TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
HIMALAYA COLLEGE OF ENGINEERING

A FINAL YEAR MAJOR PROJECT REPORT

ON

PERSONALITY EVALUATION AND
CV ANALYSIS

[CT-755]

SUBMITTED TO
**DEPARTMENT OF ELECTRONICS AND COMPUTER
ENGINEERING**
Chyasal, Lalitpur

SUBMITTED BY
Dristi Dugar (89862)

Gayatri Gharti Magar (89864)

Ichhya Amatya (89866)

Saraswati Subedi (89883)

March, 2021

PERSONALITY EVALUATION AND CV ANALYSIS

A FINAL YEAR MAJOR PROJECT REPORT

[CT-755]

"A FINAL YEAR MAJOR PROJECT REPORT SUBMITTED FOR PARTIAL FULFILLMENT OF THE DEGREE OF BACHELORS' IN COMPUTER ENGINEERING"

SUPERVISOR

Er. Ashok GM

SUBMITTED TO

TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

HIMALAYA COLLEGE OF ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

Chyasal, Lalitpur

SUBMITTED BY

Dristi Dugar (89862)

Gayatri Gharti Magar (89864)

Ichhya Amatya (89866)

Saraswati Subedi (89883)

March, 2021

ACKNOWLEDGEMENT

ABSTRACT

This project will enable a more effective way to grade submitted candidate CVs from a large number of applicants providing a consistent and fair CV ranking policy. System will rank the experience, qualifications, projects and key skills required for particular job position. This system also employs personality evaluation (Big5 Personality test) through standard questionnaires. Finally, combining the results of both CV analysis and personality evaluation all the candidates are graded accordingly which can be viewed by the HR department to select the suitable candidate. This system will help the human resource department to select right candidate for particular job profile, which in turn provide expert workforce for the organization. Candidates are required to upload their own CV into the system, which will analyze the professional eligibility based on the information of CV. The system employs a NER model, in which CV and job description are annotated manually to build a trained JSON set. This model then parses and extracts chief entities to summarize the resume with only relevant information which the evaluator is looking for and also to summarize the skillset from recruiter description. The system then performs pre-processing to the summarized skillset and also does synonyms detection and replacement using the WordNet dictionary. Then, the system employs a NLTK approach using TF-IDF algorithm to determine most relevant keywords in the CV for the given job and compare the CV with job description skillset using cosine similarity to finally rank the CV. Candidates also may perform an aptitude test, which will be conducted on questionnaire provided by the recruiter to add more value to the judgement. Thus, the system provides a helping hand for the recruitment process so that the candidate's CV will be shortlisted and fair decision will be made.

Keywords: *workforce, recruitment, WordNet, NER, NLTK, TF-IDF*

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ATS   Applicant Tracking System

CV   Curriculum Vitae

DFD   Data Flow Diagram

ER   Entity Relationship

HR   Human Resources

JSON   Java Script Object Notation

LLC   Limited Liability Company

ML   Machine Learning

NER   Named Entity Recognition

NLP   Natural Language Processing

NLTK   Natural Language Tool Kit

NoSQL   Not Only SQL

SQL   Structured Query Language

TF-IDF   Term Frequency – Inverse Document Frequency

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Background

After completing education the next phase that comes in a person's life is job. While searching for jobs the most important thing to represent an applicant is Curriculum Vitae (CV) or Resume. In this era of technology, job searching has become more smart and easier at the same time. However, there are more than enough applicants for a single job and it is really tough for HR to select candidates only based on their CV/Resume. To solve this problem, there are companies who provide specific format for their applicants so that they can make this process a little bit easier. Even after doing that the process is still pretty boring and most of the cases full of errors. This project will basically benefit the organizations to easily shortlist the candidate based on the CV ranking policy.

Human personality has played a vital role in an individual's life as well as in the development of an organization. One of the ways to judge human personality is by using standard questionnaires or by analyzing the Curriculum Vitae (CV). Traditionally, recruiters manually shortlist/filters a candidate's CV as per their requirements. In our project, a system is a built that automates the eligibility check and personality evaluation of candidates in a recruitment process. An online application is developed for the analysis of aptitude test, personality test and candidate's CV.

Recommendation using machine learning techniques have been used for the analysis of the CV. In literature, various evaluation tools have been used. One of the approaches has been mentioned in which, a tool called "Career Mapper" is used for the recommendations of the CV. It checks for the completeness of the user profile. The recommendation usually involves the use of various filters. Content and Collaboration are among them. One of the approaches of Content-Based Recommender is mentioned which uses Fo-DRA for the recommendation which is based on content.

'Jobscan' is another tool that gives job seekers an instant analysis of how well their resume is tailored for a particular job, along with how it can be even better optimized for an applicant tracking system (also known as an ATS).

Such parsing tools facilitate evaluation of resumes at a quick glance, thereby simplifying the effort required in shortlisting candidates among a pile of resumes.

1.2 Problem Statement

Finding the right person for the right job has never been an easy feat for companies, whose value is very often to a large degree derived from their manpower. With the increased mobility of job seekers in recent years, more and more jobs are seeing rapidly growing pools of potential candidates, requiring respective recruiters have to wade through hundreds if not thousands of CVs to find the perfect match. When there is a large number of candidates, the automation can also include a more challenging task: scoring and ranking candidates' CVs according to their match to a job posting.

Most of these parsing tools, parse relevant skillset from CV and provide how relevant the CV is in comparison to the set skillset. But, such parsing and ranking tasks involve a great amount of NLTK processing to exactly find similarities between the CV and set constraints. It also need to take into consideration similar words of keywords while performing keyword matching. And, since many recruiter now also perform pre-employment personality testing, it will be more effective of a judgement in adding another layer to the recruitment task.

This system strives to solve these problems by providing a fair CV ranking or grading policy for an automated recruitment system which shortlists the suitable candidates for the job. The major objective of our system is to take the current resume ranking system to another level by adding in more layers to nltk processing and adding in personality evaluating feature to the system and making it more flexible for both employee and employers.

1.3 Objectives

The general objective of our project is to develop a web based application to evaluate personality and analyze CV. The specific objective of our project is:

- To provide an efficient CV grading system accompanied by personality evaluation with the use of NER model and TF-IDF algorithm.

1.4 Project Scope and Applications

Hiring the right person for the right job is a common challenge faced by all companies. Especially for positions with a large number of applicants the search for the right candidate can feel like looking for a needle in the haystack. In these situations traditional methods of recruitment can be too expensive and time-consuming to be a viable option. Hence, not surprisingly, recruitment technology that can facilitate this process are in high demand making this system a very useful and applicable system with a wide scope in this field.

This can be applied in job hunting websites, parsing applicants' resume in large companies, third generation hiring systems, pre-employment testing, and commercial sectors having intake and eliminations in bulk like Government Jobs.
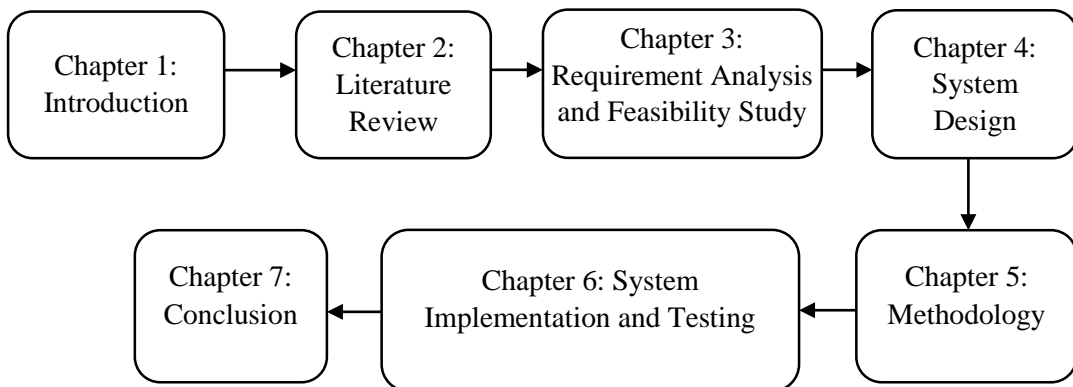
1.5 Report Organization

Fig 1.1 Report Organization

Chapter 1 includes the INTRODUCTION part and provides a general idea about the project is given here along with its objectives to be achieved with their problem statement, and scope and application. Chapter 2 contains the LITERATURE REVIEW part and provides a general research on Personality evaluation test, CV analysis and implementation of different previous ATS websites and system. Chapter 3 consists of REQUIREMENT ANALYSIS AND FEASIBILITY STUDY part and contains a discussion about the project requirements and feasibility study done during requirement analysis phase. Chapter 4 contains the SYSTEM DESIGN part and different diagrams depicting the analysis and design of the system is given here. The building blocks of the system and system flow is explained. Chapter 5 includes the METHODOLOGY part and a step-wise implementation of the working model is described in this chapter. Detailed explanation of CV Ranking and Personality Test system is given. Chapter 6 contains the SYSTEM IMPLEMENTATION AND TESTING part and here the test cases of the implementation phase is provided. Also, here the obtained results from the model is analyzed and accuracy of the trained model is measured in this chapter. Chapter 7 is the final section and contains the CONCLUSION part. It consists of project conclusion and discussion on the limitations as well as future enhancements.

2. LITERATURE REVIEW

A CV (short for the Latin phrase curriculum vitae, which means "course of life") is a detailed document highlighting your professional and academic history. CVs typically include information like work experience, achievements and awards, scholarships or grants you've earned, coursework, research projects and publications of your work. [1] CV are submitted while applying for any job by a job seeker and it plays the most important role in being called up for an interview for that job. [2] And, a job description is a document intended to provide job applicants with an outline of the main duties and responsibilities of the role for which they are applying. A job description is an essential part of the job application process as, with the right information, it should help applicants to determine whether the role is in line with their skill set and whether it is a job they actually want to do. From the organization's perspective, the job description is vital in ensuring that the applications received for the position closely match the needs of the role itself. It helps HR departments and external recruiters to streamline the selection process and receive a high concentration of candidates who are suitable for interview or further selection.

[3]Job profiles can affect job applicants and recruiters alike. For job seekers, writing a resume that doesn't meet the vital duties listed in the profile is akin to a futile job search. Since most firms today use applicant tracking systems (ATS) to narrow down resumes, you must include the right set of resume keyword into your copy. Checking the job profile is a smart way to beat the ATS. You can use these lists to know which keywords shall be included in your resume. For recruiters, a poorly written job description could attract the wrong set of job seekers. Not only will this slow down the entire hiring process; it could also lead to job mismatch. This could also lead to a higher turnover rate, since most of the hired staff can't meet the demands of the job. In turn, this will require you to repeat the hiring process all over again. Hence, it is important to publish a clear and complete job profile.

Resume writing websites like [4] 'resumeprofessionalwriters' provide ATS-friendly interview-winning professional-looking and keyword-optimized resumes.

Also, [5] Resume parsers analyze a resume, extract the desired information, and insert the information into a database with a unique entry for each candidate. Recruitment agencies work with CV/Resume Parsing tools to automate the storage and analysis of CV/Resume data. This saves recruiters hours of work by eliminating manual processing of each job application and CV they receive. [6] Once the resume has been analyzed, a recruiter can search the database for keywords and phrases and get a list of relevant candidates. Many parsers support semantic search, which adds context to the search terms and tries to understand intent in order to make the results more reliable and comprehensive. The candidates returned are ranked based on how closely they match the keywords and job profile.

Machine learning is extremely important for resume parsing. Each block of information needs to be given a label and sorted into the correct category, whether that's education, work history, or contact information. [7] Rule-based parsers use a predefined set of rules to parse the text. This method does not work for resumes because the parser needs to "understand the context in which words occur and the relationship between them." [8] For example, if the word "Harvey" appears on a resume, it could be the name of an applicant, refer to the college Harvey Mudd, or reference the company Harvey & Company LLC. The abbreviation MD could mean "Medical Doctor" or "Maryland". A rule-based parser would require incredibly complex rules to account for all the ambiguity and would provide limited coverage.

[9] This leads us to Machine Learning and specifically Natural Language Processing (NLP). NLP is a branch of Artificial Intelligence and it uses Machine Learning to understand content and context as well as make predictions. [7] Many of the features of NLP are extremely important in resume parsing. Acronym normalization and tagging accounts for the different possible formats of acronyms and normalizes them. Lemmatization reduces words to their root using a language dictionary and Stemming removes "s", "ing", etc. Entity extraction uses regex expressions, dictionaries, statistical analysis and complex pattern-based extraction to identify people, places, companies, phone numbers, email addresses, important phrases and more.

The blog [10] speaks about a field in Natural language Processing and Information Retrieval called Named Entity Recognition and how we can apply it for automatically generating summaries of resumes by extracting only chief entities like name, education background, skills, etc. It is often observed that resumes may be populated with excess information, often irrelevant to what the evaluator is looking for in it. Therefore, the process of evaluation of resumes in bulk often becomes tedious and hectic. Through our NER model, we could facilitate evaluation of resumes at a quick glance, thereby simplifying the effort required in shortlisting candidates among a pile of resumes. The first task at hand of course is to create manually annotated training data to train the model. For this purpose, 220 resumes were downloaded from an online jobs platform. These documents were uploaded to our online annotation tool and manually annotated. We use python's spaCy module for training the NER model. The model is then shown the unlabeled text and will make a prediction.

In the thesis [11], Parse information from a resume using natural language processing, find the keywords, cluster them onto sectors based on their keywords and lastly show the most relevant resume to the employer based on keyword matching. First, the user uploads a resume to the web platform. The parser parses all the necessary information from the resume and auto fills a form for the user to proofread. Once the user confirms, the resume is saved into our NoSQL database ready to show itself to the employers. Also, the user gets their resume in both JSON format and pdf.

The website [12], Jobscan is a tool that gives job seekers an instant analysis of how well their resume is tailored for a particular job, along with how it can be even better optimized for an applicant tracking system (also known as an ATS).

But, keyword matching doesn't take the synonyms of the keyword into consideration. For that purpose, a dictionary comes into play, to match similar words while keyword matching. Hence, similar words matching can provide a better accuracy for eventual ranking of such applicant tracking system (ATS).

[13] With each passing year, it becomes harder and harder to get a job — companies are inundated with thousands of candidates for one open position. In response to the overwhelming increase in potential candidates, employers have added an extra step in the recruitment process: pre-employment personality testing. A personality test is an assessment used by employers to help find a candidate whose character traits are best suited for a specific position. The pre-employment testing is designed to reveal particular aspects of a candidate's personality and estimate the likelihood that he or she will excel in such a position. Research shows that if an employee is placed in a position that doesn't match his or her personality, it often leads to lower engagement. Low employee engagement results in 21 percent lower productivity and about 45 percent higher turnover, and replacing employees is expensive. Pre-employment job personality tests are now delivered online, where they are processed instantaneously. Results are then verified and normed against thousands of other candidates, speeding up the hiring process and ensuring that the candidates who move forward are compatible with the company.

Personality prediction involves determining personality traits based on the big-five model. [14] Many contemporary personality psychologists believe that there are five basic dimensions of personality, often referred to as the "Big 5" personality traits. The five broad personality traits described by the theory are extraversion (also often spelled extroversion), agreeableness, openness, conscientiousness, and neuroticism (OCEAN). [15] The test consists of fifty items that you must rate on how true they are about you on a five point scale where 1=Disagree, 3=Neutral and 5=Agree. It takes most people 3-8 minutes to complete.

Personality test score gives a measure for the CV recruiters to select the best suited one for the job whose character traits are best match for a specific position. Adding an extra step to the resume filtering process, helps to consider both resume and personality of the job seekers.

3. REQUIREMENT ANALYSIS AND FEASIBILITY STUDY

Our system implements the agile methodology of software development lifecycle. Agile methodology promotes continuous iteration of development and testing throughout the software development lifecycle of the project. Both development and testing activities are concurrent. Our project was based on the core values of agile methodology as:

- Individual and team interactions over processes and tools
- Working software over comprehensive documentation
- Responding to change over following a plan

3.1 Functional Requirements

Functional requirements describe the behaviors (functions or services) of the system that support the user's goals, tasks or activities. The functional requirements of our application are:

- **Registration and Log In**

  Users need to create a new account and register into the system either as an employee or an employer. System has to make sure that the user is registered in the application so the user must be authenticated by login process.

- **Job creation**

  Users logged into the system as employers can create job objects along with requirements for the employees to apply to.

- **Job searching, viewing and applying**

  The users of employee modules can search for and view jobs uploaded by the employers and also apply to the respective jobs.

- **Personality and Aptitude test**

  The employees can also take standard personality tests and aptitude tests uploaded by the employers.

- **CV and Description annotation and analysis**

  Uploaded CVs and description are annotated by the help of spaCy models and analyzed using TF-IDF algorithm with cosine similarity to rank the CVs on the basis of their similarities.
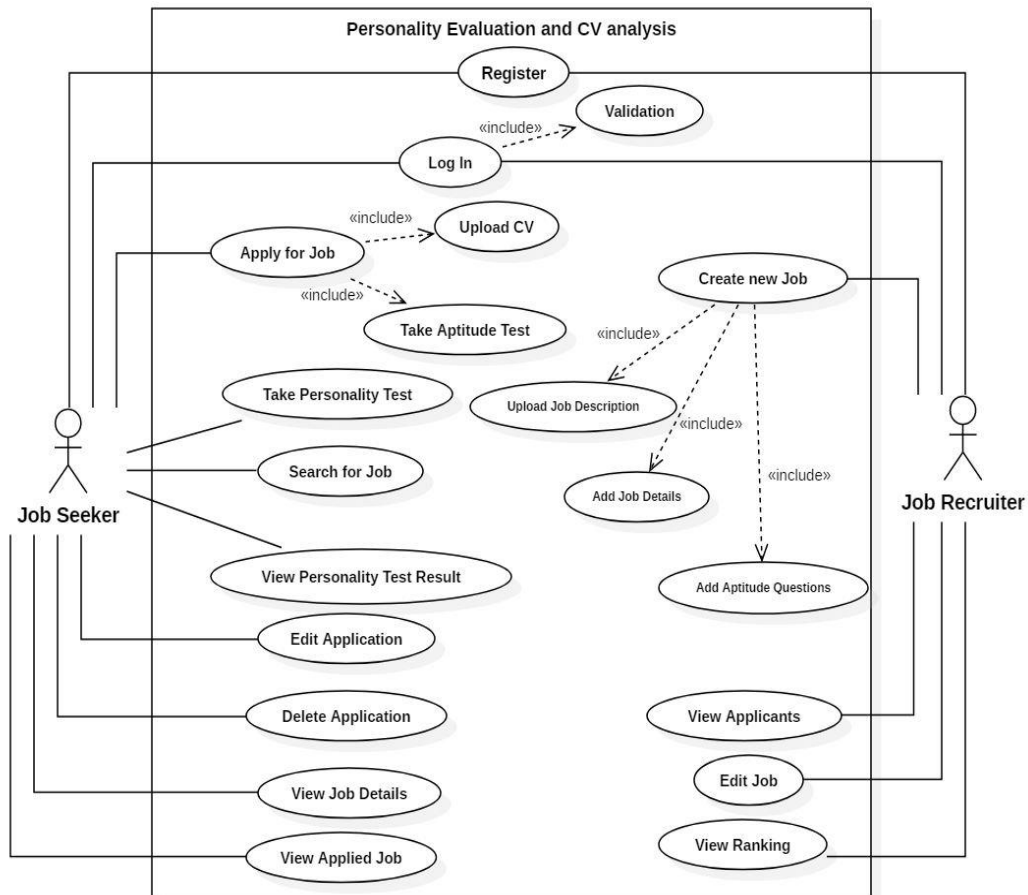
3.1.1 Use Case Diagram

Fig 3.1 Use Case Diagram

3.2 Non-Functional Requirements

Non-functional requirements include the constraints and qualities of the application. The non-functional requirements of the system are:

- **Security**

  To ensure the security of the user data, the system needs to authenticate on every user before the user can access their data in the system. This is achieved by registering the users and allowing access to only registered users via login.

- **Reliability and Accuracy**

  The system should provide accurate similarity values of CVs so that similar or related CVs to the description are ranked at the top making the system more reliable. Also, scores of personality and aptitude tests need to be accurate and reliable as well.

- **Speed**

  The speed of similarity calculation needs to be as fast as possible for better user experience. For this, the similarity values are calculated only once and stored in the database for quick retrieval rather than complex calculations.

3.3 Software/Hardware Requirements

The software requirement of the system are:
- Web browser (Chrome, Firefox, Explorer etc.)
- Django
- PostgreSQL

The hardware requirement of the system are:
- Computers supporting any web browsers (Chrome, Firefox, Explorer etc.)
- Android or iOS devices able to support any web browsers (Chrome, Firefox, Explorer etc.)

3.4 Feasibility Study

3.4.1 Technical Feasibility

The proposed system is technically feasible. We are completely capable of carrying out the project along with using the NLP approach for analyzing the documented CV using the TF-IDF algorithm. All the resource and technology required for the project is readily available.

3.4.2 Operational Feasibility

The project turns out to be operationally feasible as well as evaluating and analyzing suitable CVs is a commonly faced problem in the recruitment process of any organization or company. The system can be used to minimize the work of HR departments in these organizations by providing an efficient, time-saving and less tedious way to manage, rank, grade and select the suitable candidates on the basis of their requirements and needs helping to elevate their workforce for proper functioning of the organization.

4. SYSTEM DESIGN

4.1 Block Diagram of the system

This is the basic block diagram of the entire system. Here, main CV analysis part takes place in the CV Ranking model. Also, a personality evaluator system exist which predicts the personality of the job seekers. The system supplies output of CV Ranking model, aptitude test and personality evaluator to the job recruiter.
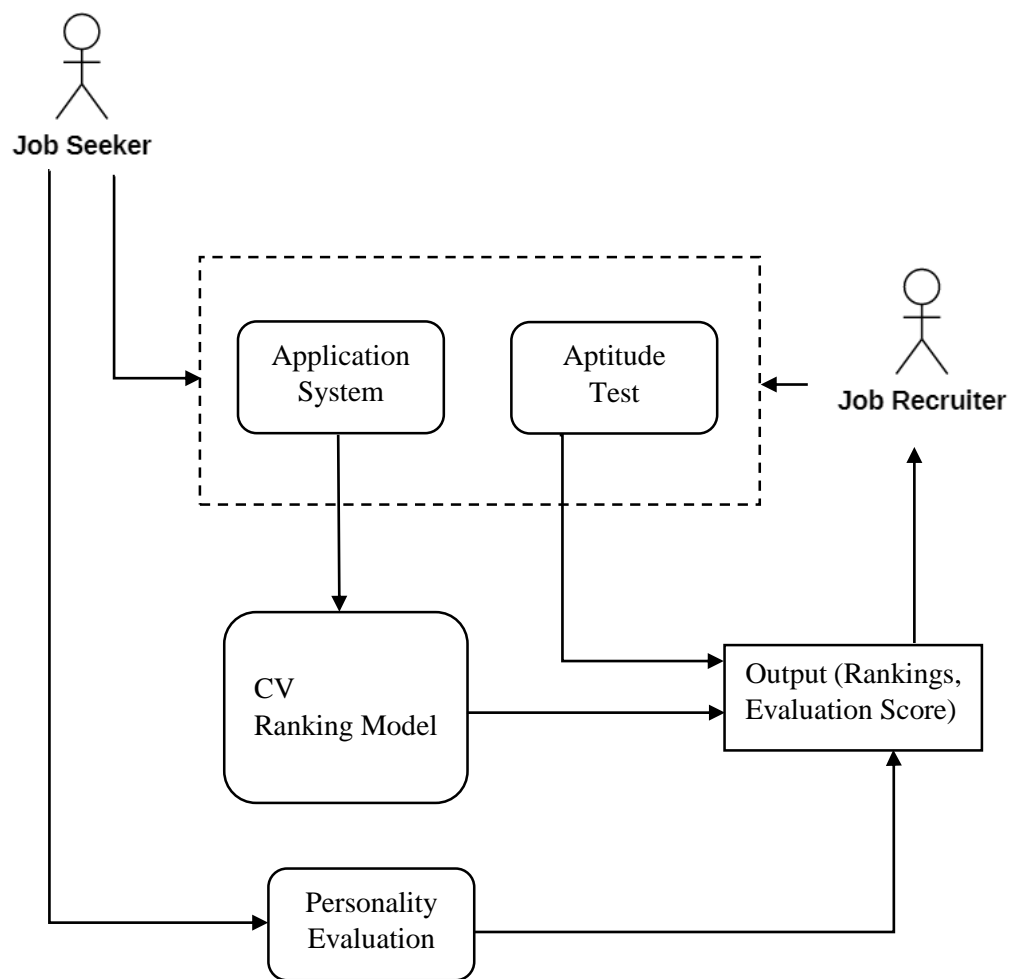
Fig 4.1 Block Diagram of the System

4.2 System Flow Diagram

The figure below represents the flow of the system. The system can be logged onto by both the employer or admin and the employee or candidate. The employer can upload a job description along with aptitude questionnaire if necessary in case of a vacancy and publish it in the system. Then, the candidates can view those vacancies and upload their own CV information and perform a test for the jobs of their interest. Also, they are required to fill up a questionnaire for personality evaluation and submit it into the system. The system then compares the CV with the job descriptions and starts to grade and rank them according to the requirements and qualifications required for the job. Finally, it provides the final ranking of the CVs in accordance to the requirements along with the results of personality test and aptitude test. This final result can be viewed by the employers to short list the candidates as per their need providing a time-saving way for the conventional interviewing system by shortening the list of candidates.
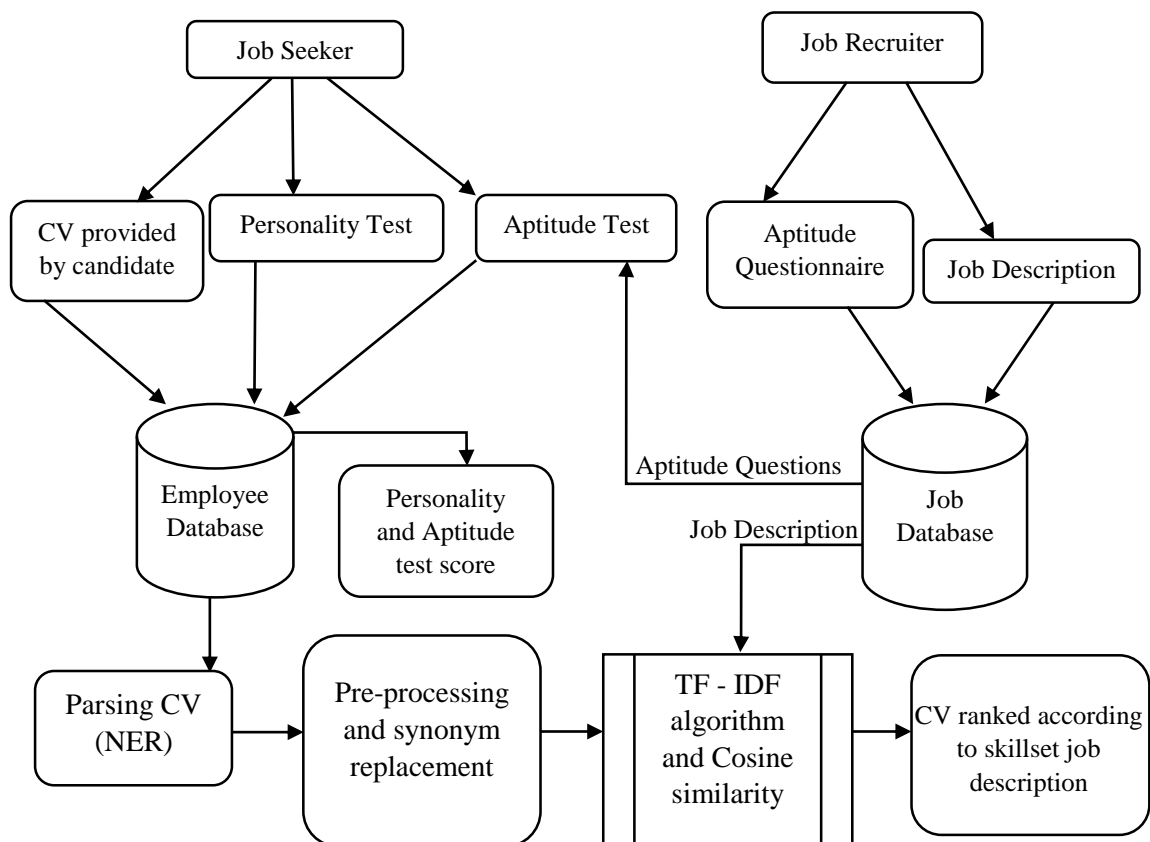
Fig 4.2 System Flow Diagram

14

4.3 ER Model
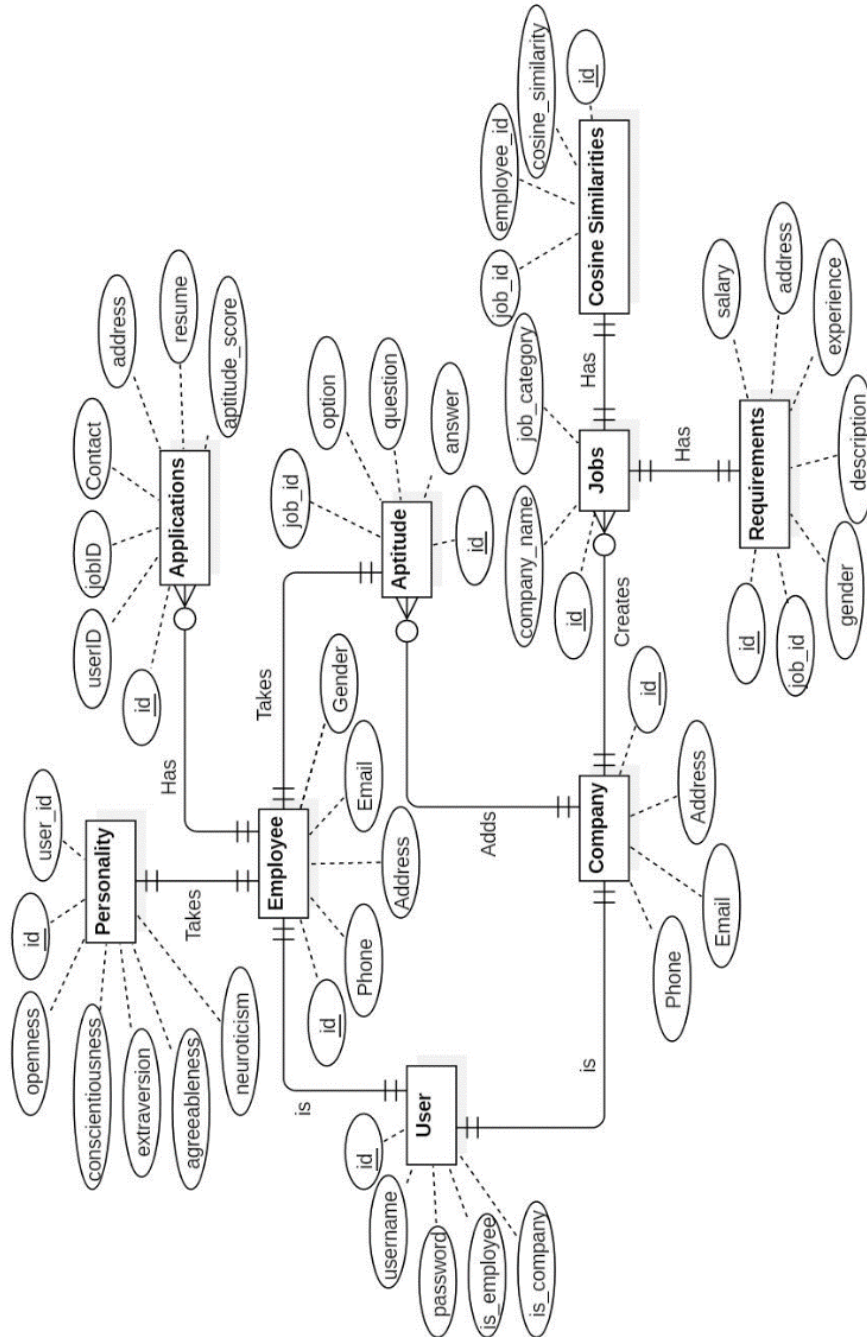
Fig 4.3 ER Diagram

4.4 Schema Diagram
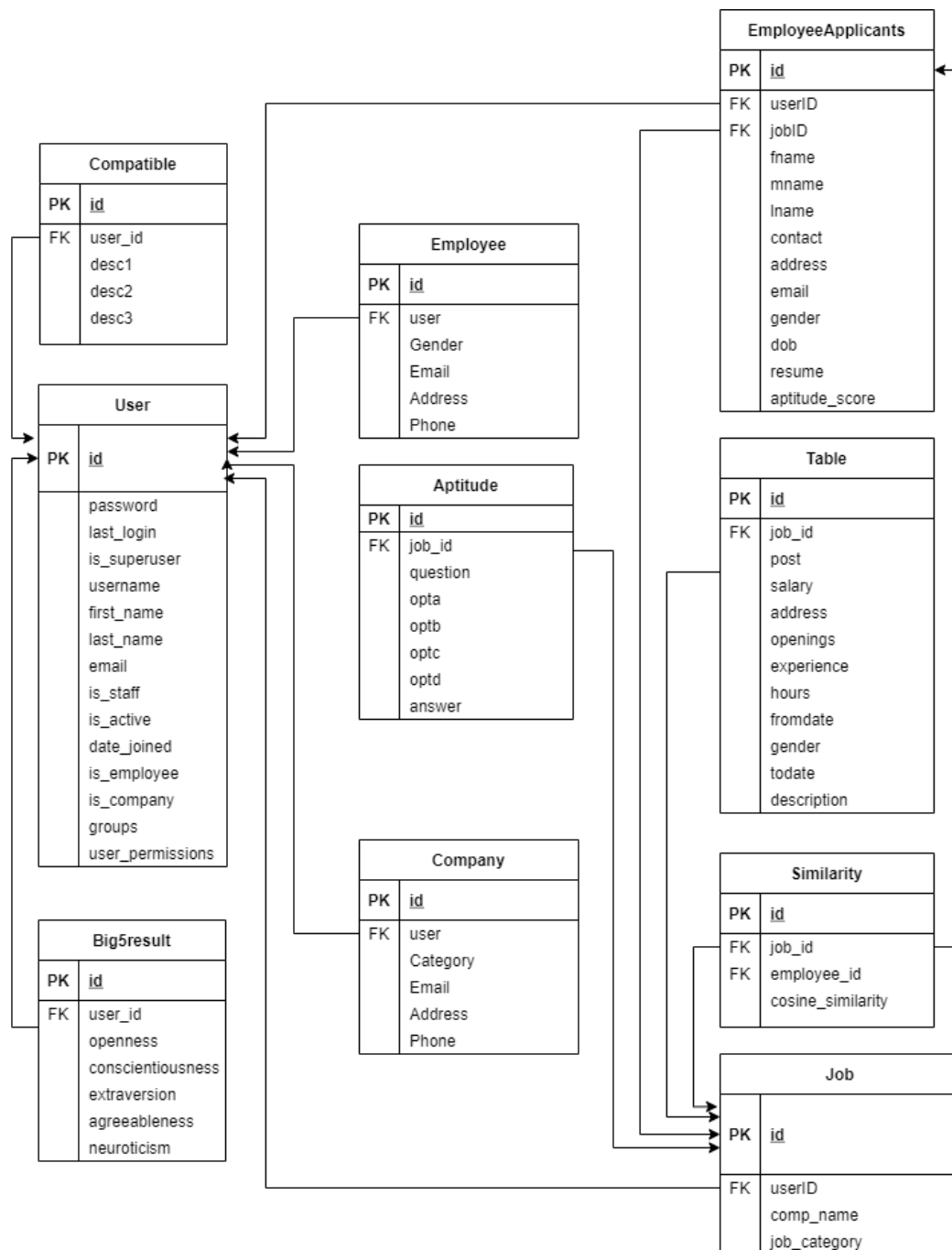
Fig 4.4 Schema Diagram

4.5 Process Model
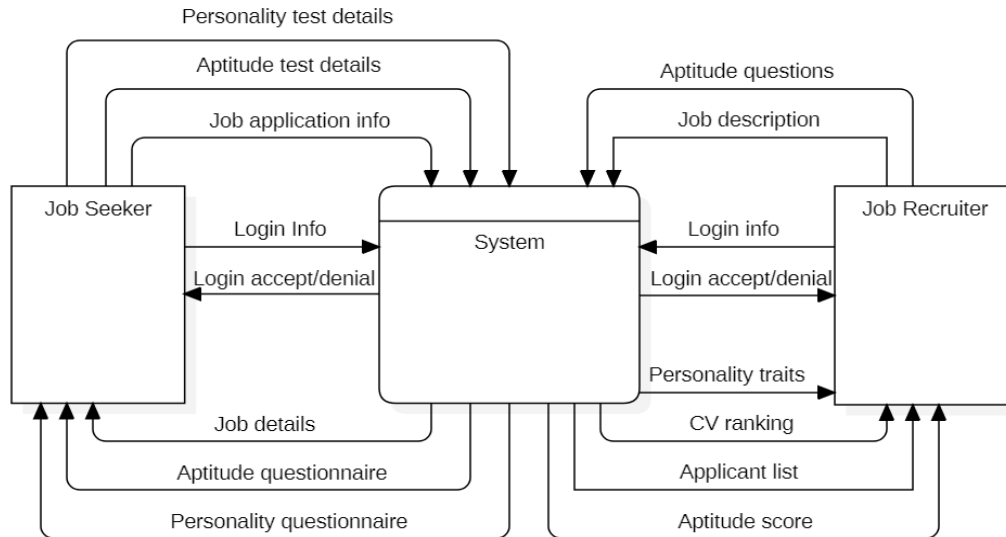
Fig 4.5 Context Diagram (DFD Level 0)
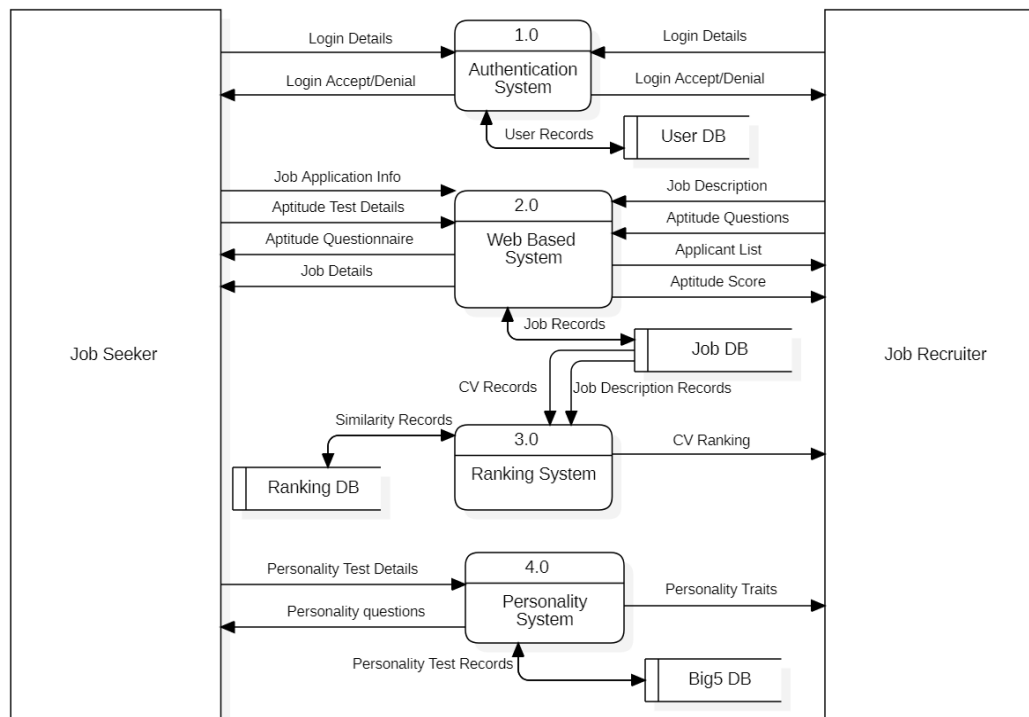
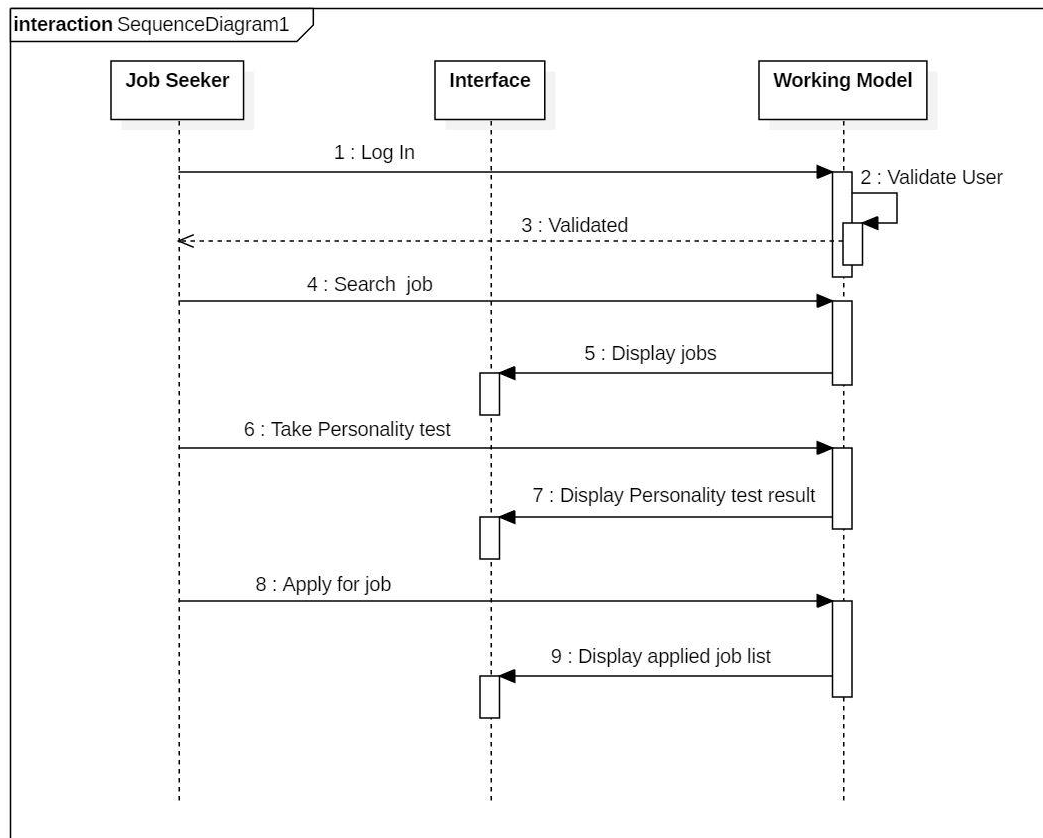*Fig 4.6 DFD Level 1*

4.6 Sequence Diagram

interaction SequenceDiagram1

| Job Seeker | Interface | Working Model |

1 : Log In

2 : Validate User

3 : Validated

4 : Search job

5 : Display jobs

6 : Take Personality test

7 : Display Personality test result

8 : Apply for job

9 : Display applied job list

Fig 4.7 Sequence Diagram for Job Seeker
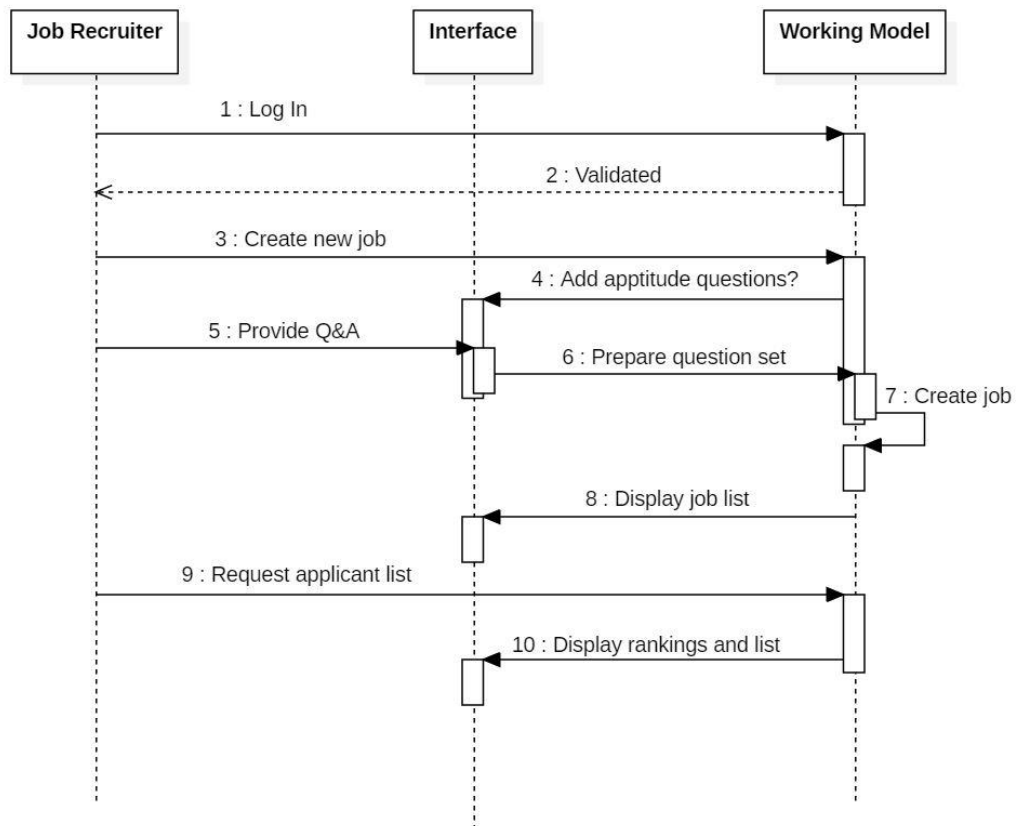
Fig 4.8 Sequence Diagram for Job Recruiter

The diagram shows three participants: **Job Recruiter**, **Interface**, and **Working Model**, with the following messages:

1 : Log In
2 : Validated
3 : Create new job
4 : Add apptitude questions?
5 : Provide Q&A
6 : Prepare question set
7 : Create job
8 : Display job list
9 : Request applicant list
10 : Display rankings and list

4.7 Activity Diagram
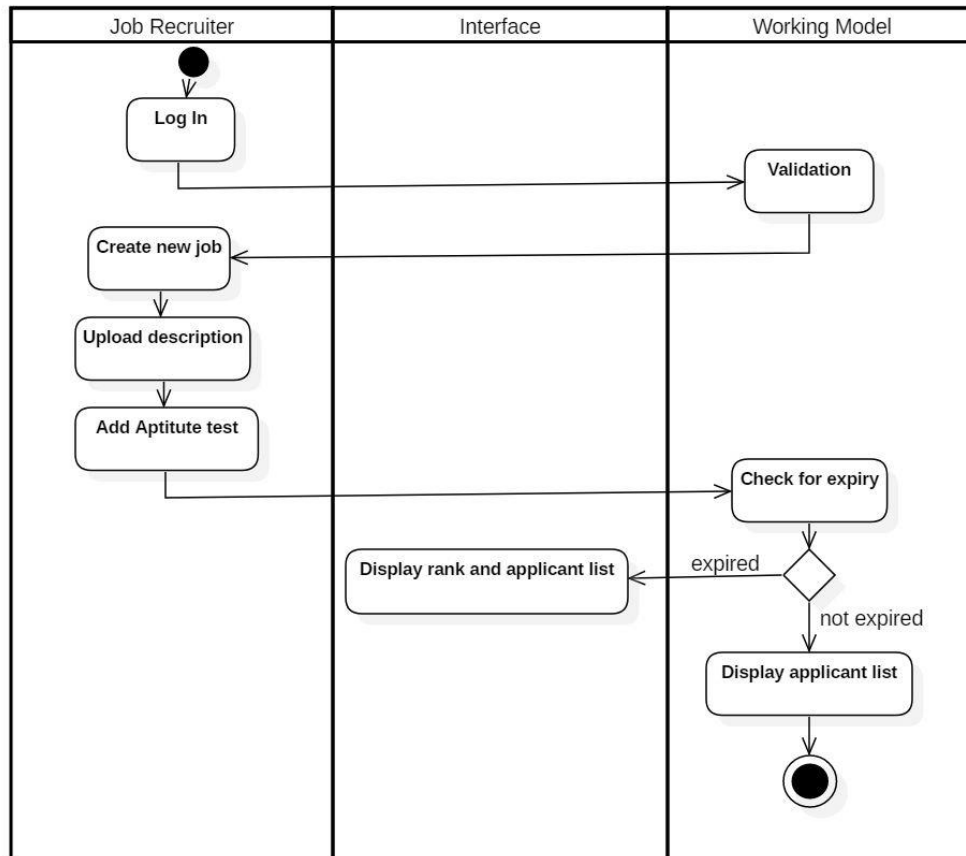
Fig 4.9 Activity Diagram for Job Recruiter

Job Seeker	Interface	Working module
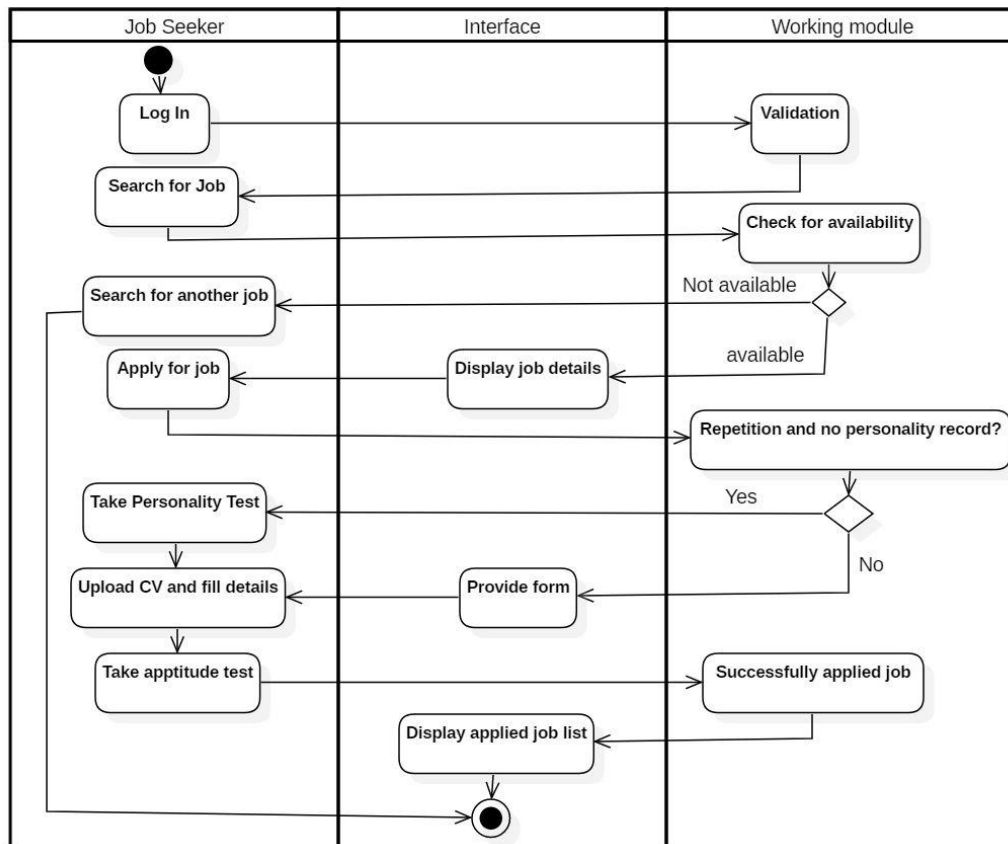
Fig 4.10 Activity Diagram for Job Seeker

21

5. METHODOLOGY

5.1 Working Principle

The system is an efficient application for grading and ranking of various candidates for a particular job according to the requirement of the employer on the basis of CV analysis and personality evaluation. The main working of the model is as shown in the diagram below.
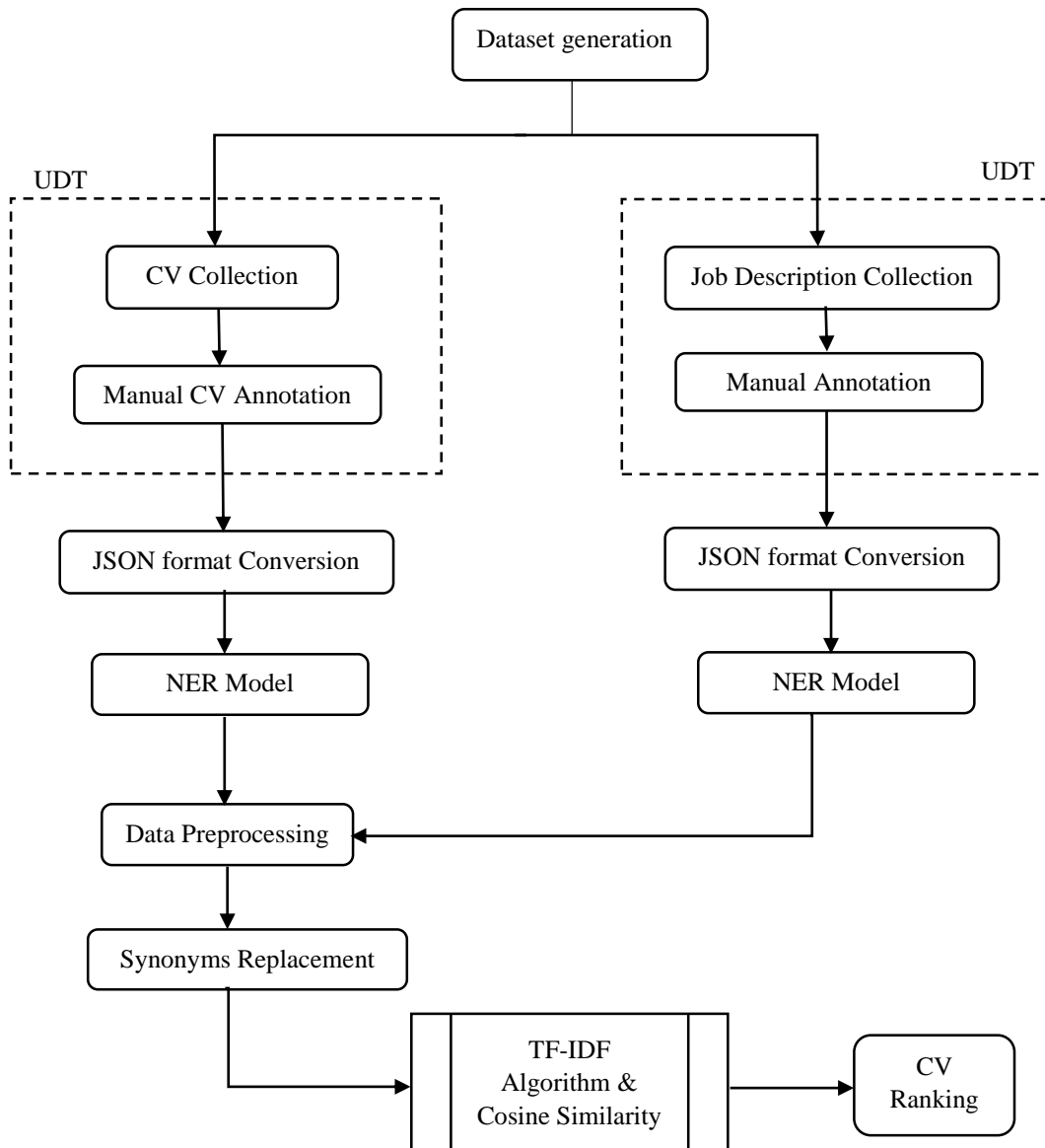
Fig 5.1 Block Diagram of CV Ranking Model

5.2 CV Ranking System

The basic flow of the CV ranking system is as given below.
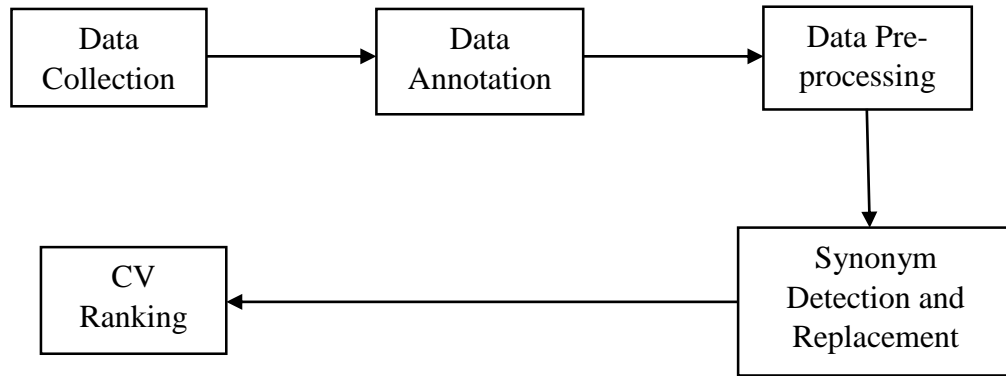
Fig 5.2 System Flow Diagram of CV Ranking

5.2.1 Data Collection

We collected different CV in different format from various sources. We collected some from:

Google Drive link:

https://drive.google.com/file/d/17M9oDPip5JFFFNJhDCBQKy8BMqyxajU/view

Kaggle Dataset link: https://www.kaggle.com/avanisiddhapura27/resume-dataset

We also received some materials with the help of our supervisor.

And, for job description, we collected our description in electronic format from websites like jobsnepal, careerbuilder and monster.

We collected standard questionnaire set for Big 5 Personality test from Big 5 Factor Markers: https://ipip.ori.org/newBigFive5broadKey.htm and selected our standard 50 questionnaire from a GIT link: https://github.com/jcl132/personality-prediction-from-text/blob/master/data/BIG5/codebook.txt

5.2.2 Data Annotation
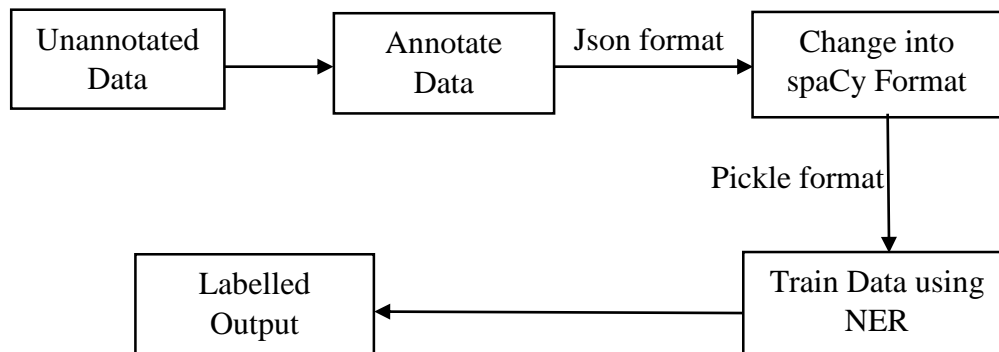
Fig 5.3 Data Annotation Block

- **Unannotated Dataset**

  Firstly, CVs and Job descriptions are collected for the purpose of analysis and training. These are raw data with no pre-processing or labelling done.

- **Annotate Data**

  The unannotated or raw data are manually annotated according to the designated labels for filtering of relevant information. This is done by the use of a data annotation tool UDT (Universal Data Tool). For analyzing our unstructured documents, entities are predefined and for annotating data we use these entities to label our documents.

  The output of this process is a labelled data in JSON format including only necessary or relevant information.

  JSON format example:

  "samples": [
  {
         "_id": "s_wm6vkf08",
         "document": "Priyanka Mehra. Certified Full Stack Web Developer.
                priyanka.mehra@gmail.com. (718) 212-6466.
                linkedin.com/in/priyankamehra…………………"

```
"annotation": {
    "entities": [
     {
      "text": "specialized skills in Java and JavaScript",
      "label": "Skills",
      "start": 140,
      "end": 181
     },
     {
      "text": "Diverse experience utilizing Java tools in business, Web, and
client server environments",
      "label": "Experience",
      "start": 216,
      "end": 304
     },
```

- **Change into spaCy format**

For the annotated data to be trained, it needs to be changed into a format accepted by the spaCy model i.e. is the pickle format which is also the output of this block.

spaCy format example:

```
[
    ('Specialized skills in Java and JavaScript', {'entities': [(23, 26, 'Skills'), (28, 37,
    'Skills')]})
]
```

- **Train data using NER**

The final dataset in pickle format is then used for training with the help of NER (Named Entity Recognition) feature of spaCy to get a trained model with the ability to label and filter relevant information from additional CVs and descriptions.

Every "decision" made – for example, which part-of-speech tag to assign, or whether a word is a named entity – is a prediction based on the model's current weight values. The weight values are estimated based on examples the model has seen during training. To train a model, we first need training

data – examples of text, and the labels we want the model to predict. This could be a part-of-speech tag, a named entity or any other information.
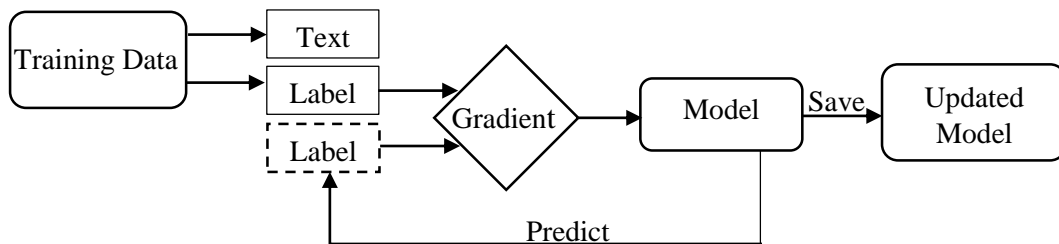
Fig 5.4 NER Training

Training is an iterative process in which the model's predictions are compared against the reference annotations in order to estimate the gradient of the loss. The gradient of the loss is then used to calculate the gradient of the weights through backpropagation. The gradients indicate how the weight values should be changed so that the model's predictions become more similar to the reference labels over time.

When training a model, we don't just want it to memorize our examples – we want it to come up with a theory that can be generalized across unseen data.

- **Labelled Output**

After the model is trained, we are able to get labelled outputs to any CV or description fed into it.

For CV,

```
"labels": [
    {
     "id": "Designation",
     "displayName": "Designation"
    },
    {
     "id": "Responsibilities",
     "displayName": "Responsibilities"
    },
    {
     "id": "Qualifications",
     "displayName": "Qualifications"
    } ]
```

For Job Description,

```
"labels": [
{
"id": "Skills",
"displayName": "Skills",
"description": ""
},
{
"id": "Experience",
"displayName": "Experience",
"description": "Something worn on the head."
},
{
"id": "Qualification"
},
{
"id": "Projects"
}]
```


Fig 5.5 Data Annotation (NER) of CV using universaldatatool

```
EXPERIENCE                          - Part of team leading the IPO of Worldpay, a multi-billion-pound global FinTech c
nt  Global Coordinator; drafted Analyst Presentation and prospectus and managed analyst Q&A  2014  THE BOSTON CON
Milan, Italy    Senior Associate
EXPERIENCE                          - Consumer Goods and Financial Institutions Practice  •
SKILLS                              - Assessed an Italian bank's CIB division positioning through wallet-sizing analys
ent  interviews; led cross-functional team in portfolio strategy and coverage model redesign  •
EXPERIENCE                          - Built 3-year strategy and investments plan for a global retail company; built an
iness  cases for entry into new markets and facilitated resource allocation negotiations across projects  2012  A
EXPERIENCE                          - Built an Excel-based scheduling tool to optimize labor cost for a global travel
upervised  roll-out in over 300 stores, resulting in 8% labor cost reduction from previous year  •
EXPERIENCE                          - Analyzed distributed energy and energy efficiency market in EU countries for a p
identifying €200M opportunity in underserved hotel segment in Italy and France  2011  Visiting Associate  •
EXPERIENCE                          - Developed extensive comparative study on healthcare system in 5 EU countries, wh
cited in the press and used by BCG to sell 3 additional projects that earned over €1M  OTHER BUSINESS EXPERIENCE
PS)  2016  2R CAPITAL, London, UK  Value fund, European Equities (launch phase
EXPERIENCE                          - Worked on the setup of the fund, defining strategy and investment philosophy and
searched and presented investment ideas, applying fundamental analysis to identify undervalued  companies based o
eption vs. market  2015  C5 CAPITAL, London, UK    Specialist technology Private Equity fund, focused on cybersec
data
EXPERIENCE                          - Supported investment team in origination
EXPERIENCE                          - Bocconi Outstanding Scholar
QUALIFICATION                       - Bachelor of Business Administration and Management  •
QUALIFICATION                       - Mentor at Mentors4U (coaching 3 students interested in a career in management
SKILLS                              - Fluent English
```

Fig 5.6 NER test result

Fig 5.7 Data Annotation (NER) of job description using universaldatatool

```
on  of resources, staff performance, annual appraisals, development and mentoring
RESPONSIBILITIES                - To attend to clients' requests including distributions, asset purchase or disposal, amen
he  Trust Deed and updating Letters of Wishes
RESPONSIBILITIES                - To liaise with bankers, investment managers, agents and other professionals as necessary
e  administration of client entities
RESPONSIBILITIES                - To liaise with internal departments as necessary in the administration of client entitie
as  compliance, business development and finance
RESPONSIBILITIES                - ensure proper records are maintained and controls adhered to
RESPONSIBILITIES                - To observe and reinforce internal control policies and procedures and statutory guidelin
applicable to the role
RESPONSIBILITIES                - To ensure that the database is accurate and kept up to date
RESPONSIBILITIES                - To ensure that file notes raised in the system are cleared by the respective administrat
RESPONSIBILITIES                - To check annual financial accounts
RESPONSIBILITIES                - To check annual trust reviews
RESPONSIBILITIES                - To review bank advices to identify any irregular transactions and/or any injections of a
RESPONSIBILITIES                - To attend to queries in relation to trust accounting
RESPONSIBILITIES                - To manage safe custody items and maintain register
RESPONSIBILITIES                - To check and approve reimbursement of disbursements
RESPONSIBILITIES                - To check client invoices for accuracy monitor outstanding invoices
RESPONSIBILITIES                - To send out fee invoices and collect fees
RESPONSIBILITIES                - To meet chargeable time requirements
RESPONSIBILITIES                - To assist Trident with meetings its regulatory requirements, including but not limited t
k
RESPONSIBILITIES                - To either comply with any CPD requirements of your professional body
RESPONSIBILITIES                - To observe and comply with Trident's internal policies and procedures
RESPONSIBILITIES                - To participate in training and tests provided by Trident pursuant to Trident's regulator
gations  to train and test its employees and to demonstrate competency in the topics tested
RESPONSIBILITIES                - Anything else reasonably required by Trident in the administration of client entities an
maintenance of Trident's licences and regulatory compliance
QUALIFICATIONS                  - minimum of 5 years' experience in a trust environment
QUALIFICATIONS                  - High degree of professional ethics and integrity
QUALIFICATIONS                  - Ability to establish and maintain good relationships with clients and colleagues
```

Fig 5.8 NER test result for job description

5.2.3 Data Pre-processing

Data preprocessing is an essential step in building a Machine Learning model and depending on how well the data has been preprocessed; the results are seen. Text preprocessing steps are widely used for dimensionality reduction. In the vector space model, each word/term is an axis/dimension. The number of unique words means the number of dimensions.

Tokenization: Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens. The tokens could be words, numbers or punctuation marks.
Input: "SLC from Nobel Academy"
Output: ['SLC', 'from', 'Nobel', 'Academy']

Lower casing: Converting a word to lower case (NLP -> nlp).

Output: ['slc', 'from', 'nobel', 'academy']

29

Stop words removal: Stop words are very commonly used words (a, an, the, etc.) in the documents. These words do not really signify any importance as they do not help in distinguishing two documents.

Input: "Machine learning is cool"

Output: ['Machine', 'Learning', 'cool', '!']

Explanation: Stop word 'is' has been removed

Removing punctuation: In python, we remove punctuation characters using translate() method. This method makes a copy of a string with a specific set of values substituted.

text.translate(str.maketrans(", ", string.punctuation)

Lemmatization: Lemmatization reduces the words to a word existing in the language. For lemmatization to resolve a word to its lemma, part of speech of the word is required. This helps in transforming the word into a proper root form.

(change, changing, changes) -> change

5.2.4 Synonym Detection and Replacement

As the TF-IDF is unable to match words with the same meaning and count them as one, the accuracy of the result is greatly affected. This also creates a vector larger than required as synonyms are treated as different words. So, synonyms should be detected and replaced beforehand so this problem could be avoided and accuracy could be increased. For this, WordNet is used.

The WordNet is a part of Python's Natural Language Toolkit. It is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. WordNet's structure makes it a useful tool for computational linguistics and natural language processing. WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings.

```
In [6]: modified_arr=[['poor', 'boat'], ['good', 'misfortunate', 'fruit'],['bad', 'best']]

In [7]: skip=[]
        for d in modified_arr:
            for i in range(len(d)):
                synonyms=[]
                for syn in wn.synsets(d[i]):
                    for l in syn.lemmas():
                        synonyms.append(l.name())

                for doc in modified_arr:
                    for j in range(len(doc)):
                        if doc[j] not in skip:
                            if doc[j] in synonyms:
                                if doc[j]!=d[i]:
                                    doc[j]=d[i]
                                    skip.append(doc[j])

In [8]: print(modified_arr)

        [['poor', 'boat'], ['best', 'poor', 'fruit'], ['bad', 'best']]
```

Fig 5.9 WordNet synonym replacement example

As can be seen in the figure, the word list of each document is scanned to find the similar words and replace them. The word good and best are recognized as synonyms so are replaced by a single and common word i.e. best. The replaced words are than skipped during the scan so as to avoid double replacement.

Thus, by the use of WordNet Synonyms of the words were detected and replaced by a common and single word to reduce the vector size of the documents, solve the problem of not being able to detect and distinguish synonyms and improve accuracy of the TF-IDF result.
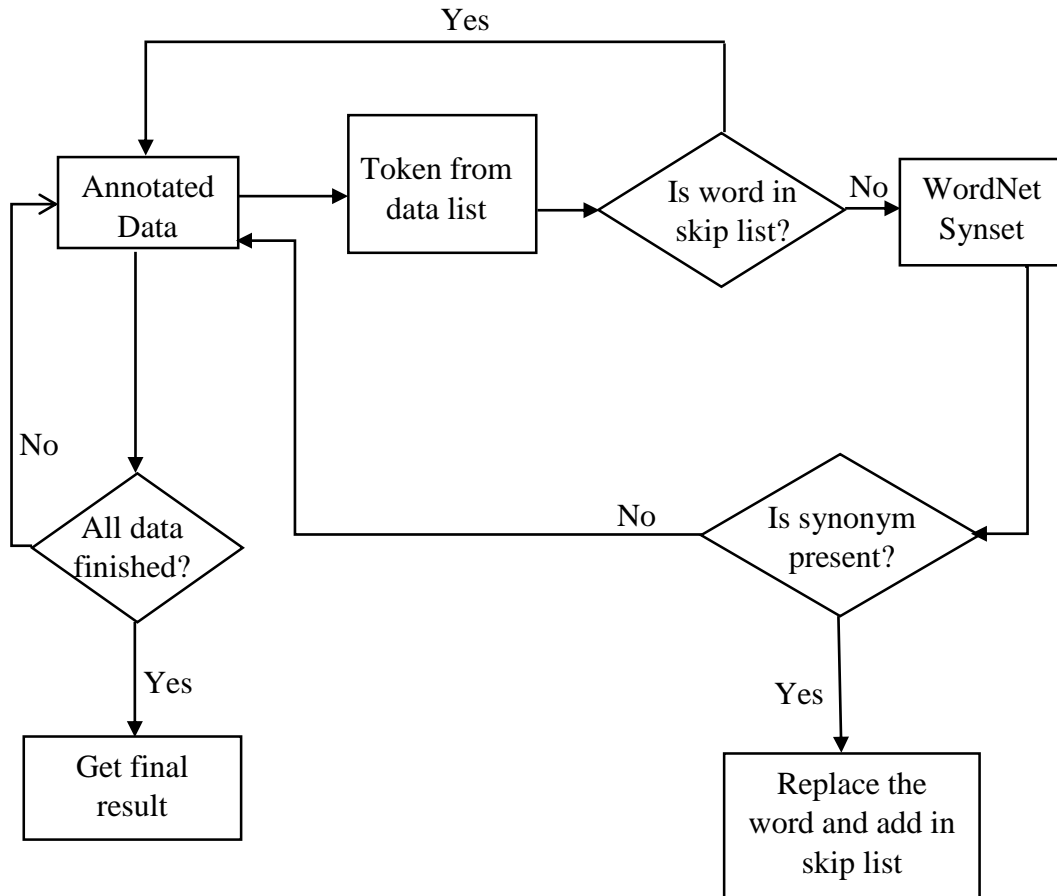
Fig 5.10 Synonym Replacement using WordNet

5.2.5 CV ranking using TF-IDF Algorithm

After applying WordNet to detect and replace synonyms we can move on to calculating the TF-IDF weight of the documents and finally achieving the similarity score using cosine similarity. TF-IDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. Cosine similarity is a metric used to determine how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. In this context, the two vectors I am talking about are arrays containing the word counts of two documents.
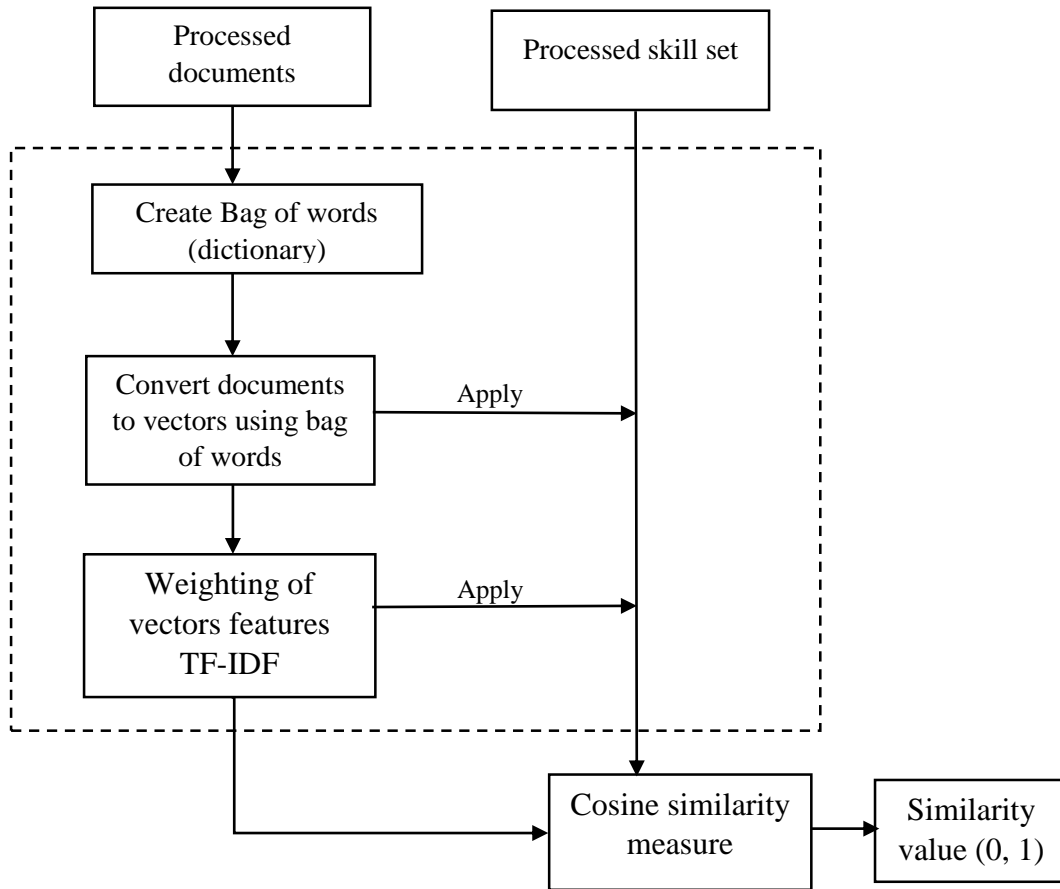
Fig 5.11 TF-IDF Ranking Algorithm

The TF-IDF weight can be calculated as:

Step 1:

$$TF\ (\text{'keyword'}) = \frac{Number\ of\ times\ \text{'keyword'}\ appears\ in\ document}{Total\ number\ of\ keywords\ in\ the\ document.} \qquad \textbf{\textit{Equation 5.1}}$$

Step 2:

$$IDF\ (\text{'keyword'}) = log\left(\frac{Total\ number\ of\ CV}{Number\ of\ document\ with\ term\ \text{'keyword'}}\right) \qquad \textbf{\textit{Equation 5.2}}$$

Step 3:

$$Weight = \ TF('keyword') * IDF('keyword') \qquad \textbf{\textit{Equation 5.3}}$$

Cosine Similarity

Cosine similarity is a metric used to determine how similar the documents are irrespective of their size. Cosine similarity is calculated as follows,
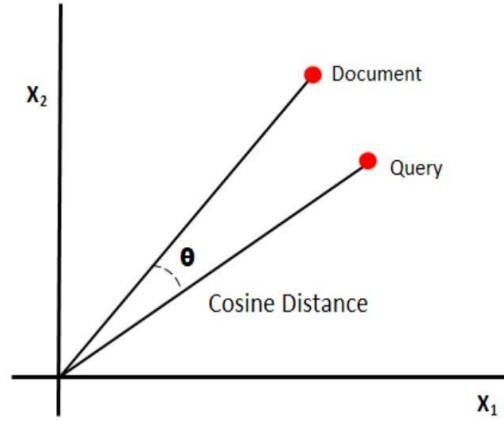

Fig 5.12 Vector plot of tf-idf tokens

Cosine similarity can be achieved as:

$$\text{Similarity} = \cos(\Theta) = \frac{\boldsymbol{Doc1} \cdot \boldsymbol{Doc2}}{|\boldsymbol{Doc1}| \, |\boldsymbol{Doc2}|} \qquad \qquad \textit{Equation 5.4}$$

Where, Doc1 and Doc2 are the TF-IDF weight vectors of the documents.

Cosine similarity measures the similarity between two vectors of an inner product space. The greater the value of θ, the less the value of cos θ, thus the less the similarity between two documents.

```
In [47]: length = len(ndocs) - 1
         for i in range(length+1):
             cosine = cosine_similarity(tf_idf[length], tf_idf[i])
             print(cosine)
[[0.14904657]]
[[0.15063282]]
[[0.05616422]]
[[0.17997112]]
[[0.22725202]]
[[0.13282125]]
[[0.12771337]]
[[0.07791326]]
[[0.02517972]]
[[0.03252057]]
[[0.04254702]]
[[0.08686569]]
[[0.11849896]]
[[0.0586095]]
[[0.07444243]]
[[1.]]
```

Fig 5.13 Cosine Similarity Output

5.3 Personality Test System

For personality evaluation we used BIG 5 or OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) personality test. The test was performed by providing a standard set of 50 questions (10 questions for each trait), both positive and negative keyed questions, to the job seekers to which the answers ranged from strongly agree(5) to strongly disagree(1).
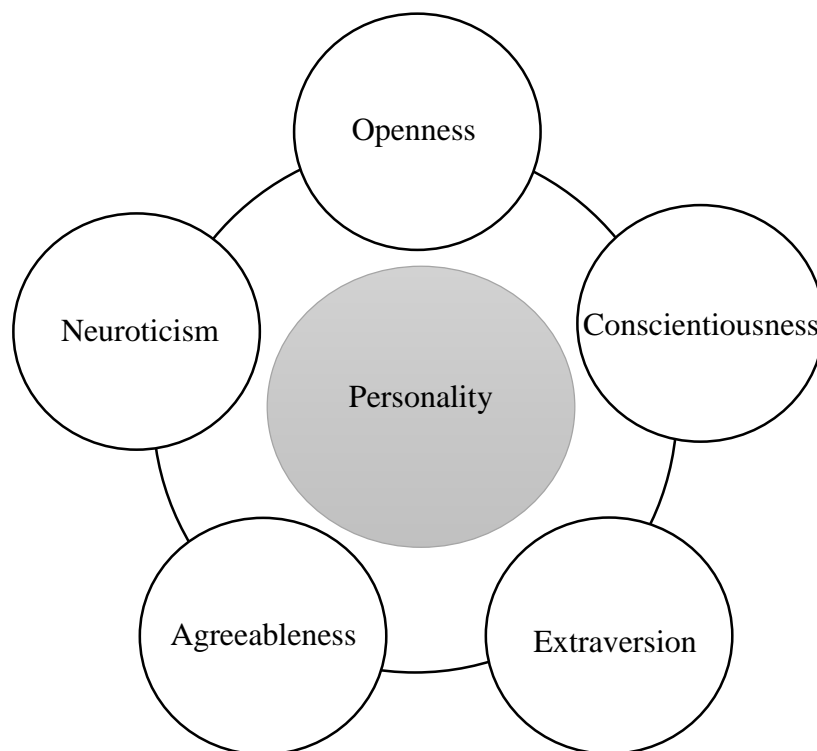

Fig 5.14 Big 5 OCEAN traits

Negative keyed questions

For example, the factor Extraversion describes someone who is outgoing, energetic, talkative, and enjoys human interaction. The first Extraversion item [EXT1] is "I am the life of the party." a positively-keyed item; whereas the second item [EXT2] is "I don't talk a lot." a negatively-keyed item.

Reverse Coding

The value for the answers were recorded and reverse-coded by subtracting 6 from the value of answers for all negative keyed questions. This results in a dataset where the item values all have a common direction and interpretation (i.e., a higher value corresponds with more of that trait).
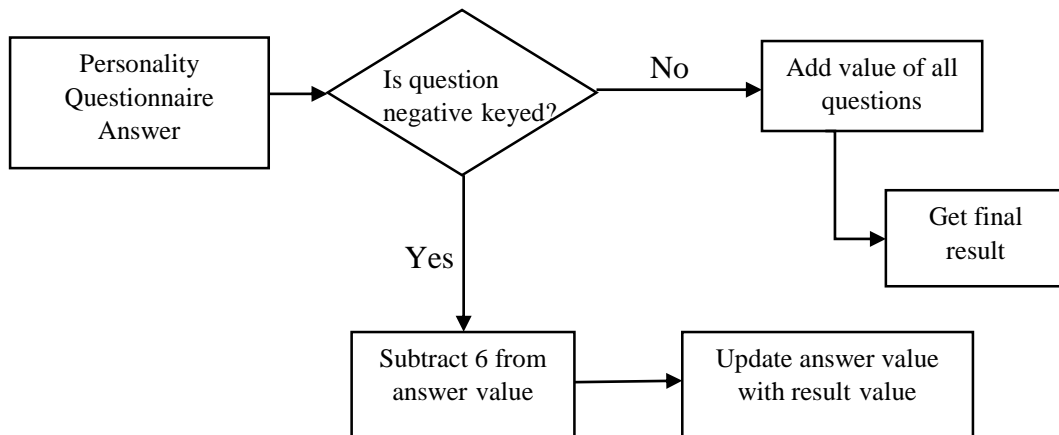

Fig 5.15 System Flow of Big 5 Personality Test

Finally, the resultant value were added together and percentage value was calculated to determine what traits exist in what amount in the individuals.

5.4 Tools Used

PostgreSQL

It is a free and open-source relational database management system (RDBMS) emphasizing extensibility and SQL compliance. It was used for fulfilling the database requirements of the system.

PyCharm

PyCharm is an integrated development environment (IDE) used in computer programming, specifically for the Python language. It was used for coding, editing and debugging of the system.

UDT (Universal Data Tool)

The Universal Data Tool is a web/desktop app for editing and annotating images, text, audio, documents and to view and edit any data defined in the extensible .udt.json and .udt.csv standard. It was used to manually annotate and label the CVs and descriptions and to convert them into json format.

NLTK (Natural Language Tool Kit) library

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. It was used for pre-processing functions such as tokenization, lemmatization and stop words removal.

spaCy

It is an open-source software library for advanced natural language processing, written in the programming languages Python and Cython. It was used for processing the annotated data by the use of NER (Named Entity Recognition).

WordNet

WordNet is a lexical database of semantic relations between words in more than 200 languages. It was used to detect and replace synonyms to increase the effectiveness of TF-IDF algorithm.

6. SYSTEM IMPLEMENTATION AND TESTING

The final system was completely tested for any faults and errors and the final result was analyzed as well. The UI of the system was interactive for better user experience. The working was also presented in an easy to understand way to the client so it can be used by anyone. All the system modules were giving the desired results individually as well as in an integrated manner. The web front-end displayed all job and applicant related details along with the ranking of CVs based on the calculated similarity. Likewise, the back-end system handled the task of efficiently managing the different databases and also the accurate calculation of similarity between CVs and job description. So, a fully functional and tested system was achieved as the result.

6.1 Accuracy

The accuracy was measured for the proper annotation of data into its various labels for both the CVs and descriptions. For this, we compared the manually annotated and model annotated output for the same set of data to test how accurately the data was annotated. The result of comparison and testing can be seen in the tables below:

Entity	Precision	Recall	F-Score	Accuracy
Designation	0.9976	0.9953	0.9960	99.60%
Responsibility	0.9452	0.9527	0.9476	95.27%
Qualification	1.0	0.9677	0.9836	96.77%

Table 6.1 Accuracy for Job Description Annotation

Entity	Precision	Recall	F-Score	Accuracy
Skills	0.8138	0.7971	0.7327	79.71%
Experience	0.8073	0.7626	0.7814	76.26%
Qualification	0.9067	0.9164	0.8957	91.64%
Project	0.6663	0.7374	0.6870	73.74%

Table 6.2 Accuracy for CV Annotation

6.2 Testing

The system was tested at different levels to assure its reliability, accuracy and performance. The various tests performed were as:

6.2.1 Unit testing

Unit testing refers to the process of testing modules against the detailed design. For our system, unit test was done by independently testing the different modules i.e. login, job application, job requirements, aptitude test, personality test and CV analysis to ensure proper working.

Registration

The registration module should take the necessary details of the user to be added to the database under user table.

ID	TC01
TITLE	Register
PREREQUISITE	Username Contact Password Email
TEST ACTION	Enter details for registration and save to database.
EXPECTED RESULT	The user details should be added to the database in the user table.
RESULT	New users are successfully registered.

Table 6.3 Test Case for Registration

Log In

The login module should provide access to only registered users to their account whereas the unregistered users should be asked to register as a user first.

ID	TC02
TITLE	Log In
PREREQUISITE	Registered Username and Password
TEST ACTION	Log in with registered username and password
EXPECTED RESULT	Should check whether the given username and password matches the registered ones and based on the result provide or deny access to further features for the user.
RESULT	Registered users are logged in to the system successfully.

Table 6.4 Test Case for Log In

Personality test

This module should ensure that personality test is taken only once with no overwriting of the previous test scores.

ID	TC03
TITLE	Personality Test
PREREQUISITE	Personality Questionnaire
TEST ACTION	Take personality test
EXPECTED RESULT	The personality test results should be calculated and stored. Also, the personality form should not be displayed again after the results are shown.
RESULT	The personality test should not be retaken.

Table 6.5 Test Case for Personality Test

CV Ranking

This module activates only when the job has expired, and performs the analysis and ranking functions of the system.

ID	TC04
TITLE	CV Ranking
PREREQUISITE	Expired Job with Applicant List CVs matching the Job Requirement CVs not matching the Job Requirement
TEST ACTION	Rank CVs for expired jobs.
EXPECTED RESULT	During ranking, CVs matching the job requirement are ranked at the top whereas the CVs not matching the job requirement are ranked at the bottom.
RESULT	The ranking starts only after the job has expired and proper ranking is displayed where CVs not matching the requirement are ranked at the bottom.

Table 6.6 Test Case for CV Ranking

6.1.2 Integration Testing

For integration testing each individual software modules was combined and tested as a group. Integration testing was conducted to evaluate the compliance of our system or component with specified functional requirements. It was done after unit testing, taking the various software modules that have been unit tested as input, grouped them in larger aggregates, applied various tests to those aggregates, and delivered its output, the integrated system, ready for system testing. The application modules passed the integration test and were moved on for system testing. We integrated NER model and the TF-IDF model, where the summarized output from NER model after some preprocessing was passed as input to TF-IDF model for further dimensionality reduction.

6.1.3 System Testing

System testing was conducted on the complete integrated system to evaluate the system's compliance with its specified requirements. System testing took, as its input, all of the integrated components that have passed integration testing. For system testing, we manually annotated the test data and compared it with the output of the data gained from the model. Hence, we got the accuracy for each entities used in annotation. The system test was a success as the system performance as a whole was as the expected results.

7. CONCLUSION

7.1 Conclusion

Manually sorting through hundreds of CVs during the recruitment process is a laborious task that requires great effort and time. This project aimed to develop a much better alternative in the form of a CV grading system based on the similarity between the CVs and job description. This result was obtained by implementing different algorithms and methods which consequently generated the desired output. So, a platform for both job seekers and recruiters was created with added functionality of CV analysis and grading.

Enormous knowledge was gained through the project. Discovering and solving various predicted and unpredicted problems as a team, we learnt a lot on the subject of NLP processing and also learned the use and implementation of various other tools used while development of project. We understood the importance of background research as well. Thus, we hope that the system developed will certainly assist in the field of recruitment by saving both time and effort spent on sorting the CVs and provide a new advanced method for better performance.

7.2 Limitations

Even though the project meets most of our set goals, it has some limitations to improve upon.
- The system cannot take in scanned images as input for CV/Description document.

7.3 Future Enhancements

Due to time constraint, many features couldn't be incorporated in the project. The system can be upgraded in many aspects such as:

- Providing a feature for comparison of CVs so as to modify and improve own CV for better chances at landing the job.
- Feature of automatically detecting all the required information by just uploading the CV with no need of entering other additional information.

REFERENCES

[1]   Indeed Career Guide, "What Is a CV? Curriculum Vitae Definition and Examples," 24 November 2020. [Online]. Available: https://www.indeed.com/career-advice/resumes-cover-letters/what-is-a-cv. [Accessed 28 November 2020].

[2]   WikiJob, "What Is a Job Description?," 16 October 2020. [Online]. Available: https://www.wikijob.co.uk/content/application-advice/job-applications/what-job-description. [Accessed 1 November 2020].

[3]   resumeprofesionalwriters, "Matching Your Resume with the Job Description to Ace Your Job Search," 8 June 2020. [Online]. Available: https://www.resumeprofessionalwriters.com/resume-vs-job-description-comparison/. [Accessed 17 September 2020].

[4]   [Online]. Available: https://www.resumeprofessionalwriters.com/. [Accessed 8 December 2020].

[5]   "What Is CV/Resume Parsing?," Daxtra Technologies Ltd, 18 October 2016. [Online]. Available: www.daxtra.com/2016/10/18/what-is-cvresume-parsing/. [Accessed january 2020].

[6]   C. Ratcliff, "What is semantic search and why does it matter?," ClickZ Group Limited, 21 October 2015. [Online]. Available: https://www.searchenginewatch.com/2015/10/21/what-is-semantic-search-and-why-does-it-matter/. [Accessed 5 January 2020].

[7]   "Is Your Resume Ready for Automated Screening?," 2 January 2016. [Online]. Available: www.resumehacking.com/ready-for-automated-resume-screening. [Accessed 15 January 2020].

[8]   P. Nelson, "Natural Language Processing (NLP) Techniques for Extracting Information," Search Technologies, [Online]. Available:

https://www.searchtechnologies.com/blog/natural-language-processing-techniques.
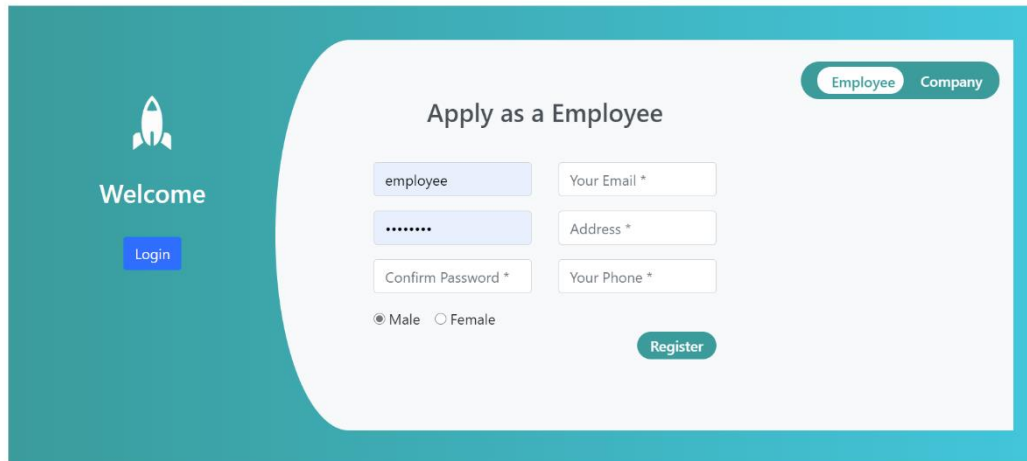
[9]   B. Reynolds, "The Terrible Trouble with Natural Language Processing (It's Us.)," Salesforce.com, Inc., 17 August 2017. [Online]. Available: www.salesforce.com/blog/2016/08/trouble-with-natural-language-processing.html. [Accessed January 2020].

[10]   a. Narayanan, "Automatic Summarization of Resumes with NER," DataTurks, 14 July 2018. [Online]. Available: https://medium.com/@dataturks/automatic-summarization-of-resumes-with-ner-8b97a5f562b. [Accessed 22 August 2020].

[11]   S. Sanyal, "Resume Parser with Natural Language Processing," March 2017. [Online]. Available: https://www.researchgate.net/publication/313851778_Resume_Parser_with__Natural_Language_Processing. [Accessed December 2019].

[12]   F. Patankar, "Compare To See How Your Resume Ranks Against Any Job Description with Jobscan," [Online]. Available: https://www.careergeekblog.com/jobscan-compare-resume/. [Accessed 17 January 2020].

[13]   J. Feldman, "Get to Know the 5 Most Popular Pre-Employment Personality Tests," 20 February 2019. [Online]. Available: https://www.topresume.com/career-advice/how-to-pass-the-pre-employment-personality-test. [Accessed 13 May 2020].

[14]   K. Cherry, "The Big Five Personality Traits," 13 July 2020. [Online]. Available: https://www.verywellmind.com/the-big-five-personality-dimensions-2795422#:~:text=The%20five%20broad%20personality%20traits,openness%2C%20conscientiousness%2C%20and%20neuroticism.&text=As%20a%

20result%2C%20the%20five,the%20building%20blocks%20of%20persona l. [Accessed 15 August 2020].

[15] Open-Source Psychometrics Project, "Big Five Personality Test," 2 August 2019. [Online]. Available: https://openpsychometrics.org/tests/IPIP-BFFM/. [Accessed 27 March 2020].

APPENDICES

Snapshots

Fig 9.1 Signup Page

Fig 9.2 Login Page

Fig 9.3 Home Page

Fig 9.4 Recent Jobs

Fig 9.5 Job Details Page

Fig 9.6 Personality test

Fig 9.7 Forms

Fig 9.8 Aptitude test form

Fig 9.9 Aptitude test

Fig 9.10 Company Dashboard

3	Amit Musale	9843765778	amit@gmail.coom	Lazimpat	Male	08 Sep, 1850	0	View Resume
4	Supriya Yadav	9860923051	supriya@gmail.com	Sanogaucharan	Female	08 Sep, 1998	3	View Resume
5	Shashank Vishwakarma	9804706675	shashank@gmail.com	Lalitpur	Male	08 Sep, 1990	1	View Resume

CV RANKINGS

1 Rajeev Jain	**2** Priyanka Mehra	**3** Shashank Vishwakarma	**4** Supriya Yadav	**5** Amit M
tude Score: 2	Aptitude Score: 2	Aptitude Score: 1	Aptitude Score: 3	Aptitude Score: 0
onality	Personality	Personality	Personality	Personality
nness: 58.0%	Openness: 50.0%	Openness: 64.0%	Openness: 48.0%	Openness:
cientiousness: 56.0%	Conscientiousness: 62.0%	Conscientiousness: 66.0%	Conscientiousness: 62.0%	Conscientiousness:
aversion: 66.0%	Extraversion: 58.0%	Extraversion: 60.0%	Extraversion: 54.0%	Extraversion:
eableness: 68.0%	Agreeableness: 66.0%	Agreeableness: 58.0%	Agreeableness: 70.0%	Agreeableness:
roticism: 66.0%	Neuroticism: 54.0%	Neuroticism: 50.0%	Neuroticism: 56.0%	Neuroticism:

Fig 9.11 CV Ranking

SHASHANK

CONTACT : 9860923051

EMAIL : shashank@gmail.com

ADDRESS : Sanogaucharan

Personality

openness:	64.0%
conscientiousness:	66.0%
extraversion:	60.0%
agreeableness:	58.0%
neuroticism:	50.0%

MOST COMPATIBLE JOBS:

Full stack developer
Product Manager
Assistant Manager

Reload Compatible Jobs

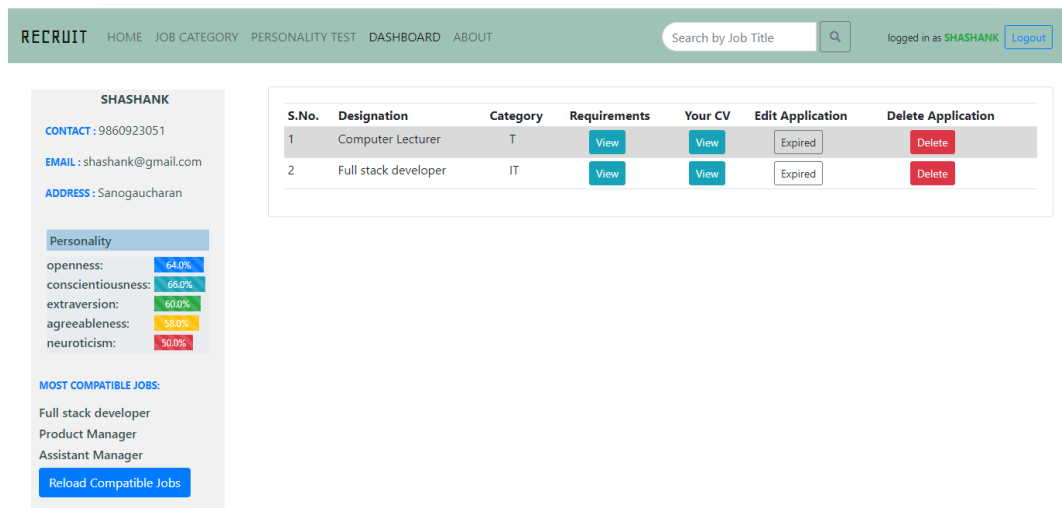| S.No.	Designation	Category	Requirements	Your CV	Edit Application	Delete Application
1	Computer Lecturer	T	View	View	Expired	Delete
2	Full stack developer	IT	View	View	Expired	Delete

Fig 9.12 Employee Dashboard

52