sophisticated Named Entity Recognition (NER) model is created, specifically trained on the job description, by parsing the uploaded resumes into structured text and utilizing extensive Natural Language Processing (NLP) techniques like segmentation, tokenization, stemming, lemmatization, part-of-speech tagging, and NER. The following resume shortlisting procedure heavily relies on this NER model. By comparing the entities and skills taken from resumes with those listed in the job description, the objective is to automate and optimize the process of identifying the most appropriate profiles.This approach strives to enhance the testing process's accuracy as well as effectiveness by decreasing gaps and preventing the need for extensive manual processing. This allows for to address the difficulties that come with handling a high volume of resumes.

II. LITERATURE REVIEW

There are hundreds of papers already that show how important is the problem of screening resumes automatically in the hiring domain. The process of shortlisting resumes is the critical step in talent acquisition which aids in selection of better candidates to the particular job role. Automated resume shortlisting saves time, ensures consistency, and reduces bias, leading to efficient and fair candidate selection. Several studies have been conducted and numerous papers have been published over the past few years highlighting the ways to perform resume screening and analyse various machine learning algorithms to compare its performance with respect to algorithms.

979-8-3503-9335-4/24/$31.00 ©2024 IEEE

Table 1. Insight about various papers published in previous years.

Study/ Year	Title	Dataset	Algorithm / Approaches	Remarks
[1] B.Surendiran et al. (2023)	Resume Classification Using ML Techniques	3446 resumes in 48 different professions from a variety of fields.	Various Classification Techniques: RF, DT, KNN, and SVM.	Random forest classifiers gave the highest accuracy (91%).
[2] Rishabh Bathija et al. (2023)	SVM, Weighted KNN, and KNN are compared for resume screening.	Manual collection of resumes from various categories	Evaluating Three Machine Learning Models: KNN, Weighted KNN, and SVM-KNN.	Highest accuracy was achieved by weighted KNN.
[3] S Bharadwaj et al. (2022)	Utilizing LSTM and NLP to Screen Resumes Effectively	25 job categories are addressed by 962 resumes.	Natural Language Processing toolkit	The goal of improvement is to centralize 90% of the picking phase.
[4] Vishnu S. Pendyala et al. (2022)	NLP and ML for full resumes that enhance placement.	Consists of various resumes with job categories	Regression models (KNN regression, Support vector regression),Sentiment analysis module using NLP.	Provides psychological ability, employment scores, and rated resumes.
[5] Tumula Mani, et al. (2022)	NLP-automated resume assessment tool.	Csv(including resumes and required skills for job)	Natural language processing	The entire PDF is converted to plain text using NLP for extensive screening.
[6] Rasika Ransing et al. (2021)	Resume Evaluation and Prioritization Using Stacking Methods	From kaggle (including role and resume)	Stacked categorization XG-Boost, KNN, and Linear SVC.	With 85% accuracy and a validated rating of 0.7523, XG-Boost beats the rivals.
[7] M.F. Mridha, et al. (2021)	Evaluation Job Overviews: Using CNN and ML	IT Job Knowledge Bases Integration.	Random Forest Classifier	The accuracy rate of the BDJOBS site is an impressive 74%.

III. METHODOLOGY

A. Dataset:

A collection of 2452 resume is gathered in the dataset. Each resume includes details like name, contact, email address, education, skills, work experience, project, languages, hobbies, and many other related to the work profile candidate applying for. The work profile belongs to various felids such as engineering, medical, sales, finance, logistics etc.

B. Model Architecture:

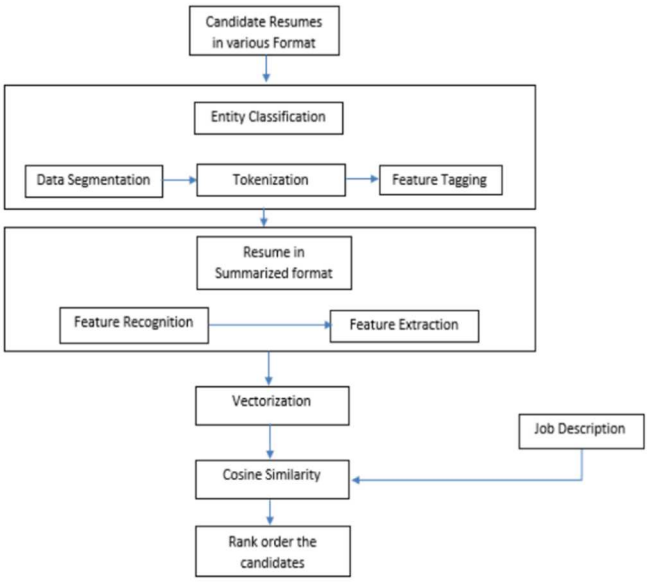

Figure 1. Model architecture of resume short listing

C. Data Pre-processing:

In figure 1, first the texts from the resumes in pdf format are extracted. After that the following steps are followed to process the text obtained from the resumes. The steps are:

1. Segmentation:
The text from the resume is segmented. The text is divided into various segments based on punctuations like full stop (.), comma (,) or line breaks, page components in html.

Example: "(I am frontend developer) – part1, (not a backend developer) –part2

1.

2. Tokenization:
In this step the sentences are separated into individual words. Also, the punctuation or any special characters are removed.

Example: "I am frontend developer," – (I) (am) (frontend) (developer)

3. Stemming:
Stemming is process of obtaining stem word from a word. In this step stem of the word is obtained by removing affixes like "ing", "un" etc. from the words.

Example: jumping, jumped – jump

4. Lemmatization:
In this step the root stem of the words in obtained. Root stem gives new base from of the words.

Example: went, going, gone – go (lemma)

5. Part of speech tagging:
In this step the sentences is converted to different forms like list of words, list of tuples etc. In this step which part of the speech a word belongs to which tag word such as verb, noun, pronoun, etc. is identified.

Example: I am a frontend developer.

I (pronoun), am (verb), a (article), frontend (noun) and developer (noun).

6. Entity Annotation (or Named Entity Recognition/ Identification):
Entity annotation or Named Entity Recognition is the process of classifying the words in sub-categories which is one of the subtasks of information extraction of resumes. The category represents what the meanings of the given words are. It classifies the named entities of unstructured text to predefined categories. The NER (Named Entity Recognition) model is used and trained using annotated words to identify and create required categories.

The sub categories that are available in NER model by default are:

a. Person
b. Quantity
c. Location
d. Organization
e. Movie
f. Monetary Value

Spacy is used for training the NER model.

7. Modelling
The steps involved in resume shortlisting using NLP are as follows:
Firstly, all the data (i.e. resumes) are collected.
The dataset is divided into two subsets: the test set contains the remaining 20% of the data, and the training set includes 80% of the data in its entirety.
The required libraries and modules are loaded.
The training data is extracted and pre-processed.
The training data is annotated using the automatic annotation tool known as "Prodigy".
The Spacy configuration file is downloaded and edited to configure the training environment.

Then the annotated data is fed into the NER model using "Spacy" tool to train the model (Refer figure 2).

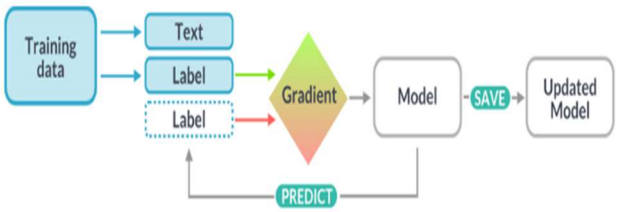

Figure 2. NER model using “Spacy” tool

The Spacy tool's operation is dependent on numerical models; each of the decisions made by its constituent parts is a forecast made using the model's current weight values. In order to estimate the gradient of the loss, the model's predictions are compared against reference descriptions during the training process, which consists of iterations. The gradient of the weights is then calculated using this loss gradient in back propagation, enabling the model to continuously improve its predictions through iterative learning. The architecture of training model of spacy is as follows:

8. The accuracy of the NER model is checked using three parameters.

The parameters are:

a. P (Precision)
By determining the percentage of positive predictions that are sincerely accurate, precision evaluates the model's accuracy. Mathematically,
Precision (P) = Number of Correct Predictions / Total Number of Predictions

b. R (Recall):
In the framework of classification, recall evaluates the percentage of actual positive experiences in the dataset that the classifier accurately identified as positive instances. Mathematically,
Recall (R) = Total Number of True Positives/ Total Number of Actual Positives in the Dataset

c. F (F1 score):
Often thought of as the balance of recall and precision, the F1-Score presents a single metric that takes both into consideration when assessing performance. Mathematically,
F (F1 score) = 2 * (precision * recall)/ (precision + recall)
F1 score gives the accuracy of the model. If F1 score of the model is less, then the model is trained again with more relevant dataset and valid entities in NER.

9. The model is tested against the test data.

10. The training of the model is completed with good F1 score.

11. A test job description is created.

12. A resume data is selected from test dataset and passed through the trained NER model to get annotated data.

13. Then a cosine similarity is used to measure the similarity between the resume data and the job description and similarity score is generated.

Formula used by cosine similarity is as follows:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Here, A – Resume data and B – Job description.

14. For resume short listing, each resume is passed through the created model and each resume gets a similarity score for the given job description.

15. The resumes are sorted based on the similarity score.

16. The resumes can be shortlisted based on the required number of candidates (Refer figure 3).

IV. Results:

```
================================== Results ==================================

TOK     100.00
NER P   96.49
NER R   76.39
NER F   85.27
SPEED   3420


============================== NER (per type) ==============================

                       P       R       F
NAME                100.00  100.00  100.00
CONTACT NUMBER      100.00  100.00  100.00
EMAIL ADDRESS        92.31  100.00   96.00
DESIGNATION         100.00  100.00  100.00
COMPANIES WORKED AT 100.00  100.00  100.00
SKILLS               92.63  100.00   96.17
GRADUATION YEAR     100.00  100.00  100.00
INSTITUTE           100.00  100.00  100.00
DEGREE              100.00  100.00  100.00
LOCATION            100.00   88.89   94.12
```

Figure 3. Result of NER model using “Spacy” tool

V. CONCLUSION

Natural Language processing (NLP) based resume shortlisting research has shown great potential for accelerating the hiring process. It is obvious that NLP combined with machine learning can be successful in early resume screening based on insights gathered from various relevant studies. These methods offer a productive approach to organize candidates, extract useful information from their resumes, and cut down on the time and expenses related to manual evaluations. The precision and dependability of these technologies have the potential to change the work that organizations conduct, as prior study has demonstrated.

VI. FUTURE WORK

Improvement in the field of resume shortlisting using NLP has a wide range of potential directions. By creating the database and testing various types and models in various scenarios, it will be possible to expand the application of these methods beyond engineering to various industries and fields of practice, improving the flexibility and robustness of these systems. Additionally, methods to develop dynamic profiles of users based on data retrieved from their resumes, as suggested in one research, may provide job searchers with useful insights. Finally, as technology advances, it is crucial to create effective algorithms for training NLP models that will eventually result in outcomes that are well-matched.

The process of resume shortlisting using NLP may be made even better by adding the insights and goals from previous research, giving both employers and job seekers a productivity face they can rely on.

REFERENCES

[1] B. Surendiran, T. Paturu, H. V. Chirumamilla and M. N. R. Reddy, "Resume Classification Using ML Techniques," 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), Karaikal, India, 2023, pp. 1-5. doi: 10.1109/IConSCEPT57958.2023.10169907.

[2] R. Bathija, V. Bajaj, C. Megnani, J. Sawara and S. Mirchandani, "Revolutionizing Recruitment: A Comparative Study Of KNN, Weighted KNN, and SVM - KNN for Resume Screening," 2023 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2023, pp. 834-840. doi: 10.1109/ICCES57224.2023.10192665.

[3] V. S. Pendyala, N. Atrey, T. Aggarwal and S. Goyal, "Enhanced Algorithmic Job Matching based on a Comprehensive Candidate Profile using NLP and Machine Learning," 2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService), Newark, CA, USA, 2022, pp. 183-184. doi: 10.1109/BigDataService55688.2022.00040

[4] 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

[5] T. M. Harsha, G. S. Moukthika, D. S. Sai, M. N. R. Pravallika, S. Anamalamudi and M. Enduri, "Automated Resume Screener using Natural Language Processing(NLP)," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 1772-1777. doi: 10.1109/ICOEI53556.2022.9777194.

[6] 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

[7] M. F. Mridha, R. Basri, M. M. Monowar and M. A. Hamid, "A Machine Learning Approach for Screening Individual's Job Profile Using Convolutional Neural Network," 2021 International Conference on Science & Contemporary Technologies (ICSCT), Dhaka, Bangladesh, 2021, pp. 1-6.