Job Retrieval Model
Matching Resumes with Job Openings

Anay Chauhan
BTech (CSE) 2021013
anay21013@iiitd.ac.in

Divyansh Mishra
BTech (CSE) 2021040
divyansh21040@iiitd.ac.in

Pranjal Bharti
BTech (CSE) 2021080
pranjal21080@iiitd.ac.in

Dhruv Sood
BTech (CSSS) 2021387
dhruv21387@iiitd.ac.in

Hardik Singh
BTech (CSSS) 2021390
hardik21390@iiitd.ac.in

Priyash Shah
BTech (CSAI) 2021553
priyash21553@iiitd.ac.in

Abstract

In today's dynamic job market, traditional methods of job searching often prove inadequate in efficiently connecting job seekers with relevant opportunities, primarily due to the limitations of keyword-based searches and manual screening processes. This inefficiency leads to mismatches and underutilization of skills, impacting both job seekers and employers. To address this critical issue, our proposed solution harnesses a combination of information retrieval techniques and advanced Natural Language Processing (NLP) methodologies.

Our approach involves parsing resumes to extract essential information such as skills and experience, utilizing job platform APIs to retrieve comprehensive job listings and requirements, and employing a sophisticated matching algorithm to compare resume profiles with job descriptions. By incorporating algorithms like TF-IDF with Random Forest for probabilistic results, and integrating advanced techniques such as BERT embeddings for contextual understanding, we aim to significantly enhance job matching accuracy and streamline the overall job search process.

Moreover, we introduce a unique feature to our platform: the capability to optimize resumes using Generative AI API. This functionality empowers users to upload their resumes and receive personalized recommendations for optimization, ensuring that resumes align closely with the requirements of specific job postings.

Through rigorous experimentation and a commitment to continuous improvement, our platform endeavors to optimize resume-job posting matching, thereby increasing user satisfaction and contributing to economic growth by reducing unemployment rates and fostering skill development. With a focus on user-friendly web interfaces and a dedication to expanding our global impact, our platform aims to revolutionize the job market by facilitating better job matches and promoting diversity and inclusion in the workforce.

1 Problem Statement

In today's job market, nearly 60 percent of job seekers encounter difficulties in finding relevant job openings that align with their skills and preferences. this challenge stems from the inherent limitations of traditional job search methods, particularly on online job boards like indeed, where individuals primarily rely on keyword-based searches. while job seekers often input specific job titles they desire, such as "software developer" or "senior project manager," the narrow focus on these titles overlooks numerous other positions that may be suitable based on their skill-sets and experiences. consequently, job seekers may miss out on potential opportunities, leading to frustration, extended job search durations, and mismatches between candidates and positions.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

This problem is further exacerbated by the variability in job titles across companies and industries. for instance, a software developer may encounter job listings with titles like "CS programmer" or "Front-end Engineer," while a senior project manager might overlook positions titled "Product manager" or "Program manager," despite possessing transferable skills. as a result, job seekers must expend additional time generating alternative title variations or seeking roles emphasizing transferable skills, adding complexity and inefficiency to the job search process.

Addressing this problem is crucial for several reasons. firstly, it addresses a significant inefficiency in the job market, where mismatches between job seekers and opportunities hinder economic growth and stability. secondly, by leveraging information retrieval (IR) techniques and natural language processing(NLP), we can streamline the job search process, enhance job matching accuracy, and optimize resumes, thereby improving the overall user experience for both job seekers and employers.

Finally, solving this problem benefits various stakeholders, including job seekers, employers, and recruiters, by increasing the likelihood of finding suitable matches, reducing recruitment costs, and improving efficiency. In summary, the problem at hand involves improving job matching accuracy and efficiency by addressing the limitations of keyword-based searches and manual screening processes on job boards. by doing so, we empower job seekers to discover a broader range of relevant job opportunities that align with their skills and preferences, ultimately contributing to economic growth, skills development, and diversity and inclusion in the workforce.

2 Motivation

The problem of inefficient job matching is of paramount importance due to its widespread impact on various stakeholders and its implications for economic growth, workforce development, and diversity and inclusion.

2.1 Improved job seeker experience:

By addressing job matching inefficiencies, we enhance the job search experience, empowering individuals to find roles aligned with their skills and goals. This boosts job satisfaction, work-life balance, and financial outcomes, while also increasing motivation. When individuals apply only to relevant jobs, they feel more confident in their applications and have higher chances of selection. This targeted approach reduces the loss of motivation from rejections, encouraging job seekers to stay engaged in the search and strive for advancement.

2.2 Enhanced productivity and innovation:

Aligning individuals with roles that match their skills and interests not only benefits job seekers but also leads to increased productivity and innovation within organizations. When employees are in roles that capitalize on their strengths and passions, they are more engaged, motivated, and likely to contribute meaningfully to their work.

2.3 Efficiency for recruiters and hr professionals:

By leveraging advanced technologies such as information retrieval (IR) and Natural Language Processing(NLP) techniques, recruitment agencies and HR professionals also stand to benefit from streamlined processes credited to access to a larger pool of qualified candidates.

2.4 Contribution to economic growth:

Addressing the inefficiencies in job matching contributes to economic growth and stability by reducing unemployment rates and facilitating better job matches. when individuals are employed in roles that utilize their skills and talents effectively, they contribute to increased productivity, consumer spending, and overall economic activity.

3 Objective

The primary objectives of this project are as follows:

Resume Parsing To parse resumes to extract relevant information such as skills, experience, and education which will be fed into the model.

Exploratory Data Analysis To gain insights into the distribution and characteristics of Resume Dataset, providing a foundational understanding of the data.

Data Preprocessing It involves cleaning the data to handle missing values, removing duplicates, and addressing inconsistencies. additionally, data preprocessing includes transforming the dataset into a structured format suitable for analysis and model training.

Front-End Testing To develop and test the front-end framework for input and output of information. this involves designing user interfaces that are intuitive and user-friendly, allowing users to easily input their resume information and view the output generated by the system.

Matching algorithm To use a matching algorithm to compare the profiles of resumes and job openings which calculates a similarity score based on factors such as job description, user skills, experience etc. and recommends changes in the resume based on different matched jobs and their relevant scores.

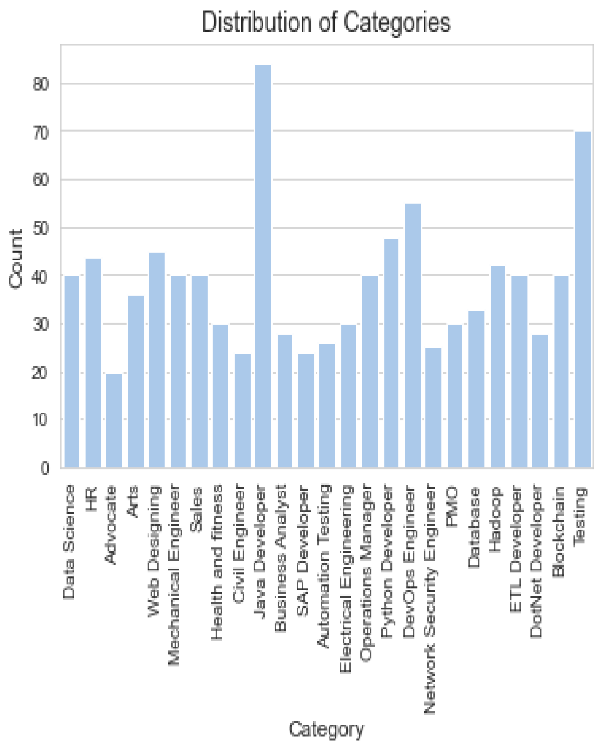
4 Dataset Description

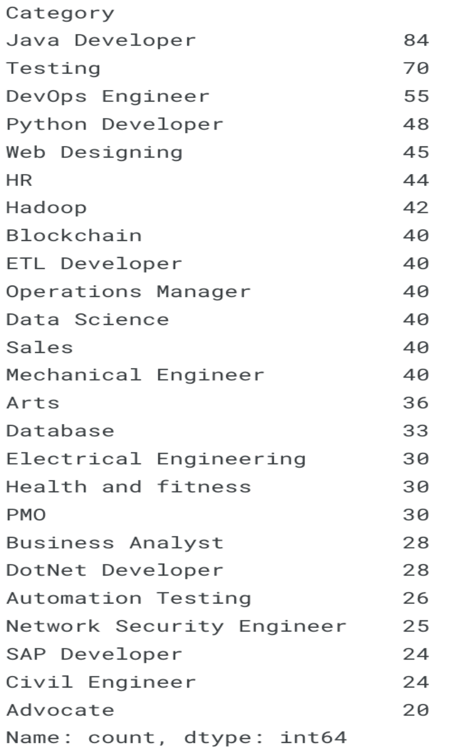
The dataset consists of resumes categorized under various fields where each entry in the dataset represents a resume of an individual with relevant information regarding their skills, experience, education, and professional background etc.

Companies often receive thousands of resumes for each job posting and employ dedicated screening officers to screen qualified candidates.Hiring the right talent is a challenge for all businesses. This challenge is magnified by the high volume of applicants if the business is labour-intensive, growing, and facing high attrition rates.

IT departments are short of growing markets. In a typical service organization, professionals with a variety of technical skills and business domain expertise are hired and assigned to projects to resolve customer issues. This task of selecting the best talent among many others is known as Resume Screening.Typically, large companies do not have enough time to open each CV, so they use machine learning algorithms for the Resume Screening task.

4.1 Exploratory Data Analysis

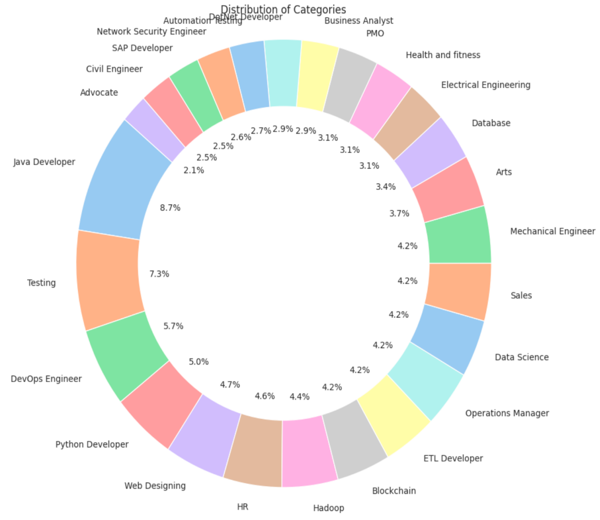




```
Category
Java Developer                84
Testing                       70
DevOps Engineer               55
Python Developer              48
Web Designing                 45
HR                            44
Hadoop                        42
Blockchain                    40
ETL Developer                 40
Operations Manager            40
Data Science                  40
Sales                         40
Mechanical Engineer           40
Arts                          36
Database                      33
Electrical Engineering        30
Health and fitness            30
PMO                           30
Business Analyst              28
DotNet Developer              28
Automation Testing            26
Network Security Engineer     25
SAP Developer                 24
Civil Engineer                24
Advocate                      20
Name: count, dtype: int64
```

The dataset contains about 900+ resume categorised under different Job profiles.Every Category contains vital information for the job such as Skills,Experience,Educational Background, projects etc.

4.2 Data Distribution



Ltd description Details January
monthsCompany Details
Exprience months
Education Details
Pvt Ltd
SQL Server
Details company
Exprience monthsCompany
Exprience Less
Skill Details
Mumbai Maharashtra
Project Description
Pune Maharashtra
year months
Project
Project Name
client
team member
application
knowledge
experience
Machine Learning
Operating System
customer
Company Details
test case
Worked
Management System
data
user
team
Java Developer
Pune University
business requirement
Maharashtra State
description company

- The data encompasses various professional fields or job roles, ranging from technical domains like HR, Java Developer, and Testing, to non-technical domains such as Chef, Fitness, and Arts.
- The distribution of categories is not uniform, as some categories appear more frequently than others. This non-uniform distribution is typical in real-world datasets, reflecting the prevalence or demand for certain professions over others.

- By examining the distribution, one can gain insights into the job market or the composition of a particular workforce. For example, categories like HR, Java Developer, and Testing have relatively high frequencies, suggesting a significant presence or demand in the job market, while categories like BPO and AUTOMOBILE have lower frequencies, indicating a comparatively smaller presence or demand.

5 Literature Review

5.1 End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT

The escalating volume of job applications poses a manual challenge for employers in identifying suitable candidates. an end-to-end solution is presented to rank candidates based on their suitability to a job description. this involves two primary stages. initially, a resume parser is constructed to extract comprehensive information from candidate resumes. this parser is subsequently deployed as a web application for public use. following this, Bert sentence pair classification is employed to rank candidates according to their alignment with the job description. to approximate the job requirements, the descriptions of candidates' past job experiences as outlined in their resumes are utilized. the dataset encompasses resumes in both linkedin and general non-linkedin formats. perfect parsing accuracy is achieved for linkedin resumes, establishing a robust baseline with 73 percent accuracy for assessing candidate suitability. [1]

5.1.1 Method

Resume parsing: Differentiate between linkedin and non-linkedin resumes using heuristic methods based on font size frequency and order of occurrence. Structure extracted text into predefined segments using heuristics and Bert-based sequence classification.
Standard format conversion: Attempt to convert non-linkedin resumes into linkedin format using bert for sequence classification. Segment text based on heuristics and classify it into linkedin format sections.
Candidate ranking: Utilize Bert for sequence pair classification to rank candidates based on their suitability to a job description.simulate job descriptions using a candidate's past job experiences and train bert to predict the similarity between a job description and a candidate's work experience.

5.1.2 Results

- 100% accuracy in differentiating between linkedin and non-linkedin resumes.
- 100% accuracy in structuring extracted text from linkedin resumes into predefined segments.
- 97% accuracy in converting non-linkedin resumes to linkedin format.
- 72.77% accuracy in predicting the similarity between job descriptions and candidates' work experiences using Bert-based sequence pair classification.

5.1.3 Limitations

Format dependency: the system heavily relies on structured linkedin resumes, potentially leading to information loss and reduced accuracy when parsing resumes in varied formats.
Limited generalisation: focusing primarily on linkedin format resumes and simulating job descriptions from candidate work experiences may limit the system's ability to generalise across diverse resume styles and job requirements.

5.2 Resume Classification using various Machine Learning Algorithms

The shift to online platforms due to the pandemic has underscored the importance of automating processes like hiring to improve efficiency and reduce manual effort that can be digitized. automating the initial stage of resume categorization and verification would significantly streamline the recruitment process, saving time and mitigating human errors associated with manual handling of paperwork.

Machine learning algorithms, including naive bayes, random forest, and support vector machines (svm), can be leveraged for resume classification. these algorithms facilitate the extraction of skills and demonstrate diverse capabilities within appropriate job profile categories. as resumes are categorized and pre-processed, the system can retrieve suitable job profiles and present them to interviewers, expediting applicant selection.

This automated process proves particularly beneficial during video interviews, where interviewers can readily access relevant candidate profiles. by providing interviewers with real-time insights into candidate skills and qualifications, this system enhances decision-making and accelerates the selection process. [4]

5.2.1 Method

Data collection and preprocessing: gather datasets from sources like kaggle, glassdoor, and indeed. clean the unstructured data by removing spaces, converting text to lowercase, and eliminating stop words. tokenize documents, and apply stemming and lemmatization to standardise vocabulary and simplify word variations.
POS tagging and tf-idf vectorization: associate grammatical information with words based on context and relationships within sentences. simultaneously calculate term frequency-inverse document frequency (tf-idf) to assess word importance. this assigns weights to words based on their frequency and rarity in the dataset, aiding in understanding the text's syntactic structure.
Applying classification algorithm: utilise classification algorithms such as naïve bayes, support vector machine (SVM), and random forest to train the model. train the models with cleaned and classified data, evaluating performance metrics including accuracy, precision, recall, and f1 score.
Confusion matrix analysis: generate confusion matrices for each classification algorithm to evaluate true positive, true negative, false positive, and false negative values. random forest demonstrates excellent true positive values across various job classes.

5.2.2 Results

- Naïve Bayes: accuracy - 45%, precision - 0.521, recall - 0.452, f1 score - 0.448.
- SVM: accuracy - 60%, precision - 0.598, recall - 0.597, f1 score - 0.594.
- Random forest: accuracy - 70%, precision - 0.687, recall - 0.683, f1 score - 0.678.

Classification algorithms:

- Naïve Bayes: accuracy - 45%
- Support Vector Machine (SVM): accuracy - 60%
- Random Forest: accuracy - 70%
  Random Forest exhibited the best performance with high true positive values across various job classes.

5.2.3 Limitations

Data bias: Biases in the training data, such as under representation of certain demographics or job profiles, may lead to skewed results and inaccurate classifications.
Algorithm selection: Although Naïve Bayes, SVM, and Random Forest are frequently employed for resume classification, other algorithms or ensemble methods may yield superior performance, yet were not investigated in this study. While these conventional algorithms demonstrate satisfactory performance, they may not fully capture the semantic nuances of the text.

6 Novelty

This solution stands out for its comprehensive approach to resume optimization and job suggestions, uniquely tailored for the Indian job market. Unlike existing paid services like jobscan.co, which often offer limited features, clunky interfaces, and lack optimization for the nuances of the Indian context, our platform aims to fill these gaps and provide a seamless, user-friendly experience.

Integration Across Platforms: Unlike single-platform solutions, our platform integrates data from various job portals, recruitment websites, and professional networks. By aggregating job listings from multiple sources, users gain access to a diverse array of opportunities across different industries and sectors.

Dynamic Job Recommendations: A key innovation of our platform is its ability to dynamically update job suggestions in response to changes in a user's resume and evolving job market trends. By continuously monitoring job postings and analyzing user profiles, the platform delivers personalized job recommendations aligned with each user's unique career goals and aspirations.

Tailored for the Indian Market: Acknowledging the distinct features of the Indian job market, our platform is specifically crafted to meet the needs and preferences of Indian job seekers and employers. From language and cultural considerations to industry-specific requirements, our platform takes into account factors relevant to the Indian context, ensuring users receive pertinent and actionable insights.

7 Methodology

7.1 Model Overview

The job retriever model aims to match resumes with job openings by leveraging parsing techniques and job platform APIs. Two distinct approaches are explored: one utilizing TF-IDF encoding coupled with a random forest algorithm for probabilistic results, and the other employing the BERT model to capture the semantic nuances of words.

7.2 Resume Parsing

Resumes are parsed to extract pertinent information such as skills, experience, and education. This extracted data serves as input for the model.

7.3 Job Platform APIs

The model interacts with job platform APIs to retrieve job openings and their corresponding requirements. It collects data on job titles, skills, and other criteria.

7.4 Matching Algorithm

The matching algorithm employs BERT (Bidirectional Encoder Representations from Transformers) to compare resume and job opening profiles. It calculates a similarity score based on factors such as job descriptions, user skills, experience, and other relevant attributes.

7.5 Resume Optimization

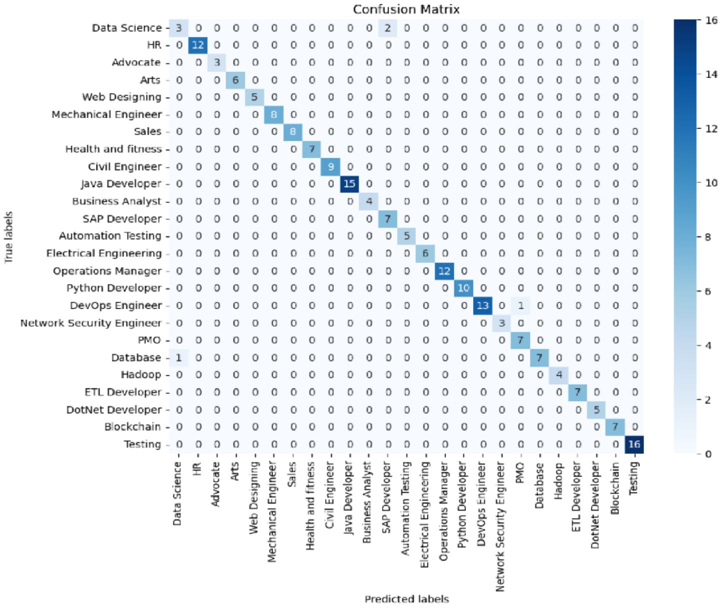
Additionally, the model leverages Generative AI API (Gemini AI) for resume optimization. This component enables users to upload their resumes and receive personalized recommendations for optimizing them based on semantic analysis, ensuring alignment with job postings and enhancing the overall job matching process.

7.6 Baseline

7.6.1 Base Model

TF-IDF vectors were utilized to encode textual values into numerical representations. Additionally, label encoding was performed on all categories, assigning them numerical values from 1 to 25. The dataset was then split into training and testing sets in an 80:20 ratio.

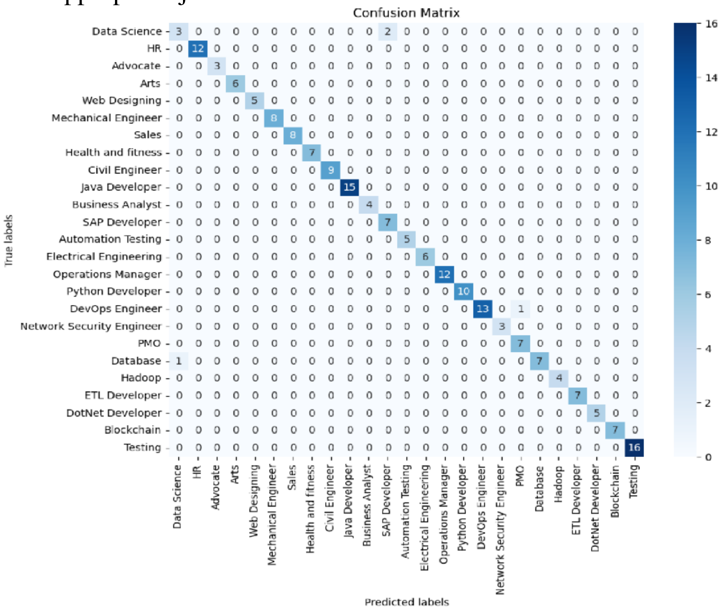
- **KNN:** Initially we used a K Nearest Neighbours algorithm based model in order to make the predictions of the job titles suitable for the user.



TF-IDF Scores were used in order to encode the textual values to numbers. We used the following evaluation metrics in order to test the accuracy of the model and the results were as follows:

- Precision : 0.98
- F1 score : 0.98
- Recall : 0.98

- **Random Forest:** Later, we changed the model being used to make the predictions from KNN to a Random Forest based machine learning model. We used Random Forest to classify the textual values from the resume and get probabilistic matching. Used this probability to rank appropriate jobs.



We used the same evaluation metrics as before and made the comparison between these two models based on them:

- Precision : 0.99
- F1 score : 0.98
- Recall : 0.98

7.6.2 Website Development:

We created a web application that is built using Flask, a Python web framework, which provides routing and rendering capabilities. The frontend part of the website makes use of the ReactJS framework which allowed us to create a much more interactive, user friendly and intuitive website. The frontend consists of mainly 5 components:

- Homepage
- Anout page
- File Upload page
- Optimisation page
- Job similarity page

We simply made use of /GET and /POST methods to fetch and display the relevant data. Moreover, we used Tailwind CSS to beautify our website and make it more appealing.
The backend part of the website is made using Flask (as mentioned before) and helps us to integrate the model that we created with the frontend allowing us to make predictions and finally making use of APIs to fetch the relevant job openings.

8 CodeBase

Following is the link to the codebase Job Retrieval.

9 Evaluation

9.0.1 Comparison with Baselines

Our final model, powered by BERT, revolutionizes the job matching process by leveraging its advanced NLP capabilities. BERT's contextual understanding of language enables it to precisely classify resumes into sections, ensuring accurate job matching. By encoding the semantic content of both resumes and job descriptions, our model captures the intricate relationships between words and phrases.

Unlike traditional methods such as TF-IDF with Random Forest, which rely on keyword-based matching, BERT embeddings provide a more nuanced representation of text. BERT considers the contextual meaning of words, allowing for a deeper understanding of language semantics. This contextual understanding enables our model to identify the most suitable job matches based on skills and qualifications with greater accuracy. Additionally, BERT-based matching offers improved performance compared to TF-IDF with Random Forest due to its ability to capture semantic similarities between resumes and job descriptions. By considering the semantic context captured by BERT embeddings, our algorithm ensures that job matches are not solely based on keyword matches but also on the broader meaning and context of the text. This results in more relevant and precise job matches, ultimately enhancing the effectiveness of the job matching process.

9.0.2 Performance on existing data/SOTA evaluation metrics

The evaluation was done on the outcomes of the final model that was powered by BERT (Bidirectional Encoder Representaions from Transformers). Using the BERT model allows us to classify resumes into sections in a better manner as compared to the previous models. Moreover, unlike the Random Forest and KNN based models which make use of keyword based classification, the BERT model uses contextual based classification, which leads to better results. Not only did this model give a better accuracy as compared to the previous models, it had a better performance with respect to other metrics as well. Once again we used the usual evaluation metrics of Precision, F1 Score and Recall along with a few additional metrics that are mentioned below:

- Precision : 1
- F1 score : 1
  (F1 Score is one of the most valuable SOTA evaluation metric. Since this model has a F1 Score of 1, it is safe to assume that this model gives the best output)

- Recall : 1

9.0.3 Performance on New Dataset

Evaluating the performance of a model on new data is crucial for assessing its generalization capabilities and real-world applicability. In our case, since the model is designed specifically for technical job matching, finding a similar new dataset might be challenging.

Since finding a similar new dataset was not feasible due to the specialized nature of the model, we couldn't directly evaluate its performance on entirely new data. However, by rigorously evaluating its performance on the test data from the existing dataset, we can still gain valuable insights into its effectiveness and robustness.

[3] [6] [5] [2]

References

[1] Vedant Bhatia et al. “End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT”. In: *ArXiv* abs/1910.03089 (2019). URL: https://api.semanticscholar.org/CorpusID:203905435.

[2] Jie Chen, Chunxia Zhang, and Zhendong Niu. “A Two-Step Resume Information Extraction Algorithm”. In: *Mathematical Problems in Engineering* 2018 (May 2018), pp. 1–8. DOI: 10.1155/2018/5761287.

[3] Yiou Lin et al. “Machine learned resume-job matching solution”. In: *arXiv preprint arXiv:1607.07657* (2016).

[4] Pal, Riya et al. “Resume Classification using various Machine Learning Algorithms”. In: *ITM Web Conf.* 44 (2022), p. 03011. DOI: 10.1051/itmconf/20224403011. URL: https://doi.org/10.1051/itmconf/20224403011.

[5] Rui Yan et al. “Interview Choice Reveals Your Preference on the Market: To Improve Job-Resume Matching through Profiling Memories”. In: July 2019, pp. 914–922. ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330963.

[6] Peng Yi et al. “A job recommendation method optimized by position descriptions and resume information”. In: Oct. 2016, pp. 761–764. DOI: 10.1109/IMCEC.2016.7867312.