JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

RESUME ANALYZER USING TRANSFORMERS

Prof. Savita Sawant[1], Mrunal Kulkarni[2] , Pranjali Polawar[3], Maheshwari Subramaniyan[4], Avanti Thale[5]

[1]Department of Computer Engineering, Pillai College of Engineering, New Panvel, Navi Mumbai, India
[2]Department of Computer Engineering, Pillai College of Engineering, New Panvel, Navi Mumbai, India

[3]Department of Computer Engineering, Pillai College of Engineering, New Panvel, Navi Mumbai, India

[4]Department of Computer Engineering, Pillai College of Engineering, New Panvel, Navi Mumbai, India

[5]Department of Computer Engineering, Pillai College of Engineering, New Panvel, Navi Mumbai, India

Abstract
Over the past five years, there has been a noticeable increase in the Indian recruitment market owing to the escalating demand for cost-effective labor and an uptick in job opportunities. Concurrently, with the growing job market, there's been a corresponding rise in the recruitment industry, adopting a novel approach to hiring by outsourcing the recruitment process to specialized firms dedicated to sourcing the right talent for businesses. As these entities come to the fore, manually sifting through all candidate resumes becomes both time-consuming and challenging, prompting the need for automated resume analysis systems. These systems utilize natural language processing techniques to extract relevant information such as qualifications, skills, education, and experience, subsequently matching them with company requirements to identify optimal candidates for positions. Additionally, the website utilizes APIs to scour the web for current job listings, aligning user resumes with job descriptions. This developed platform offers advantages to both job seekers and HR professionals. Additionally, the website utilizes APIs to scour the web for current job listings, aligning user resumes with job descriptions. This developed platform offers advantages to both job seekers and HR professionals.

Keywords

Resume, Transformers, Natural Language Processing, Indian Recruitment Market, Automated Resume Analysis, Job opportunities

1. INTRODUCTION

Finding a job in today's scenario is a strenuous task. Analyzing a resume manually is a complex work for HR employees as they have to go through hundreds of resumes at a time. Automatic resume analyzer using machine learning techniques helps to parse the resumes efficiently where NLP is used to tokenize the text and perform semantic and lexical analysis on the data. The Resume Analyzer is an automated system that extracts relevant information from a job seeker's resume and provides feedback based on the job requirements. It analyzes the resume's content and identifies the skills, education, experience, and other important information required for a job. This project is developed to help recruiters automate the process of filtering and screening candidates.

A. Fundamentals

In the context of the current job market, the need for efficient resume analysis and classification has become increasingly evident. This project leverages advanced Natural Language Processing (NLP) techniques,including lexical analysis, semantic analysis, and syntax analysis, to construct resume parsers. After a thorough analysis of various techniques, BERT has been selected as a fundamental component of the project. The proposed system involves the conversion of resumes in various formats into raw text, followed by the application of NER using BERT and keyword extraction.The requirements for a role are matched with the skills extracted from the resume and the percentage of matching is found. These extracted keywords play a crucial role in matching candidates' qualifications with the specific requirements specified by companies, ultimately streamlining the hiring process and improving the efficiency of candidate selection.

B. Objectives

Our system aims to elevate the existing resume ranking system to enhance flexibility for both parties involved.

1) Candidates, who want to find job
2) Client company, who is hiring the candidates.

1.2.1 *Candidates, Who Want to Find Job :*

Unemployment is a huge problem in our country, with a large number of people getting graduated but without any job. The main reason behind this is that people are unable to find a job they desire. Nowadays, candidates even with low skillset want a high standard job which is difficult to find. This is simply because their resumes do not have the required skills that the company requires. Our system will help such candidates to increase the chances of getting hired by helping them compare their resume with the job requirements of their dream job.

1.2.2 *Client Company, Who is Hiring the Candidates*:

On the other hand a lot of difficulties are faced by the HR’s of the companies while finding suitable candidates for their organization. As they are searching for the best among the thousands of candidates. They have to go through many resumes and then shortlist the candidates for the interviews. This job itself is very tiring. Sometimes it may be possible that they miss a very good candidate due to human error. Here our system will help them by providing a list of suitable candidates to select from thus saving their time and efforts.

C. Organization of the Report

The structure of the report is as follows: Chapter 1 serves as the introduction, providing an overview of fundamental terms utilized in the project. It also serves to motivate the exploration and comprehension of various techniques, including Natural Language Processing, Machine Learning, and the Implementation of Resume Analyzer using Transformers. Additionally, this chapter outlines the objectives of the report. Chapter 2 conducts a review of pertinent techniques found in the literature, discussing their advantages and disadvantages. Chapter 3 outlines the theoretical framework and proposed methodology, detailing the primary approaches employed in the project.

I. LITERATURE SURVEY

In past decades a lot of research has been performed on resume analysis. Various NLP models have been developed and implemented by various researchers using different NLP algorithms, text mining techniques etc.
The papers are technically classified as follows :

1. *Systems using BERT-BiLSTM-CRF :*

XiaoWei Li, Hui Shu, et al.[1]proposed an approach for extracting resume information using named entity recognition (NER). Their method focuses on extracting personal details such as educational background, job aspirations, and skills from resumes using NER techniques. The approach utilizes the BERT language model, incorporating a multi-head self-attention mechanism to extract text features and generate word-level vector matrices. Additionally, they employ a Bidirectional Long Short-Term Memory neural network to capture contextual abstraction features from sequential text. Finally, the Conditional Random Field (CRF) algorithm is applied for decoding and annotating the globally optimal sequence, aiding in the extraction of relevant resume entity information.

2. *Systems using Text Mining :*

Yi-Chi Chou, Chun-Yen Chao, et al.[2] investigated the utilization of interview robots in the recruitment domain. Employing techniques such as web crawling, text mining, and natural language processing, the study developed a robust system for matching job candidates with recruiters. This system analyzed electronic resumes in Traditional Chinese, assessing words based on their relevance to the Internet job market and implementing big data-related techniques. Results indicated that the system effectively identified talent-seeking demands and swiftly provided candidate rankings for specific positions, catering to the requirements of both job seekers and recruiters.

On the other hand, Abhishek Singh, Shivang Pal, Shree Prakash Singh [5] proposed a system incorporating aggressive score, DISC personality, and ability analysis for job vacancy recommendation, generating a tailored talent recommendation list for companies. This system offers personalized analysis, considering the individual preferences of job seekers. However, due to its personalized nature, the system entails a more intricate and time-consuming process.

3. *Systems using Optical Character Recognition :*

Shubham Bhor, Vivek Gupta et. al. [3] proposed OpticalCharacter Recognition(OCR) to extract the data from Resume and Extract the data from Social Media like Linkedin. The proposed system gets quality applications and avoids unfair and non- discriminatory practices. This system did not take into account the privacy of the candidates whose data was accessed.

4. *Systems using classification*

Md. Tanzim Reza and Md. Sakib Zaman[6] proposed a model aimed at extracting essential information from semi-structured text formats found in resumes or curriculum vitae (CVs), and subsequently ranking it based on the preferences and requirements of associated companies. The model is structured into three main segments to accomplish this objective.

Firstly, the entire CV or resume is segmented based on the topics of each section. Secondly, data is extracted in a structured format from the initially unstructured data. Finally, the structured data is evaluated using a decision tree algorithm, and the system is trained accordingly.

The process of structured data extraction involves segmenting the CV or resume by converting it into HTML format. Subsequently, decision tree algorithm techniques are employed to classify the input into various categories, facilitating the extraction of structured information.

5. *Systems using Text analytics*

Divyanshu Chandola, Aditya Garg, and Ankit Maurya[7] introduced a Text Analytic approach for assessing resumes based on their content. They proposed leveraging Sentiment Analysis techniques to evaluate a candidate's resume by analyzing the sentiment expressed in the provided descriptions. Sentiment Analysis has proven valuable in diverse scenarios, including capturing people's responses towards services and products.

Sayed Zainul Abideen Mohd Sadiq et al.[4] introduced a method for ranking resumes using Natural Language Processing (NLP) and Machine Learning techniques. The proposed system comprises two modules, each serving specific functions within the ranking process. One being the outer world system and the other being the resume ranking system. Infographics Resume is generated with all the stats and the queries can be made to visualize the data.The system provided ranking according to company constraints and also used social profiles to get genuine information.

6. Systems using Topic modeling and K nearest Neighbors

Suhas Tangadle and Vijayaraghavan Varadharajan[8] proposed an automated tool for resume classification utilizing semantic analysis. The authors describe the development and deployment of a resume classifier application utilizing ensemble learning through a voting classifier. This application sorts candidates' profiles into relevant domains according to their expressed interests, work history, and expertise as indicated in their resumes.

The application utilizes topic modeling techniques to introduce new domains to the existing list. In the unsupervised machine learning process, it scans a set of documents, detects word and phrase patterns, and automatically clusters them into domains best suited for the set of documents.The resume classifier employs Natural Language Processing (NLP) and a classification module. Initially, unnecessary information is eliminated from the data, and tokens are formed. The classification module then analyzes these tokens and assigns them to suitable domains. Subsequently, a graph is plotted based on the candidates' work experience mentioned in their resumes, and domains are provided accordingly (e.g., HR, TEAM, ORGAN). Finally, a classifier, such as k-nearest neighbors, is added to the ensemble-based voting classifier.

7. *System using TF-IDF Vectorizer*

Nathan Green, Michelle Liu, and Diane Murphy[9] introduced E-RAP (Electronic Resume Analyzer Portal) as a solution to enhance the employability of college graduates. Their paper presents the utilization of Machine Learning (ML) and Natural Language Processing (NLP) to develop a resume analysis and reporting tool. This tool, known as E-RAP, allows students to submit their resumes and receive automatic feedback and a rating report.
The authors outline the E-RAP analysis process, which includes data curation, data cleaning, and various analysis techniques. Researchers have devoted significant efforts to employing different techniques to analyze datasets and develop models for resume analysis. These efforts have led to the effective parsing of resumes using NLP, thereby reducing the workload of HR professionals.

However, the authors note that there is room for improvement in the accuracy of the models and in correctly recognizing entities in resumes and matching them to job requirements. This indicates ongoing research and development efforts to enhance the performance and effectiveness of resume analysis tools like E-RAP.

II. *Resume analyzer*

A. *Existing system Overview*

For a long period of time, the system existing for processing the resumes was manual, which consisted of HRs reading each resume one by one and finding the perfect candidate.However, this method is strenuous and time consuming, and can be biased or inaccurate.To solve this problem, various solutions were developed, which consist of automatic resume parsing using different methods such as using Bi-LSTM, DISC properties, etc. Lets see different approaches that are currently used.

3.1.1 A Resume Evaluation System Based On Text Mining

This system performs four major stages for resume analysis and output as shown in the figure given below.

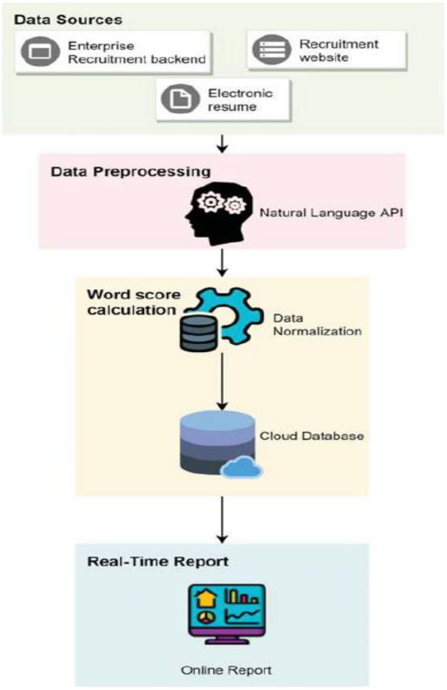


Figure 3.1: Text Mining Based System Overview

A. *Data sources and collection*

1) The DISC phrases were obtained from resumes received by the recruitment backend of a technology company in the system designed for this purpose.
2)Phrases associated with the three competency dimensions were gathered using a Python-based web crawler to extract job descriptions and requirements from major recruitment websites in Taiwan. The collected data mainly focused on recruitment information for various computer engineering positions such as network administrators, system engineers, data analysts, and database administrators.

B. *Data preprocessing:*

1) Text related to DISC and competency dimensions in Chinese underwent segmentation and stop-words filtering using the Jieba system, which is known for its ability to handle large volumes of text data effectively. Additionally, irrelevant words and phrases were removed from the Chinese text to optimize storage space and processing efficiency.

2) Electronic resumes were processed using the python-docx and pdfminer packages to extract text, followed by natural language processing. The extracted phrases were then compared with those in the database.

C. *Phrase Scoring:*

The frequency of occurrence of processed phrases in the job market was calculated, retaining only those with a count of 20 or more that were relevant to the requirements of computer engineering positions. These phrases were categorized into education and work experience, skills, and personality traits (competency dimensions). Normalization was applied to these dimensions using Formula (1) to standardize the data for convergence.

$$1+\frac{(\square-\square\square\square(\square)).9}{\square\square\square(\square)-\square\square\square(\square)} \qquad (1)$$

D. *Real-time reports*

A backend architecture was developed (Fig 3.2), utilizing a MySQL database hosted on a cloud server. Concatenation was performed using PHP. Job applicants submitted resumes to the system, where a Python-based algorithm quantified each index. Big data techniques were then employed to process the quantified user data, resulting in visualized reports on the frontend interface. This enabled candidates to identify their attributes and recruiters to assess the attributes and potential of the candidates. Consequently, the matching level between positions and suitable candidates could be improved.

3.1.2 *Resume Parsing And Finding Candidates for a Job Description using BRET*

The process of Resume Parsing and Candidate Evaluation for job descriptions using BRET (Bidirectional Encoder Representations from Transformers) involves two main stages, as outlined by Bhatia et al. (2019). Firstly, a resume parser is developed to extract essential information from candidate resumes. Subsequently, a ranking is conducted using BERT phrase pair classification, wherein the BERT algorithm predicts the correlation between the job description and candidate profiles with an accuracy of 72.77%.

The research explores the challenge of building a universal parser for all resume types, concluding that it often results in the loss of information and the unfair rejection of certain candidates. Instead, the approach of scanning LinkedIn-style resumes without information loss is adopted. Future investigations aim to incorporate vision-based site segmentation techniques to enhance structural understanding of resumes.

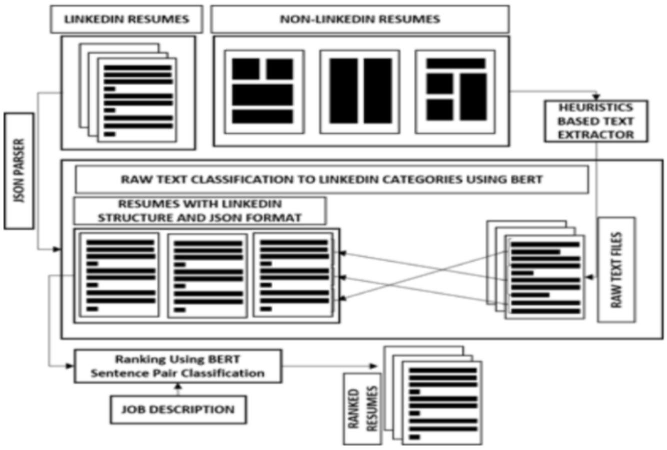

LINKEDIN RESUMES
NON-LINKEDIN RESUMES
HEURISTICS BASED TEXT EXTRACTOR
JSON PARSER
RAW TEXT CLASSIFICATION TO LINKEDIN CATEGORIES USING BERT
RESUMES WITH LINKEDIN STRUCTURE AND JSON FORMAT
RAW TEXT FILES
Ranking Using BERT Sentence Pair Classification
JOB DESCRIPTION
RANKED RESUMES

Figure 3.2 Resume Parsing using BERT System Diagram

3.1.3. *Automatic Extraction of Usable Information from Unstructured Resumes to Aid Search*

In this method, the process involves two passes: in the first pass, the resume is segmented into labeled blocks representing broad data categories. In the second pass, detailed information is extracted. Various heuristics and pattern matching algorithms are employed for extraction, yielding formats with 91% accuracy and 88% recall based on experimental results from multiple resumes.

The implementation of algorithms such as the Bert Model, Neural Network, and NLP has proven effective for information extraction, as observed from various studies and research papers. The system diagram below illustrates the data flow and completed tasks within this process.

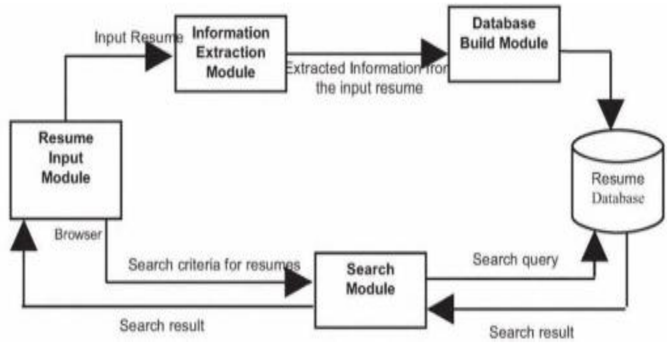


Figure 3.3 Automatic Extraction Flow diagram

3.1.4. *Information Extraction from Resume Documents in PDF Format*

Chen et al. (2016) examine the extraction of data from PDF resumes and propose a hierarchical extraction method. Their approach involves segmenting pages into blocks using heuristic criteria, assigning categories to each block using a Conditional Random Field (CRF) model, and treating detailed information extraction as a sequence labeling task. The authors highlight the importance of layout-based features, which have shown significant effectiveness in addressing challenges related to semi-structured information extraction, resulting in a notable improvement in average F1 Score of over 20% in testing. Unlike HTML resumes, PDF resumes generally contain more detailed information. The authors express their intention to investigate various page segmentation techniques in future experiments to further their understanding of document layout and content.

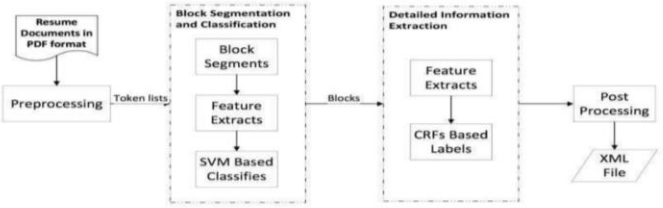


Figure 3.4 Resume Extraction From PDF format system Flow diagram

1) *Proposed System overview*

Screening hundreds of resumes daily is a strenuous task for HR employees, and finding the right job based on their resumes is a difficult job for the candidates. The HR employees may commit an error in analyzing the resumes and finding the right person to hire.

To resolve this problem, the proposed system is a resume analyzer, which is a website that screens the resumes of candidates and finds the best fit candidates for the job according to the job description provided by the company.

Proposed system uses natural language processing to perform the task of screening resumes, i.e the tasks such as lexical analysis, semantic analysis and syntax analysis are carried out by the NLP models such as NLTK. Preprocessing steps such as tokenization, stopword removal and cleaning such as hashtag removal, link removal etc. followed by Name entity Recognition are used.

Transformers like BERT are used to perform the task of tokenizing the sentences as it has been proven to be excellent for masking the tokens and using the autoregressive and sequential model to predict the next tokens. The system then matches the skills ,educational qualifications and experience etc. of the candidate with those required by the company and finds those candidates which are most similar to them.

Proposed website can be used by the candidates to upload their resumes and find out which job they are most suitable for and can apply for, and which skills they should acquire more to get their desired positions.

Proposed system architecture

The proposed system architecture goes through various stages as shown below.

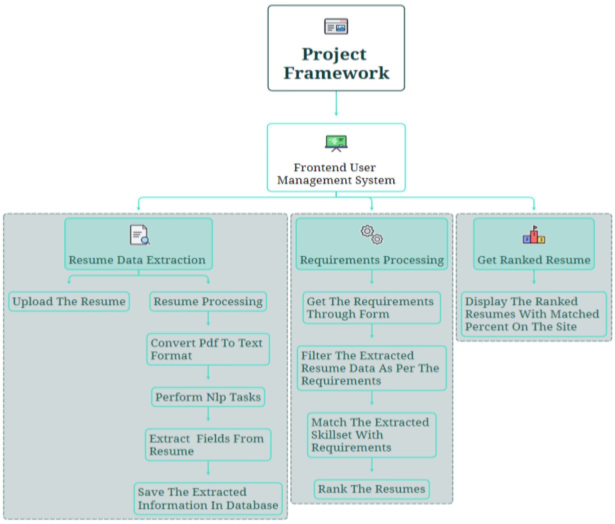

Project Framework
Frontend User Management System
Resume Data Extraction
Upload The Resume
Resume Processing
Convert Pdf To Text Format
Perform Nlp Tasks
Extract Fields From Resume
Save The Extracted Information In Database
Requirements Processing
Get The Requirements Through Form
Filter The Extracted Resume Data As Per The Requirements
Match The Extracted Skillset With Requirements
Rank The Resumes
Get Ranked Resume
Display The Ranked Resumes With Matched Percent On The Site

Figure 3.5 Proposed System Architecture

The project is divided into three main domains:

1. Resume Data Extraction: This includes steps like uploading of resume, resume processing, conversion from pdf to text, Performing various nlp tasks such as tokenization, stopword removal, cleaning, and name entity recognition.Followed by data extraction from the resume. And saving the data into the database.

2. Requirement Processing: This is the part where the extracted skills are matched with the requirements specified by the HR and the resumes are ranked accordingly.

3. Get Ranked Resume: This part will actually display the ranked resumes along with the matched percentage on the website to the users.

Website framework for project:

The website contains various pages among which some are common to all users and don't need any authentication. For example About us, contact us, FAQ and homepage that is the landing page for our site.

The other pages can be accessed by authenticated users only which are divided into two services one for the candidates where they can upload a resume and get the rank according to the selected designation. And the other one is for employers where they can specify the requirements for the job and in turn get the top matching resumes for the position.

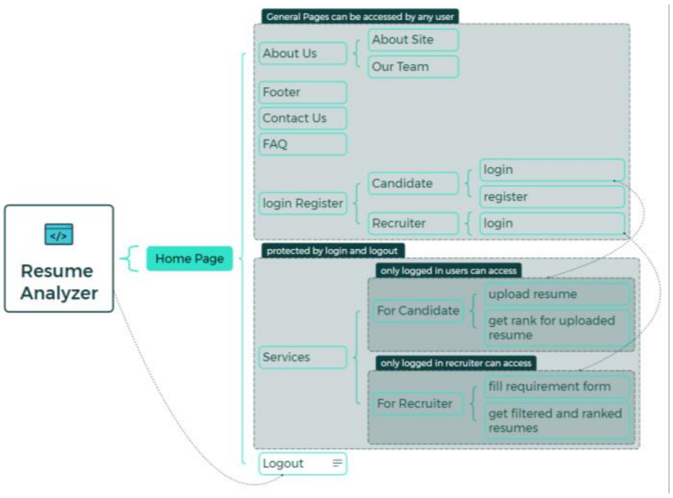

Resume Analyzer
Home Page
General Pages can be accessed by any user
About Us
About Site
Our Team
Footer
Contact Us
FAQ
login Register
Candidate
login
register
Recruiter
login
protected by login and logout
Services
only logged in users can access
For Candidate
upload resume
get rank for uploaded resume
only logged in recruiter can access
For Recruiter
fill requirement form
get filtered and ranked resumes
Logout

Figure 3.6 Website Framework

B. *Implementation Details*

The following are the implementation details of the proposed system.

3.3.1 *Methodology*

The literature review revealed that the most effective approach for accurately identifying named entities, such as skills and qualifications, is utilizing the DistilBERT transformer. In this method, tokens in the input data are randomly masked, and the model attempts to predict these masked tokens. A mask (mt) is applied where it equals 1 for removed tokens and 0 elsewhere. Pretraining the model with this task aids in learning data patterns, resulting in effective performance in downstream tasks using the acquired parameters. However, a limitation of this model is its assumption that each missing token depends on all input tokens but remains independent of other masked tokens. Furthermore, while BERT employs <mask> tokens during pre-training, real data lacks these missing tokens, creating a disparity between pre-training and fine-tuning.

In contrast, these issues are absent in AutoRegressive (AR) modeling, where the objective is to predict the next word based on the sequence of words preceding or succeeding it. Consequently, the methodology adopted in the proposed system utilizes a DistilBERT transformer. DistilBERT is a transformer employing an encoder-decoder architecture, particularly effective in named entity recognition tasks.

DistilBERT, as the name suggests, is a distilled and more compact version of the BERT (Bidirectional Encoder Representations from Transformers) model. It works based on similar principles but employs strategies to reduce model size while retaining much of BERT's language understanding capabilities. Here's an overview of how DistilBERT works:

Transformer Architecture: DistilBERT, like BERT, is based on the Transformer architecture. The Transformer architecture is designed to capture contextual information from both left and right sides of a word in a sentence. It consists of an encoder-decoder structure, but models like BERT and DistilBERT typically use only the encoder.

Distillation: The core innovation in DistilBERT is knowledge distillation. In this process, a larger, teacher model, like the original BERT, is used to teach a smaller, student model, which is DistilBERT. The teacher model's knowledge is transferred to the student model. The teacher model helps the student model understand and replicate the complex language representations it has learned during pre-training.

The input taken from the user in form of doc,docx,pdf etc form is first converted into raw text. The raw text is passed into the lexical analyzer which converts the text into tokens. Named entities such as skills,educational qualifications, age, experience etc are recognized from the text.

The entities are then matched with the available data from the company and the similarity between the required qualifications and qualifications of the candidate are found to find the most suitable candidate.

Identification of Input / Dataset /Output

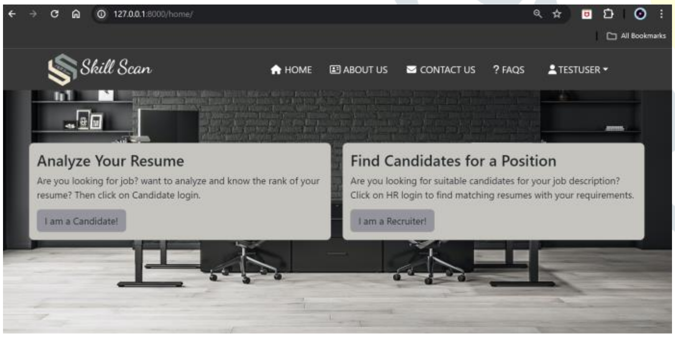
A google form was used to get the consent of the people to download their resumes from their linkedin profiles. Over all a collection of 50 resumes was formed which was further used for parsing and creating the extracted skills dataset.

III. RESULT AND DISCUSSIONS

Sample of Input and Output screenshots

A. Home page

The Home page contains two options for logging in, that is Candidate login, through which candidates can upload their resumes to match with required skills by employers, and Recruiter login, through which recruiters can find candidates which fulfill their requirements.



Analyze Your Resume

Are you looking for job? want to analyze and know the rank of your resume? Then click on Candidate login.

I am a Candidate!

Find Candidates for a Position

Are you looking for suitable candidates for your job description? Click on HR login to find matching resumes with your requirements.

I am a Recruiter!

Figure 4.1(a) Home page

B. Sign Up

In the proposed resume analyser system (SkillScan), the user should register to create an account for analyzing their resume. These details are securely stored in our database which can be accessed later for logging in.

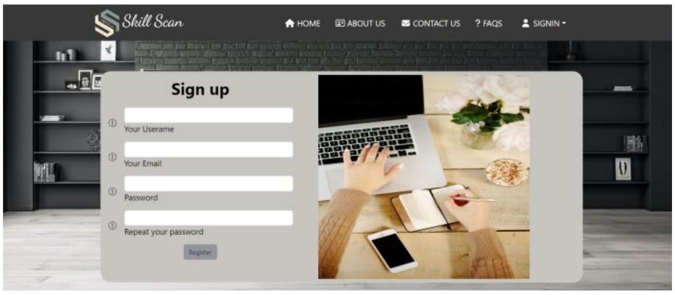


Figure 4.1(b) Sign up

C. Sign In

The Sign in page is where the user can sign-in to their account if they have already registered on the website.These details are matched for verification with the details in our database and then access is provided.

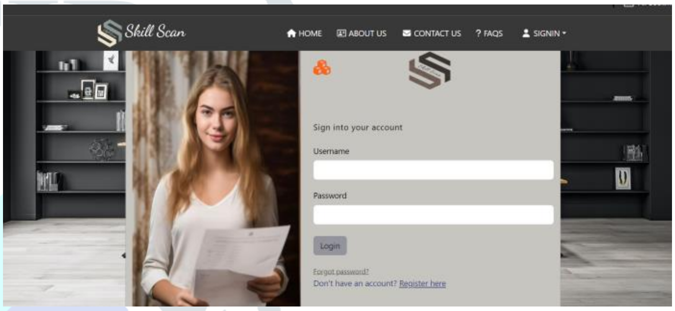


Figure 4.1(c) Login Page for candidate

D. Upload Page

The upload page is where the user can fill in the form using details like name and email id, and then upload their resume file, which can be of any format such as pdf, doc, etc.

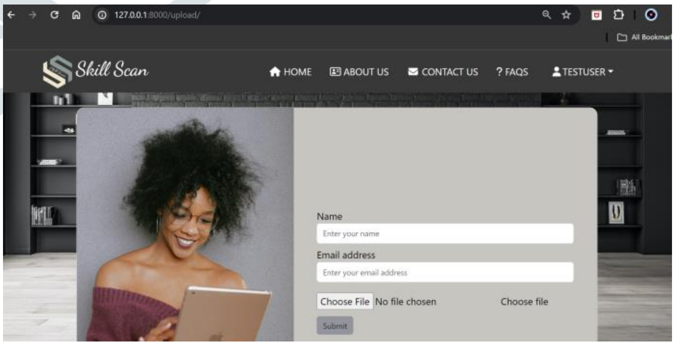
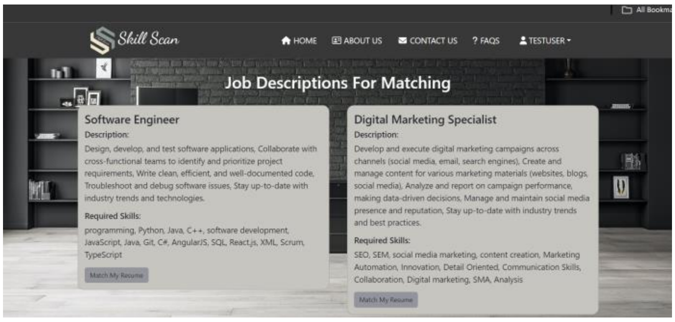


Figure 4.1(d)Page to Upload Resume

E. Select Job Description to match with

After uploading their resume the candidates can select from these job descriptions to match their resume with and get the results



Job Descriptions For Matching

Software Engineer

Description:

Design, develop, and test software applications, Collaborate with cross-functional teams to identify and prioritize project requirements, Write clean, efficient, and well-documented code, Troubleshoot and debug software issues, Stay up-to-date with industry trends and technologies.

Required Skills:

programming, Python, Java, C++, software development, JavaScript, Java, Git, C#, AngularJS, SQL, React.js, XML, Scrum, TypeScript

Digital Marketing Specialist

Description:

Develop and execute digital marketing campaigns across channels (social media, email, search engines), Create and manage content for various marketing materials (websites, blogs, social media), Analyze and report on campaign performance, making data-driven decisions, Manage and maintain social media presence and reputation, Stay up-to-date with industry trends and best practices.

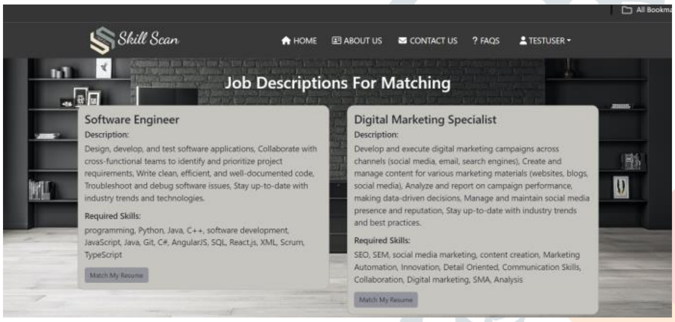
Required Skills:

SEO, SEM, social media marketing, content creation, Marketing Automation, Innovation, Detail Oriented, Communication Skills, Collaboration, Digital marketing, SMA, Analysis

Figure 4.1(e) Job description

F. Jobs From Remotive

The Remotive Jobs page is where the API to extract jobs from the website is implemented. The jobs are displayed by calling their API and getting the details of the job. Users can view the job on their website by clicking View Details.



Job Descriptions For Matching

Software Engineer

Description:

Design, develop, and test software applications, Collaborate with cross-functional teams to identify and prioritize project requirements, Write clean, efficient, and well-documented code, Troubleshoot and debug software issues, Stay up-to-date with industry trends and technologies.

Required Skills:

programming, Python, Java, C++, software development, JavaScript, Java, Git, C#, AngularJS, SQL, React.js, XML, Scrum, TypeScript

Digital Marketing Specialist

Description:

Develop and execute digital marketing campaigns across channels (social media, email, search engines), Create and manage content for various marketing materials (websites, blogs, social media), Analyze and report on campaign performance, making data-driven decisions, Manage and maintain social media presence and reputation, Stay up-to-date with industry trends and best practices.

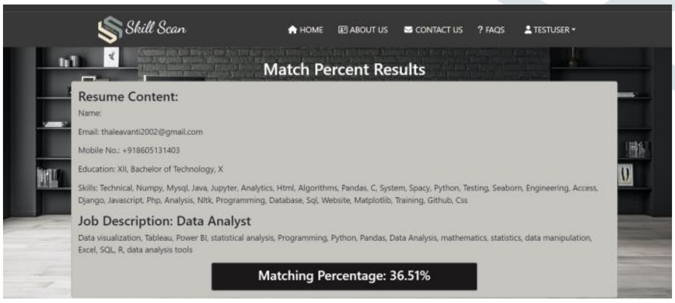
Required Skills:

SEO, SEM, social media marketing, content creation, Marketing Automation, Innovation, Detail Oriented, Communication Skills, Collaboration, Digital marketing, SMA, Analysis

Figure 4.1(f) Jobs From Remotive

G. Matching Results

The uploaded resume has been matched with the selected job description. For instance, the Data Analyst position shows a matching percentage of 36.51%.



Match Percent Results

Resume Content:

Name:

Email: thaleavanti2002@gmail.com

Mobile No.: +918605131403

Education: XII, Bachelor of Technology, X

Skills: Technical, Numpy, Mysql, Java, Jupyter, Analytics, Html, Algorithms, Pandas, C, System, Spacy, Python, Testing, Seaborn, Engineering, Access, Django, Javascript, Php, Analysis, Nltk, Programming, Database, Sql, Website, Matplotlib, Training, Github, Css

Job Description: Data Analyst

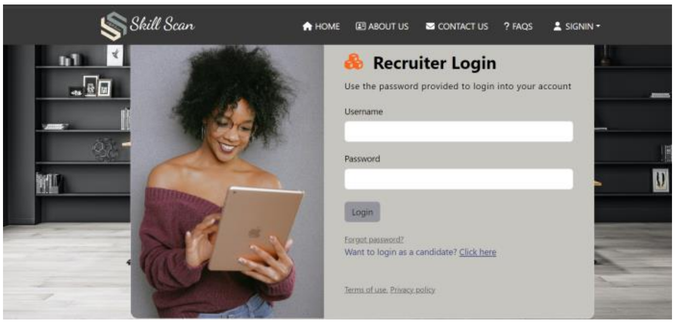
Data visualization, Tableau, Power BI, statistical analysis, Programming, Python, Pandas, Data Analysis, mathematics, statistics, data manipulation, Excel, SQL, R, data analysis tools

Matching Percentage: 36.51%

Figure 4.1(g) Matching Result

H. Recruiter Login

Another part of our project is secured for recruiters who will login through this page. Recruiters will act as super users who will have only one option that is login. One cannot register as a recruiter into our site directly.



Recruiter Login

Use the password provided to login into your account

Username

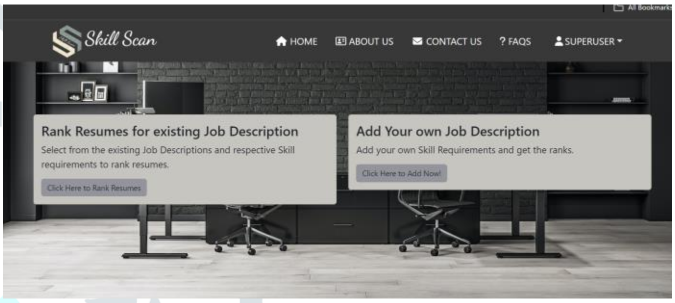
Password

Want to login as a candidate? Click here

Figure 4.1(h) Login page for recruiter

I. After Recruiter Login

Recruiter can log in as admin to access the recruiter page.



Rank Resumes for existing Job Description

Select from the existing Job Descriptions and respective Skill requirements to rank resumes.

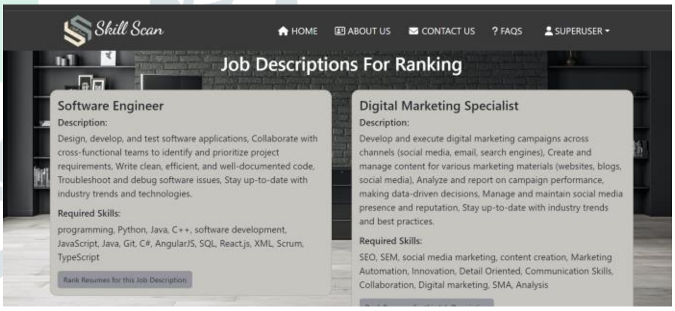
Add Your own Job Description

Add your own Skill Requirements and get the ranks.

Figure 4.1(i) After Recruiter Login

J. Job Descriptions for Ranking

The recruiter can view all the job descriptions they have uploaded or available.



Job Descriptions For Ranking

Software Engineer

Description:

Design, develop, and test software applications, Collaborate with cross-functional teams to identify and prioritize project requirements, Write clean, efficient, and well-documented code, Troubleshoot and debug software issues, Stay up-to-date with industry trends and technologies.

Required Skills:

programming, Python, Java, C++, software development, JavaScript, Java, Git, C#, AngularJS, SQL, React.js, XML, Scrum, TypeScript

Digital Marketing Specialist

Description:

Develop and execute digital marketing campaigns across channels (social media, email, search engines), Create and manage content for various marketing materials (websites, blogs, social media), Analyze and report on campaign performance, making data-driven decisions, Manage and maintain social media presence and reputation, Stay up-to-date with industry trends and best practices.

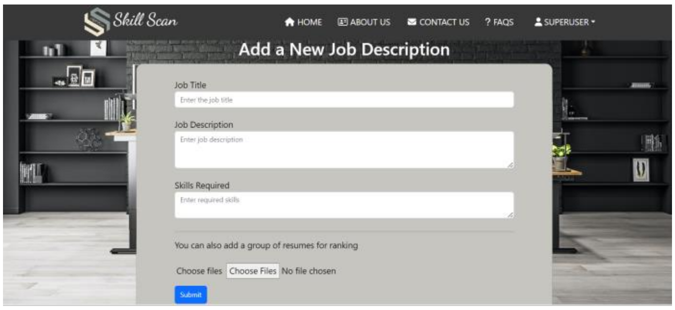
Required Skills:

SEO, SEM, social media marketing, content creation, Marketing Automation, Innovation, Detail Oriented, Communication Skills, Collaboration, Digital marketing, SMA, Analysis

Figure 4.1(j) Job description for ranking

K. New Job Description Form

The requirements form is available for the recruiter to select the parameters on which they want to filter candidates, and the details of the job they are posting, ex. title, description, etc.



Add a New Job Description

Job Title

Enter the job title

Job Description

Enter job description

Skills Required

Enter required skills

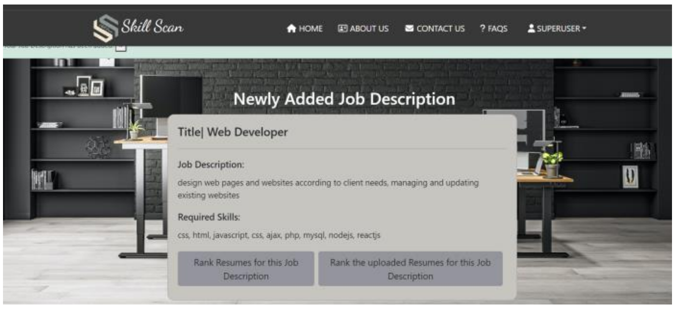
You can also add a group of resumes for ranking

Choose files Choose Files No file chosen

Figure 4.1(k) New job description

L. After New Job Description

The page after the new job description is added by the recruiter.


Skill Scan
HOME
ABOUT US
CONTACT US
? FAQS
SUPERUSER

Newly Added Job Description

Title| Web Developer

Job Description:

design web pages and websites according to client needs, managing and updating existing websites

Required Skills:

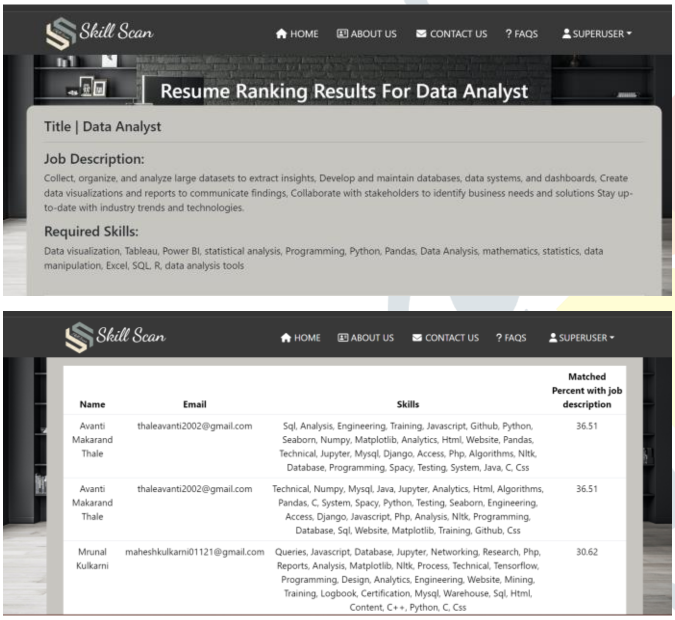
css, html, javascript, css, ajax, php, mysql, nodejs, reactjs

Rank Resumes for this Job Description
Rank the uploaded Resumes for this Job Description

Figure 4.1(l) After new job description

M. Ranking Results

After selecting the job description for which the recruiter wants to select a candidate, they can view the resumes uploaded by candidates ranked in an order of highest to lowest percentage, after finding the percentage match with the description.



Resume Ranking Results For Data Analyst

Title | Data Analyst

Job Description:

Collect, organize, and analyze large datasets to extract insights, Develop and maintain databases, data systems, and dashboards, Create data visualizations and reports to communicate findings, Collaborate with stakeholders to identify business needs and solutions Stay up-to-date with industry trends and technologies.

Required Skills:

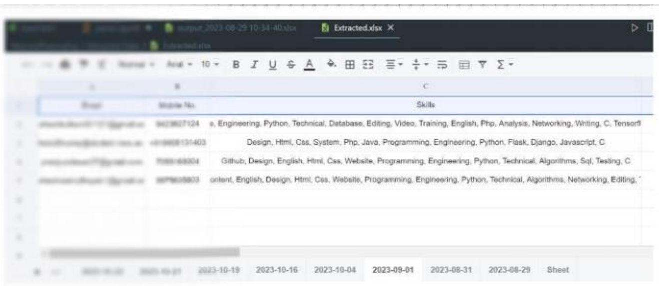
Data visualization, Tableau, Power BI, statistical analysis, Programming, Python, Pandas, Data Analysis, mathematics, statistics, data manipulation, Excel, SQL, R, data analysis tools

Name	Email	Skills	Matched Percent with job description
Avanti Makarand Thale	thaleavanti2002@gmail.com	Sql, Analysis, Engineering, Training, Javascript, Github, Python, Seaborn, Numpy, Matplotlib, Analytics, Html, Website, Pandas, Technical, Jupyter, Mysql, Django, Access, Php, Algorithms, Nltk, Database, Programming, Spacy, Testing, System, Java, C, Css	36.51
Avanti Makarand Thale	thaleavanti2002@gmail.com	Technical, Numpy, Mysql, Java, Jupyter, Analytics, Html, Algorithms, Pandas, C, System, Spacy, Python, Testing, Seaborn, Engineering, Access, Django, Javascript, Php, Analysis, Nltk, Programming, Database, Sql, Website, Matplotlib, Training, Github, Css	36.51
Mrunal Kulkarni	maheshkulkarni01121@gmail.com	Queries, Javascript, Database, Jupyter, Networking, Research, Php, Reports, Analysis, Matplotlib, Nltk, Process, Technical, Tensorflow, Programming, Design, Analytics, Engineering, Website, Mining, Training, Logbook, Certification, Mysql, Warehouse, Sql, Html, Content, C++, Python, C, Css	30.62

Figure 4.1(m) Ranking result

N. Sample of extracted Data

Data extracted form resumes stored in the form of a sheet. With a new sheet being created every day and saved with the name of that day's date.

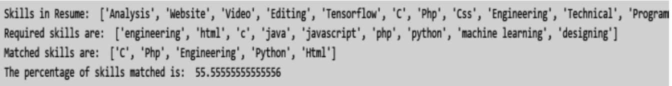


[illegible]	[illegible]	Skills
[illegible]	[illegible]	s, Engineering, Python, Technical, Database, Editing, Video, Training, English, Php, Analysis, Networking, Writing, C, Tensorfl
[illegible]	[illegible]	Design, Html, Css, System, Php, Java, Programming, Engineering, Python, Flask, Django, Javascript, C
[illegible]	[illegible]	Github, Design, English, Html, Css, Website, Programming, Engineering, Python, Technical, Algorithms, Sql, Testing, C
[illegible]	[illegible]	ontent, English, Design, Html, Css, Website, Programming, Engineering, Python, Technical, Algorithms, Networking, Editing,

Figure 4.1(n) A sample of extracted data

O. Matching the extracted skills

This is the most important part of our project where the extracted skills from the resume are matched with the specified skills or requirements in the job description and the percentage of patching is calculated.



```
Skills in Resume:  ['Analysis', 'Website', 'Video', 'Editing', 'Tensorflow', 'C', 'Php', 'Css', 'Engineering', 'Technical', 'Program
Required skills are:  ['engineering', 'html', 'c', 'java', 'javascript', 'php', 'python', 'machine learning', 'designing']
Matched skills are:  ['C', 'Php', 'Engineering', 'Python', 'Html']
The percentage of skills matched is:  55.55555555555556
```

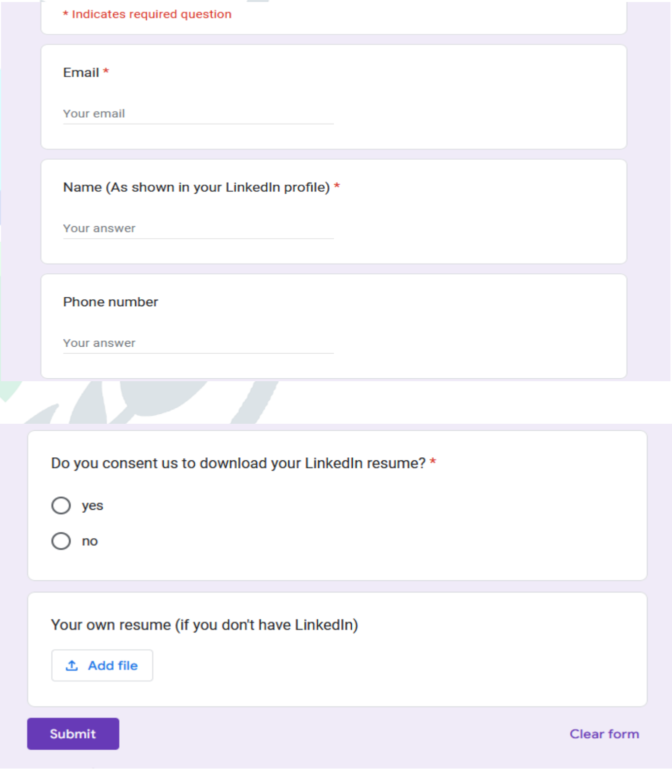
Figure 4.1(o) Matching of the extracted and

required skills

A. Evaluation Parameters

4.2.1 Datasets Used

The dataset used for testing the resume analyzer is created by collecting resumes from various subjects through online form. The snapshot of the form circulated is given below.



* Indicates required question

Email *

Your email

Name (As shown in your LinkedIn profile) *

Your answer

Phone number

Your answer

Do you consent us to download your LinkedIn resume? *

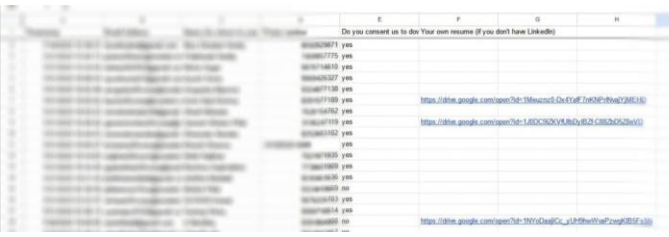
○ yes

○ no

Your own resume (if you don't have LinkedIn)

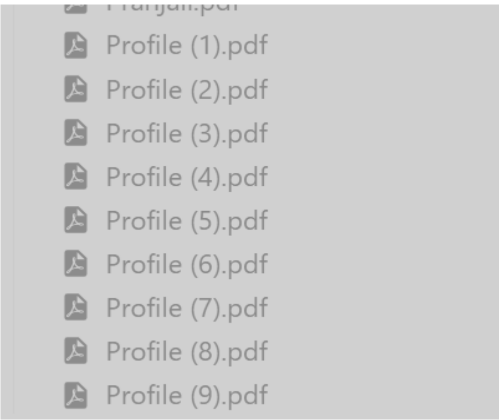
figure 4.2 (a) Form

The resumes collected are of various types such as pdf, doc, etc. The resumes collected are also of various formats such as some containing tables, images, graphics, etc.



[illegible]	[illegible]	[illegible]	[illegible]	Do you consent us to dov	Your own resume (if you don't have LinkedIn)
[illegible]	[illegible]	[illegible]	[illegible]	yes	
[illegible]	[illegible]	[illegible]	[illegible]	yes	
[illegible]	[illegible]	[illegible]	[illegible]	yes	
[illegible]	[illegible]	[illegible]	[illegible]	yes	
[illegible]	[illegible]	[illegible]	[illegible]	yes	
[illegible]	[illegible]	[illegible]	[illegible]	yes	https://drive.google.com/open?id=1Meuznz0-Dx4YafF7nKNPrlNvajYjMEHE
[illegible]	[illegible]	[illegible]	[illegible]	yes	
[illegible]	[illegible]	[illegible]	[illegible]	yes	https://drive.google.com/open?id=1J0DC9ZKVfUlbDyIBZf-C882bD5Z8eVD
[illegible]	[illegible]	[illegible]	[illegible]	yes	
[illegible]	[illegible]	[illegible]	[illegible]	yes	
[illegible]	[illegible]	[illegible]	[illegible]	yes	
[illegible]	[illegible]	[illegible]	[illegible]	yes	
[illegible]	[illegible]	[illegible]	[illegible]	yes	
[illegible]	[illegible]	[illegible]	[illegible]	no	
[illegible]	[illegible]	[illegible]	[illegible]	yes	
[illegible]	[illegible]	[illegible]	[illegible]	yes	
[illegible]	[illegible]	[illegible]	[illegible]	no	https://drive.google.com/open?id=1NYoDaaJjCc_yUH9hwWvaPzwgKIB5Fs5b

Figure 4.2 (b) Received Responses



Profile (1).pdf
Profile (2).pdf
Profile (3).pdf
Profile (4).pdf
Profile (5).pdf
Profile (6).pdf
Profile (7).pdf
Profile (8).pdf
Profile (9).pdf

Figure 4.2 (c) Resumes collected

Evaluation Metrics

The different evaluation metrics used to analysis our resume analysis system are given below

4.2.3 Precision

Precision is the ratio of true positives over the sum of false positives and true negatives. It's also referred to as positive predictive value. Precision indicates the accuracy of positive predictions made by the model.

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}} = \frac{\textit{True Positive}}{\textit{Total Predicted Positive}} \qquad (2)$$

In our resume analysis model, precision represents the ratio of correctly predicted labels for entities to all entities labeled by the model. For instance, it's the ratio of all entities correctly labeled as "Name" by the model to all entities labeled as "Name" by the model.

4.2.4 Recall

Recall, on the other hand, is the ratio of correctly predicted outcomes to all predictions. It's also known as sensitivity or specificity.

$$\text{Recall} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Negative}} = \frac{\textit{True Positive}}{\textit{Total Actual Positive}} \qquad (3)$$

Recall signifies the proportion of positive instances captured by the model through positive labeling. When the cost of False Negative outweighs that of False Positive, the best model selection is based on Recall. In our resume analysis model, recall is the ratio of correctly predicted labels for entities to all entities labeled by the model. For example, it's the ratio of all entities correctly labeled as "Name" by the model to all entities actually labeled as "Name".

4.2.5 Accuracy

Accuracy is the proportion of correct predictions out of all predictions made by an algorithm. It can be computed by dividing precision by recall or as 1 minus the quotient of false negative rate (FNR) divided by false positive rate (FPR).

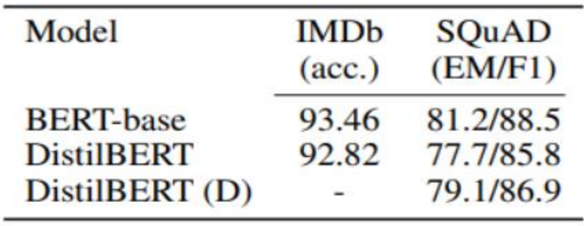
$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \qquad (4)$$

For our system, the accuracy is the ratio of all the entities labeled as " Name" by model to the total number of entities labeled.

1. Language Comprehension and Generalization Proficiency: The assessment of DistilBERT's language comprehension and generalization capabilities focused on the GLUE benchmark, which comprises a collection of nine datasets and tasks. Notably, DistilBERT consistently demonstrated superiority when compared to the ELMo baseline, consistently outperforming it. Additionally, DistilBERT surpasses BERT by retaining an impressive 97% of performance while utilizing 40% fewer parameters.

2. DistilBERT's Performance on Practical Applications: In evaluating DistilBERT's effectiveness in real-world applications, it was observed that DistilBERT exhibits a minor deficit of only 0.6% in test accuracy when compared to the IMDb benchmark, while maintaining a 40% reduction in size compared to BERT. Furthermore, it is essential to note that DistilBERT, while slightly trailing BERT in performance on SQuaD, provides noteworthy advantages in terms of efficiency.

Table 4.1: Performance of DistilBERT in Practical Applications



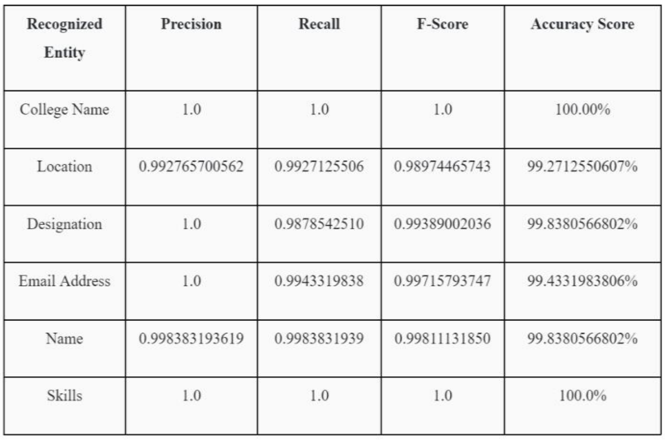
Model	IMDb (acc.)	SQuAD (EM/F1)
BERT-base	93.46	81.2/88.5
DistilBERT	92.82	77.7/85.8
DistilBERT (D)	-	79.1/86.9

3. Balancing Speed and Model Size with DistilBERT: An examination of the speed and size trade-off with DistilBERT involved a comparison of model parameters and inference times for a full pass on the STSB development set using a CPU. Notably, DistilBERT emerges as the more efficient option with a 40% reduction in parameters and approximately 60% faster inference speeds compared to BERT.

Result Analysis

To evaluate the model's performance, we employed the "Resume Entities for Named Entity Recognition (NER)" dataset provided by DataTurks. This dataset comprises 220 resumes sourced from an online job platform. These documents were then uploaded to an online annotation tool and manually annotated. The tool facilitates the creation of annotations for key entities of interest, such as email addresses, skills, phone numbers, etc., by automatically parsing the documents. Subsequently, the tool generates training data in JSON format, where each line includes the text corpus alongside the corresponding annotations, represented by labels and points. This dataset serves as the benchmark for assessing the model's accuracy and effectiveness in identifying relevant entities within resumes.

Table 4.2: Result



Recognized Entity	Precision	Recall	F-Score	Accuracy Score
College Name	1.0	1.0	1.0	100.00%
Location	0.992765700562	0.9927125506	0.98974465743	99.2712550607%
Designation	1.0	0.9878542510	0.99389002036	99.8380566802%
Email Address	1.0	0.9943319838	0.99715793747	99.4331983806%
Name	0.998383193619	0.9983831939	0.99811131850	99.8380566802%
Skills	1.0	1.0	1.0	100.0%

The result obtained shows how accurately the model predicted the various entities in the resume. For example, the accuracy score of email_id is 100.00%, which means that the email id of candidate, such as, “user@gmail.com” was identified correctly from the resume in all the samples provided for testing.Similarly, the Name of the candidate has accuracy 99.8%, which means the ratio of correctly identified names to all labeled names is 99.8%, for ex, the probability of model labeling “Sakshi Patil” as name, is 99.8%.

We find the percentage of match of resume to required qualifications is found by finding the Cosine similarity of the two lists containing required and actual skills.

$$similarity(A,B) = cos(\theta) = \frac{A \cdot B}{||A|| \, ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2}}. \quad (5)$$

For example:

Skills in Resume are : ['Analysis', 'Website', 'Video', 'Editing', 'Tensorflow', 'C', 'Php', 'Css', 'Engineering', 'Technical', 'Programming', 'Research', 'Networking', 'Content', 'Database', 'Python', 'English', 'Training', 'Writing', 'Html']

Required skills are: ['engineering', 'html', 'c', 'java', 'javascript', 'php', 'python', 'machine learning', 'designing']

Matched skills are: ['C', 'Php', 'Engineering', 'Python', 'Html']

The percentage of skills matched is: 55.55555555555556

IV. CONCLUSION

The application is able to extract valuable information from the candidate’s resume, such as name, phone number, skills and education, which are used to find how qualified the candidate is for a given job description, using Cosine similarity.The DistilBERT model of NLP is used to train the model to recognize these details from resume. The model is able to extract correct details with a 99% accuracy.

REFERENCES

[1] X. Li, H. Shu, Y. Zhai and Z. Lin, "A Method for Resume Information Extraction Using BERT-BiLSTM-CRF," 2021 IEEE 21st International Conference on Communication Technology (ICCT), Tianjin, China, 2021, pp. 1437-1442, doi: 10.1109/ICCT52962.2021.9657937.

[2] Y. -C. Chou, C. -Y. Chao and H. -Y. Yu, "A Résumé Evaluation System Based on Text Mining," 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Okinawa, Japan, 2019, pp. 052-057, doi: 10.1109/ICAIIC.2019.8669066.

[3] Bhor, Shubham, Vivek Gupta, Vishak Nair, Harish Shinde, and Manasi S. Kulkarni. "Resume parser using natural language processing techniques." *Int. J. Res. Eng. Sci* 9, no. 6 (2021).

[4] Daryani, Chirag & Chhabra, Gurneet & Patel, Harsh & Chhabra, Indrajeet & Patel, Ruchi. (2020). AN AUTOMATED RESUME SCREENING SYSTEM USING NATURAL LANGUAGE PROCESSING AND SIMILARITY. 99-103. 10.26480/etit.02.2020.99.103.

[5] Singh A, Pal S, Singh SP. RESUME RECOMMENDATION SYSTEM USING AI.

[6] Anand, Avisha & Dubey, Mr. (2022). CV Analysis Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology. 10. 1316-1322. 10.22214/ijraset.2022.42295.

[7]Chandola, Divyanshu, Aditya Kumar Garg, Ankit Maurya and Amit Kumar Kushwaha. “RESUME PARSING SYSTEM USING TEXT ANALYTICS.” (2015).

[8] Gopalakrishna, Suhas & Varadharajan, Vijayaraghavan. (2019). Automated Tool for Resume Classification Using Semantic Analysis. International Journal of Artificial Intelligence & Applications. 10. 11-23. 10.5121/ijaia.2019.10102.

[9] Green, N., Liu, X., Murphy, D., (2020). Developing an Electronic Resume Analyzer Portal (e-RAP): A Natural Language Processing Approach to Enhance College Graduates Job Readiness. *Information Systems Education Journal*18(3) pp 28-37. http://ISEDJ.org/2020-3/ ISSN : ISSN: 1545-679X.