RESUME RECOMMENDATION USING MACHINE LEARNING

[1]Ramba S Naik, [2] Shrinivas R Dhotre

[1]Student, [2]Associate Professor
[1]Department of Computer Science and engineering,
[1]KLS Gogte Institute of Technology, Belagavi, India.
[2]Department of Computer Science and engineering,
[2]KLS Gogte Institute of Technology, Belagavi, India.

Abstract: Finding the appropriate contestants for a vacant job becomes a difficult process, especially when there are many job aspirants. Manual process of evaluating resumes could obstruct the recruiter team's efforts to appoint the right individual at the right time for that particular vacancy. The main aim of this thesis is that the proposed model should be able to find the right applicant resume from the pack of curriculum vitae or the resumes and also make sense of candidate's resume based on the personal information, qualification, skills and the experience. Assured whether the hired candidate is capable of doing the job. The proposed model will work in mainly in two steps that is firstly preprocessing which includes cleaning, pickle (which keeps the track of the objects it has already serialized), Tokenization, Preparation, Feature extraction (TF – IDF) and feature mapping of the resumes. Secondly deployment and inference this includes the classifier, similarity computation, ranking, and the result. This proposed system work with number of resumes for first classifying the right class using different classifier of machine learning, once classification has been done then as per the job description, then the appropriate resume will be recommended, Naïve Bayes, Content-based Recommendation, using cosine similarity and by using k-NN to identify the CVs that are nearest to the provided job description. K-nearest neighbor classifier and Naive Bayes can perform well to this proposed model compared to other machine learning classifiers. The proposed model is able to bifurcate among the 'n' number of resumes the system and should be able to suggest a short set of subsets of resumes that are appropriate**.**

Index Terms - **Resume Recommendation, Tokenization, Content based recommendation, K-Nearest Neighbor.**

I. INTRODUCTION

Since there is a increase in usage of the internet from the past few years so this has brought a huge change in the process of job recruitments. All the job openings are conducted online on the company's website or various online job recruitment platforms or employment platforms which are available online or social media which has made job opening announcement, advertisements which is readily made available much more accessible to job aspirants or the job seekers. Job applications has drafted to online with emails and direct applications on the recruitment platform as the standard. The direct applications and emails are the two main changes that has led to the thousands of job application for the recruitment process. This proposed model has overcome the all issues which are mentioned in the research gap of the literature survey done as a part of this study of proposed model.

The direct applications and emails are the two main changes due to internet that has led to the thousands of job application from the job seekers for the recruitment process.

Today the many companies, industries face few crucial challenges:

• Picking the right candidates from the pool of resumes – India is a vast country with lot of unemployment of many graduates in various fields of education. It is manually not possible to screen each and every resume that is received in the job portal and finding the right candidate for match for the given job description. This process of hiring makes the whole recruitment process gradually slow and thus making it tedious task causing efficient resources of the companies to waste.

• To be clear about the resumes of job aspirants or to make sense – The next challenge is related to the fact that is that the structure and format of the resumes in the market are practically nonstandard different and unique. The recruiter has to screen the resumes manually particular resumes to find the suitable candidate to match to the given description details of the job. There are chances of getting exposed to error where a fair applicant or the suitable job applicant for the particular vacany might get dropped in the job recruitment process.

• Knowing that candidates that are hired can do the job –The last challenge is about mapping the resume to the job description to understand if the hired candidate can do the job that is been assigned to him or her.

II. LITERATURE SURVEY

The literature survey provides a analysis on the project report. It briefs about the various research andreview study made in the field of research interest. Since the literature survey gives the right direction for the research area it is considered as to add a great significance. Literature survey provides the description of the sources that have contributed to carry out the research successfully. The guidelines are provided through which the various input can be provided in the literature survey.

[1] In this work, they have used the linear support vector machine for mapping of the resume and the job description but the limitation of this paper is that those resumes are taken in the csv format but in the reality the resumes are always available in the pdf format and also the summary is generated by using the genism.

[2] In this study the most relevant or the top candidate resumes are rated using the cosine similarity with the help of the random forest classifier when contracted with the different procedures it makes use of the algorithm such as the probabilistic stacked denoising auto-encoder (PSDAE) and another algorithm mainly a probabilistic matrix decomposition (PMF) model for the recommendation of the relevant resume.

[3] In this study HDFC is proposed based on the CDL algorithm which is used to compare the resumes for the recommendation but the yielded accuracy is low when compared to the other algorithms.

[4] In this study which overcomes the limitation of the traditional algorithms of filtering with the data that is available in sparse and cold start items it also needs large amount of data to take the input and train it. Mongo DB and were initialized on aws. The analysis of resume final results to a comprise a competitiveness score of education, experience, skills, and personal information.

[5] In this study the research is done on combination of two different algorithms like the NLP and CFR. Voting classifier estimators have given the accuracy of 59%. The Random Forest classifier algorithm produced the best performing model for both the Job Satisfaction and Job Retention Model while the Spacy Phrase Matcher gave the best one for the NLP Model.

[6] In this study the author has discussed about the Likert scale. The evaluation of the job vacancy recommendation results on the job applicant end was conducted through a questionnaire survey. The questionnaire was designed using a Likert scale for the respondents to provide their level of agreement with each item. On the 5-point Likert scale, the complete agreement is indicated as 5, 4 indicated half agreement, 3 indicated neutral opinion, 2 indicated disagreement was indicated with 2, and 1 indicated complete disagreement. The total average score was 3.39 specifically the recommendation results might not have met the needs of every job applicant.

The literature review articles were studied for this paper and it was discovered that many of them were a combination of more than two algorithms but in spite of this the accuracy yielded is very low, then some of the study suggest that the project carried out was Rest APi and other web applications. Also, some study has gone further for the question and answer-based application like taking the review from the job aspirants after the interview process is been done which is the lengthy process.

In this project we have done everything to overcome the limitation from the other paper that is the input is taken as the resume which in the pdf format and yields the appropriate accuracy. Firstly, the system is trained with the dataset which contains the categories and the field so that machine learns from the training data set and when the new resume is imported and the execution of the program takes further process it cleans the resume with the help of text processing and the cleaned resume is taken for the further execution. After similarity computation is done the appropriate resume is recommended. It actually has no criteria of having more data to execute the recommendation of the results.

III. RESEARCH GAP

The literature review articles were studied for this paper and it was discovered that many of them were a combination of more than two algorithms but in spite of this the accuracy yielded is very low, then some of the study suggest that the project carried out was Restfull APi and other web applications. Also some study has gone further for the question and answer based application like taking the review from the job aspirants after the interview process is been done which is the lengthy process which did not give the noteworthy results.

In this project we have done everything to overcome the limitation from the other paper that is the input is taken as the resume which in the pdf format and yields the appropriate accuracy. Firstly, the system is trained with the dataset which contains the categories and the field so that machine learns from the training data set and when the new resume is imported and the execution of the program takes further process it cleans the resume with the help of text processing and the cleaned resume is taken for the further execution. After similarity computation is done the appropriate resume is recommended. It actually has no criteria of having more data to execute the recommendation of the results.

IV. MOTIVATION

To reduce the work burden on the recruiting team and also to provide the justice to the job aspirant gives us the motivation to develop the recommendation system which helps the both teams to get the worth candidate and each resume received by the team should be screened well. For this particular project we have taken the example of the engineering candidate and the relevant resumes are compared based on the skills.

V. OBJECTIVES

The main objectives of this automated machine learning model include:
• Finding the suitable candidate's resume from the huge pool of resumes
• Understanding of resumes of the job aspirant's candidate.
• Making sense of the candidate's resume so that when they are hired the recruiter be confident that the candidate, they hired can do the job that is assigned to them.

The proposed model will work in two steps:
   i) Firstly, preprocessing which includes Cleaning, Pickling (which keeps the track of the objects it has already serialized), Tokenization, Preparation, Feature extraction (TF – IDF) and feature mapping.

ii)Secondly the deployment and inference this includes the classifier, mathematical computation, similarity computation, ranking, and the suggesting of the resume which is the final result.

VI. PROBLEM STATEMENT

The proposed model should be able to bifurcate the resumes based on the requirements of the recruiter the better resume should be suggested to the recruiter which will be able to guarantee the recruiter that the hired candidate will be capable of doing the job on being hired an thus the appropriate resume is recommended.

VII. SYSTEM ARCHITECTURE

With the increase in the advancement of the research in the various filed such as artificial intelligence which led to the improvement of the developing of various packages which reduces the human manual work the most important library which is used as apart of the natural language processing in the terms of text processing which is preprocessed for the further computation of the tokenization of the words. Keeping into this consideration a new design is suggested with new tools and design. The suggested design of the working model is as given in fig 1
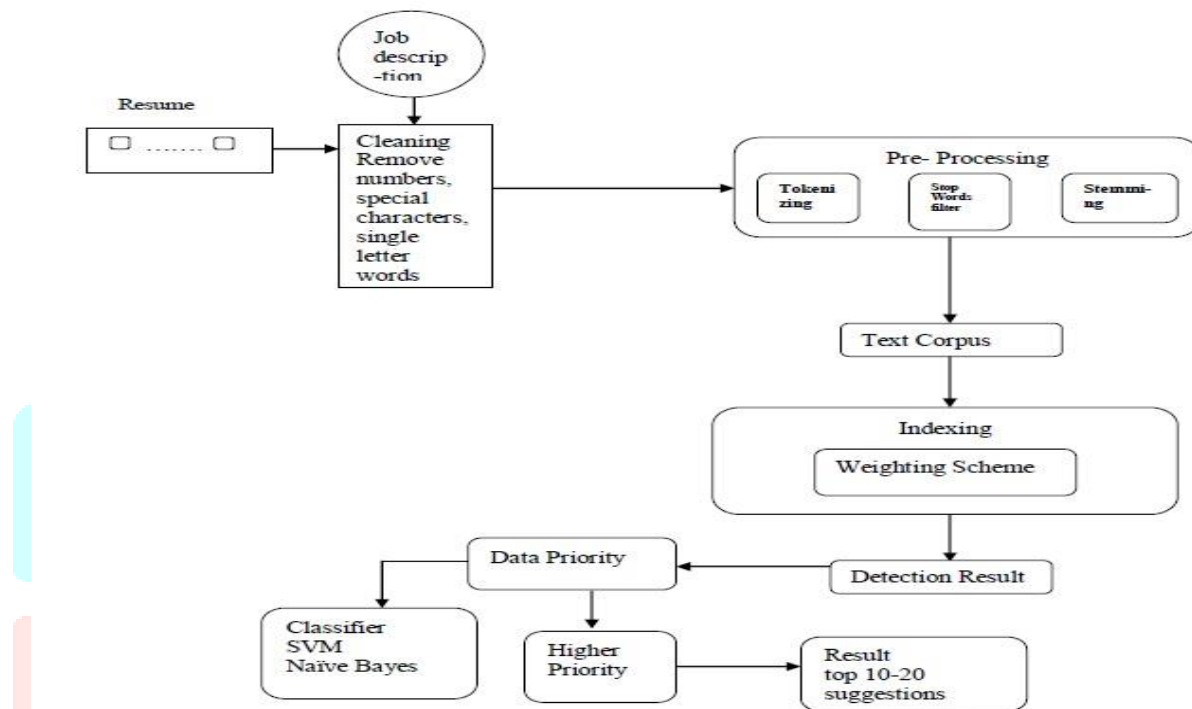
Fig 1: System architecture

This figure provides an explanation of the system's design and operation and serves as a representation of the entire system.

VIII. Hardware and software requirements

Hardware System Configuration

Processor	Above i3
Speed	1.1 G Hz
RAM	8 GB (min)
Hard Disk	20 GB

Software System Configuration

Operating System	Windows 10
Technology	Machine Learning
IDLE	Python 3.7 or higher

For a better understanding of machine learning e should be familiar with the python programming language. Python is a high-level programming language for a range of jobs (humanunderstandable languages are High level programming languages which are indeed understandable by humans)

In 1989, Guido Van Rossum created Python at the National Research Institute (NRI)in the UK.
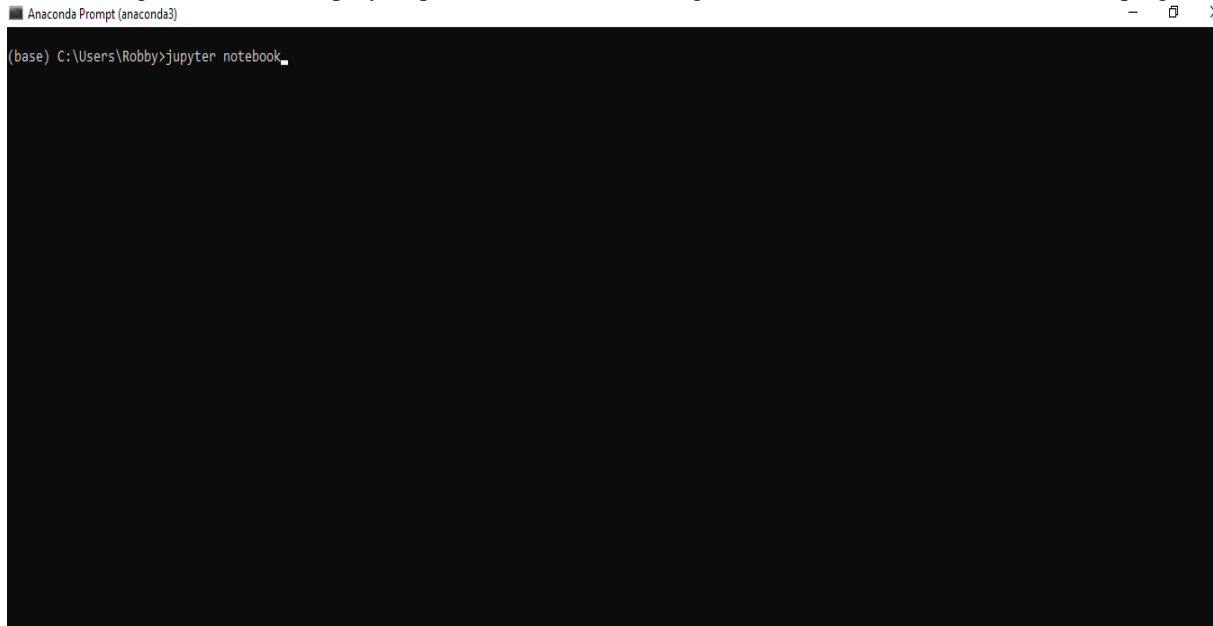
Python jupyter notebook is been used for the implementation of this project.

PYTHON JUPYTER NOTEBOOK is a open source project and is a part of Ipython kernel

IX. IMPLEMENTATION

The implementation briefs about the results obtained in the phase of implementation.

The below figures show the step-by-step execution. The below figures show the command to execute the program.
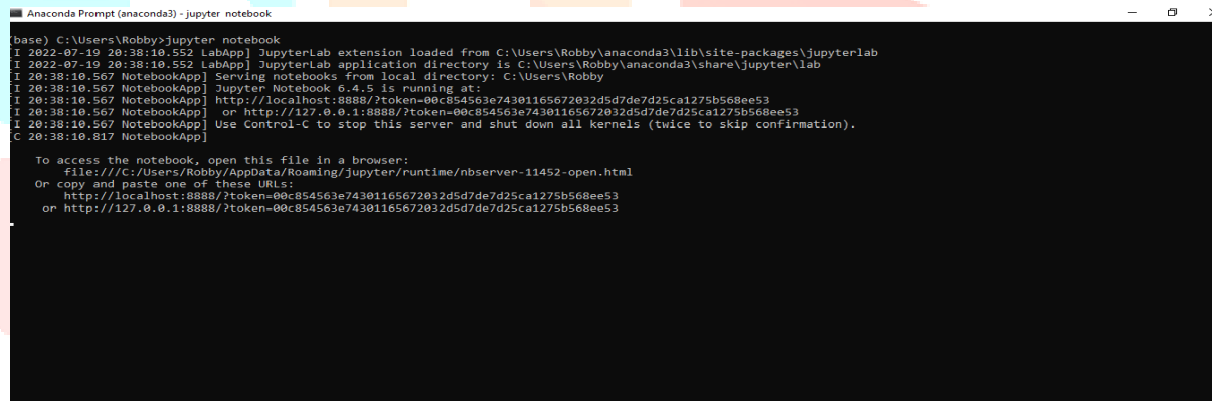

Fig 2: Command to Launch the Jupyter Notebook

For the execution of the program firstly will need to open the anaconda prompt which is similar to terminal or the command prompt(cmd).It refers to a black screen used to type in the commands by the user which can be downloaded by a conda using a GUI.

Fig 3: Jupyter Notebook runs successfully

Just by giving the command jupyter notebook in the anaconda prompt the jupyter notebook launches successfully and a new tab or the window pops up for the execution of the project

Fig 4: Jupyter Notebook runs successfully and is accessible

A particular folder is created by choosing the option as NEW and select a new folder so that all the related files are kept in the particular for the execution of the program

Fig 5: Displays the contents of the folder which are needed to be in single folder for the execution of the program

The related files like for in our case like the datasets, resumes which are used for the execution of the programs for reading the data values and evaluation of the results.

Fig 6: Displays the field of cleaned resumes

Fig 7: Classification of resumes based on the different categories

```
Displaying the distinct categories of resume and the number of records belonging to each category -
Java Developer              84
Testing                     70
DevOps Engineer             55
Python Developer            48
Web Designing               45
HR                          44
Hadoop                      42
Blockchain                  40
ETL Developer               40
Operations Manager          40
Data Science                40
Sales                       40
Mechanical Engineer         40
Arts                        36
Database                    33
Electrical Engineering      30
Health and fitness          30
PMO                         30
Business Analyst            28
DotNet Developer            28
Automation Testing          26
Network Security Engineer   25
SAP Developer               24
Civil Engineer              24
Advocate                    20
Name: Category, dtype: int64
```
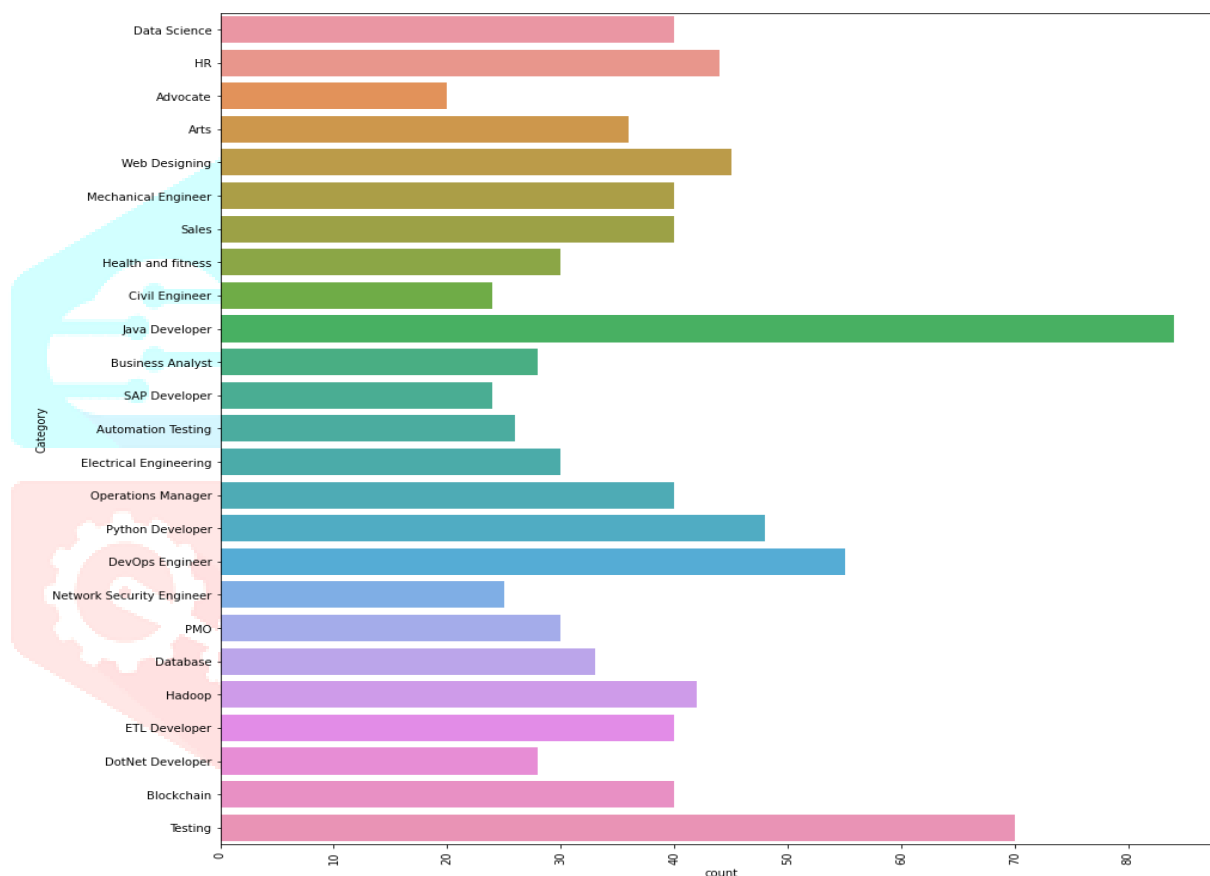
Fig 8: Total number of records of each category

Fig 9: Distribution of various areas of records based on count

Out[5]:

	Category	Resume	cleaned_resume
0	Data Science	Skills * Programming Languages: Python (pandas...	Skills Programming Languages Python pandas num...
1	Data Science	Education Details \r\nMay 2013 to May 2017 B.E...	Education Details May 2013 to May 2017 B E UIT...
2	Data Science	Areas of Interest Deep Learning, Control Syste...	Areas of Interest Deep Learning Control System...
3	Data Science	Skills â□¢ R â□¢ Python â□¢ SAP HANA â□¢ Table...	Skills R Python SAP HANA Tableau SAP HANA SQL ...
4	Data Science	Education Details \r\n MCA YMCAUST, Faridab...	Education Details MCA YMCAUST Faridabad Haryan...

Fig 10: Text mining process on resume dataset

```
[('Details', 484), ('Exprience', 446), ('months', 376), ('company', 330), ('description', 310), ('1', 290), ('year', 232), ('Ja
nuary', 216), ('Less', 204), ('Data', 200), ('data', 192), ('Skill', 166), ('Maharashtra', 166), ('6', 164), ('Python', 156),
('Science', 154), ('I', 146), ('Education', 142), ('College', 140), ('The', 126), ('project', 126), ('like', 126), ('Project',
124), ('Learning', 116), ('India', 114), ('Machine', 112), ('University', 112), ('Web', 106), ('using', 104), ('monthsCompany',
102), ('B', 98), ('C', 98), ('SQL', 96), ('time', 92), ('learning', 90), ('Mumbai', 90), ('Pune', 90), ('Arts', 90), ('A', 84),
('application', 84), ('Engineering', 78), ('24', 76), ('various', 76), ('Software', 76), ('Responsibilities', 76), ('Nagpur', 7
6), ('development', 74), ('Management', 74), ('projects', 74), ('Technologies', 72)]
```

Fig 11: Total number of counts based on resumes

```
terms = {'Quality/Six Sigma':['black belt','capability analysis','control charts','doe','dmaic','fishbone',
                              'gage r&r', 'green belt','ishikawa','iso','kaizen','kpi','lean','metrics',
                              'pdsa','performance improvement','process improvement','quality',
                              'quality circles','quality tools','root cause','six sigma',
                              'stability analysis','statistical analysis','tqm'],
         'Operations management':['automation','bottleneck','constraints','cycle time','efficiency','fmea',
                                  'machinery','maintenance','manufacture','line balancing','oee','operations',
                                  'operations research','optimization','overall equipment effectiveness',
                                  'pfmea','process','process mapping','production','resources','safety',
                                  'stoppage','value stream mapping','utilization'],
         'Supply chain':['abc analysis','apics','customer','customs','delivery','distribution','eoq','epq',
                         'fleet','forecast','inventory','logistic','materials','outsourcing','procurement',
                         'reorder point','rout','safety stock','scheduling','shipping','stock','suppliers',
                         'third party logistics','transport','transportation','traffic','supply chain',
                         'vendor','warehouse','wip','work in progress'],
         'Project management':['administration','agile','budget','cost','direction','feasibility analysis',
                               'finance','kanban','leader','leadership','management','milestones','planning',
                               'pmi','pmp','problem','project','risk','schedule','scrum','stakeholders'],
         'Data analytics':['analytics','api','aws','big data','busines intelligence','clustering','code',
                           'coding','data','database','data mining','data science','deep learning','hadoop',
                           'hypothesis test','iot','internet','machine learning','modeling','nosql','nlp',
                           'predictive','programming','python','r','sql','tableau','text mining',
                           'visualuzation'],
         'Web Development':['HTML','css','bootstrap','spring boot','php','jsp','python',
                            'coding','data','database','data mining','framework','deployment','hadoop',
                            'migrations','api','internet','UI','Front End','nosql'],
         'Healthcare':['adverse events','care','clinic','cphq','ergonomics','healthcare',
                       'health care','health','hospital','human factors','medical','near misses',
                       'patient','reporting system']
         }
```

Fig 12: Dictionary of skills

```
# Save pie chart as a .png file
pie.savefig('resume_screening_results[i].png')

Enter Resume Name with(.pdf) Extention: [_____]

In [ ]:

In [*]: import requests
        from IPython.display import Image
        from IPython.display import display

        print('First Resume Details\n\n')
        img0 = Image("resume_screening_results[i].png", width = 700)
        print('Second Resume Details\n')
        img1 = Image("resume_screening_results.png", width = 700)
        #img2 = Image("img2_path", width = 140)
        #img3 = Image("img3_path", width = 140)

        display(img0,img1)
```

Fig 13: Dialog box to enter the resume name
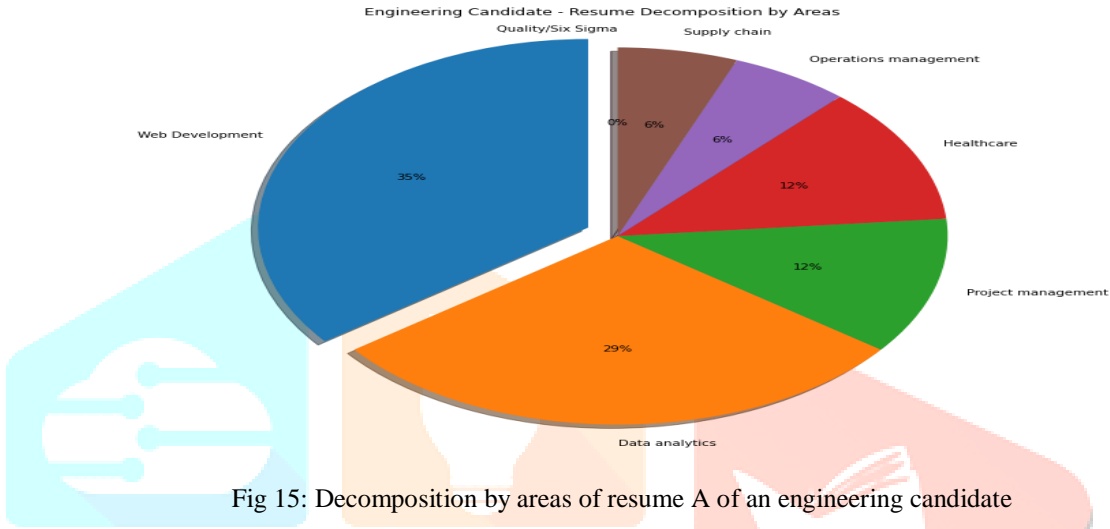
Fig 14: Resume name being entered

Fig 15: Decomposition by areas of resume A of an engineering candidate
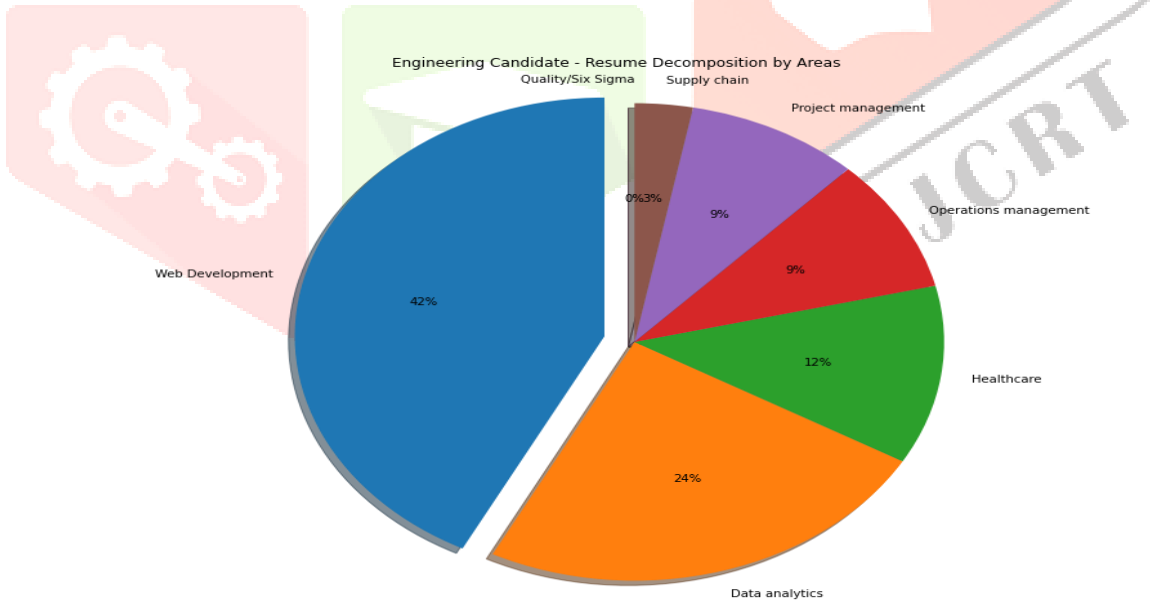

Fig 16: Decomposition by areas of resume B of an engineering candidate

Fig 16: shows the results of the recommended resume, where each resume is been mapped with the terms mentioned in the dictionary and then later on computation is done based on the dictionary terms of the individual resume and finally the resumes are compared with each other which inhibits the result of the cosine similarity found between the skills of the individual resumes.
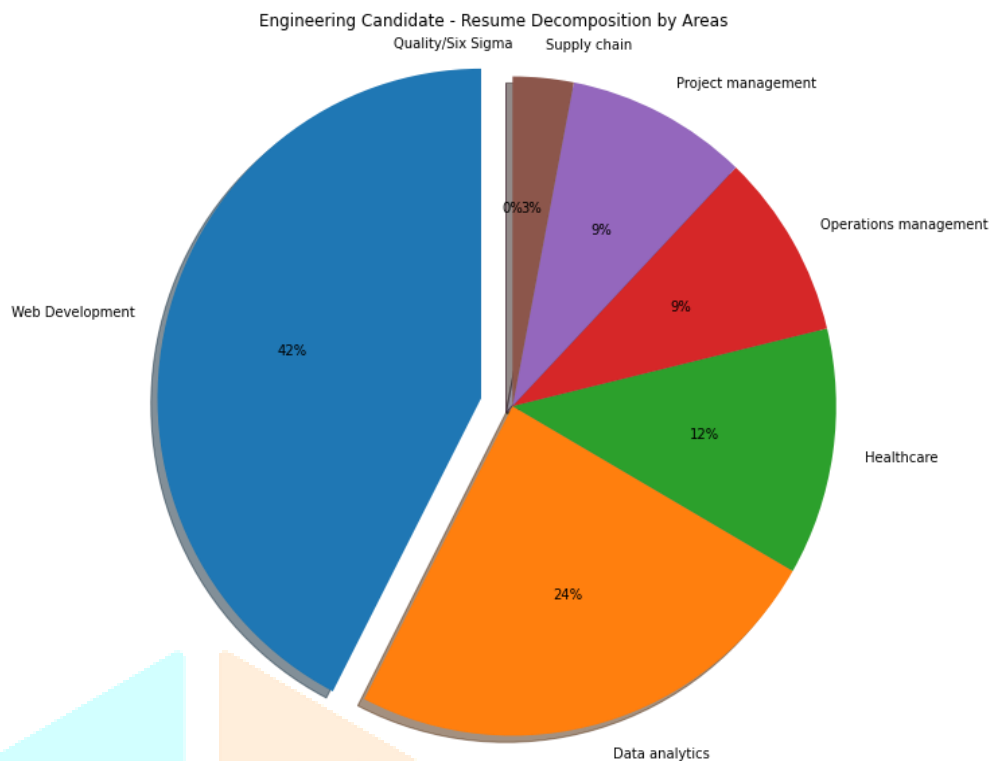
Engineering Candidate - Resume Decomposition by Areas

Fig 17: Recommends the appropriate resume

X. CONCLUSION

Many applications are received by the HR team and the recruitment team or organization for job vacancy opening. Searching the suitable or the appropriate candidate's resume from the huge pack of resumes is a monotonous task for any hiring team nowadays. The process of screening the candidate's resume is manual, time consuming, and waste of resources of the HR team. To overcome this issue, the proposed machine learning model is automated which recommends suitable candidate's resume to the recruiter based on given job requirements. The initial step towards building this model includes the cleaning of the resumes, which is quite important since a resume includes various irrelevant information, punctuation marks, special characters and so on which might cause the processing time delayed. Since the text mining process is approached for the cleaning of the resumes with the help of the NLTK library this task of cleaning resumes becomes easier. Another fact is that it will recommend the resume based on feature mapping done with the relevant areas of filed dictionary terms. With the help of the dictionaries list of all important features related to the particular areas are mapped with the relevant field of the job description and the results are depicted with the help of the pie chart and thus recommends the suitable resume to the recruiter. The proposed approach effectively captures the resume insights, their semantics and yielded an accuracy of 0.99 %. The proposed model is able to bifurcate the appropriate resume and recommends the suitable resume from the list of resumes.

XI. FUTURE SCOPE

Scheduling of the interview can be implemented as the future scope

REFERENCES

[1] Pradeep Kumar Roy, Sarabjeet Singh Chowdhary, Rocky Bhatia, "A Machine Learning approach for automation of Resume Recommendation system" 2020 and IEEE, Procedia Computer Science, Volume 167, 2020,

[2] Tejaswini K, Umadevi V, Shashank M Kadiwal, Sanjay Revanna, "Design and Development of Machine Learning based Resume Ranking System" 2021 and IEEE, (2021) ISSN 2666-4127,
https://doi.org/10.1016/j.susoc.2021.11.001.: 896-899

[3] W. Chen, X. Zhang, H. Wang and H. Xu, "Hybrid deep collaborative filtering for job recommendation," *2020 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)*, 2017, pp. 275-280, doi: 10.1109/CIAPP.2017.8167222.

[4] H. Jain and M. Kakkar, "Job Recommendation System based on Machine Learning and Data Mining Techniques using RESTful API and Android IDE," *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2019, pp. 416-421, doi: 10.1109/CONFLUENCE.2019.8776964.

[5] K. Appadoo, M. B. Soonnoo and Z. Mungloo-Dilmohamud, "Job Recommendation System, Machine Learning, Regression, Classification, Natural Language Processing," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411584.

[6] Y. -C. Chou and H. -Y. Yu, "Based on the application of AI technology in resume analysis and job recommendation," 2020 IEEE pp. 291-296, doi: 10.1109/ICCEM47450.2020.921