TERM PAPER

On

SMART TALENT RESUME RANKER

Submitted in fulfilment of the requirements for the Degree of

B.Tech in Information Technology

By

Kapil Israni (17104011)

Ayush Nagar (17104012)

Akshara Nigam (17104018)

Under the Supervision of Mahendra Gurve

Department of Computer Science Engineering and Information Technology
JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY
(Declared Deemed to be University U/S 3 of UGC Act)
A-10, Sector-62, Noida, India
October - 2020

Table of Contents

Introduction

The project aims to rank the resumes for a particular job description to predict the best suited candidate on the basis of their resumes. The set of resumes will be loaded as the input data along with the job description and analysis on the resume to predict the best suited candidate.

The summary in this report is a collection of our analysis on the basis of the research paper that was read during the course of this project. The research papers are mainly based on the concepts of how to read data of unstructured format and from different kinds of files such as pdf and document to find the relevant details from it. We also found out how entities of unstructured data are matched from another unstructured data to get the desired outcome and further learnt about feature selection models used to get better accuracy.

The ideas and algorithms are summarized together and hence our research to get the right outcome is taken forward.

Paper - 1

Paper Title	A Job Recommendation Method Optimized by Position Descriptions and Resume Information.
Author	Peng Yi, Cheng Yang, Chen Li, Yingya Zhang
Publisher	IEEE
Year	2016
Summary	Job recommendation algorithms which utilize recommendation methods to filter positions that do not meet the requirements and recommend the proper positions for job hunters play an important role in the recruitment websites. Based on the analysis of real recruitment data and the comparison of the existing recommendation methods, item based collaborative filtering algorithm has been used as the basic algorithm for a job recommendation. This paper produced an optimization algorithm to improve the accuracy of job recommendations. Historical delivery weight calculated by position descriptions and similar user weight calculated by resume information were added as two influencing factors in the preference prediction. The experiments tested on real recruitment data have shown that the optimization algorithm has greatly improved the final recommendation result. The F1-score of the optimized algorithm produced 9.6% better results than the basic algorithm.

TABLE IV. F1-MEASURE RESULTS

N	Basic Algorithm	Optimized Algorithm
1	41.21%	45.16%(+9.6%)
2	35.20%	38.75%(+10.%)
3	35.04%	38.25%(+9.2%)
4	34.57%	38.03%(+10.0%)
5	30.51%	32.18%(+5.5%)
6	28.57%	30.17%(+5.6%)

Paper - 2

Paper Title	A Job Post and Resume Classification System (JRC) for Online Recruitment
Author	Abeer Zaroor, Mohammed Maree, Muath Sabha
Publisher	International Conference on Tools with Artificial Intelligence, IEEE
Year	2017
Summary	The system exploits an integrated knowledge base for carrying out the classification task. Unlike conventional systems that attempt to search globally

	in the entire space of resumes and job posts, JRC matches resumes that only fall under their relevant occupational categories. The exploited knowledge base assists in (i) classifying resumes and job offers under their corresponding occupational categories and (ii) automatically ranking applicants that best match the announced offers. 1) Skill-Based Resume Classification Module: In this module, each skill in the skills set is submitted to the exploited knowledge base sequentially in order to obtain a list of candidate occupational categories. As a result, a list of weighted occupational categories is obtained and sorted by the highest weight (as one skill may return zero, one, or more than one occupational category) 2) Job Post Classification Module: In the Job Post Classification module, we use both the job title and the required skills from the structured job post for classification purposes. TABLE VII. COMPARATIVE EVALUATION – JRC VS. OTHER APPROACHES **Job title** / **Resume index** / **Manual score** / **Tf-idf Auto score** / **MatchingSem Auto score** / **JRC Auto score** **Back-end web developer** / CV1 / **0.38** / 0.16 / 0.30 / **0.45** CV2 / **0.26** / 0.19 / 0.19 / **0.19** CV3 / **1.0** / 0.56 / 0.70 / **1.0** **Java developer** / CV4 / **0.61** / 0.35 / 0.50 / **0.65** CV5 / **0.46** / 0.35 / 0.40 / **0.46** CV6 / **0.53** / 0.21 / 0.35 / **0.54** **Animator Designer** / CV7 / **0.35** / 0.20 / 0.20 / **0.35** CV8 / **0.70** / 0.61 / 0.70 / **0.75** CV9 / **0.20** / 0.20 / 0.25 / **0.25**

Paper - 3

Paper Title	Best Fit Resume Predictor
Author	Sujit Amin, Nikita Jayakar, M. Kiruthika, Ambarish Gurjar
Publisher	International Research Journal of Engineering and Technology (IRJET)
Year	2019
Summary	This paper focuses on the solution developed in the form of a web application to predict the best fit resumes against a given job description posted by a job recruiter. In this prototype, the web application can intelligently predict which resumes are better fit against the given job listing based on key factors of any candidate. These key factors include, but not limited to, education, number of years of experience and skills. This solution was developed on the purpose of significantly reducing the workload of the recruiters of any company who otherwise experience the pain of manually going through the details of each and every candidate's resume from the given pool of prospective candidates. The output of this will be visible only to the recruiter in the form of a rank list of all the candidates based on the overall resume scores assigned to each and every applicant on the basis of their education, work experience etc. The NLP framework used for the web application for data extraction was the SpaCy English model. The datasets used for dependency parsing on every candidate's resume were in the CSV format. The database which was used to store information of the job

	applicants including their resumes was MySQL. The accuracy achieved for the NLP model for this web application was around 67%.

Paper - 4

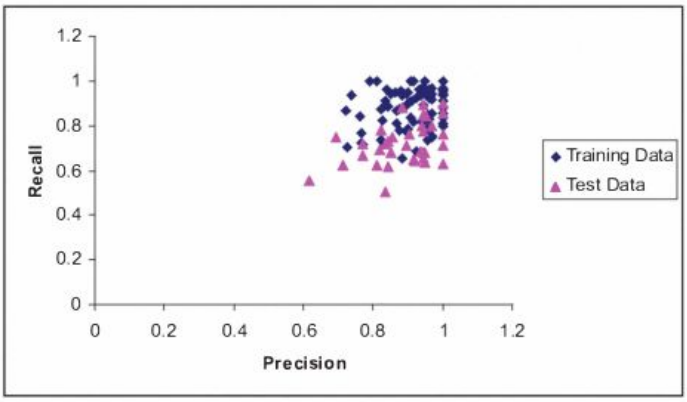
Paper Title	Towards an Information Extraction System based on Ontology to Match Resumes and Jobs
Author	Duygu Çelik, Aşkın Karakaş, Gülşen Bal, Cem Gültunca, Atilla Elçi, Başak Buluz, Murat Can Alevli
Publisher	IEEE 37th Annual Computer Software and Applications Conference Workshops
Year	2013
Summary	In this mentioned project, the system enables a free structured format of resumes to transform them into an ontological structure model. The produced system based on an ontological structure model and called Ontology based Resume Parser (ORP) is tested on a number of Turkish and English resumes. The proposed system is kept in a Semantic Web approach that provides companies to find job seekers in an efficient way. The system parses information from a resume such as general information, personal information, education information, work experience, qualifications, projects, certificates, references, other information etc and analyzes its data and infers new concepts from the written ontological rules with existing data. The system makes inference with the predefined semantic rules based on the resume knowledge that makes it differ substantially from other studies.

Paper - 5

Paper Title	Web Application for Screening Resume
Author	Sujit Amin, Nikita Jayakar, Sonia Sunny, Pheba Babu, M.Kiruthika, Ambarish Gurjar
Publisher	IEEE, International Conference on Nascent Technologies in Engineering
Year	2019
Summary	This paper focuses on a web application for screening Resumes of various candidates. The recruiters from various companies can post the details of the job openings available in their respective companies. The interactive web application allows the job applicants to submit their resume and apply for the job postings they may still be interested in. The resumes submitted by the candidates are then compared with the job profile requirement posted by the company recruiter by using techniques like machine learning and Natural Language Processing (NLP). Scores can then be given to the resumes and they can be ranked from highest match to lowest match. This ranking is made visible only to the company recruiter who is interested to select the best candidates from a

	large pool of candidates. The scores as well as the rank list will only be visible to the recruiter and not to the candidates. The recruiter can then make an informed decision on when to select for the next round of the hiring process. The job description text file is retrieved from the database. After that, the relevant entities of the candidate resume text file as well as the job description text file are then compared and a score is assigned to the candidate.

Paper - 6

Paper Title	Automatic Extraction of Usable Information from Unstructured Resumes to Aid Search
Author	Sunil Kumar Kopparapu
Publisher	IEEE
Year	2010
Summary	This paper describes a system for automated resume information extraction to support rapid resume search and management. The system is capable of extracting several important informative fields from a free format resume using a set of natural language processing (NLP) techniques. A working system is described, for automatic resume management. The system is capable of extracting six major fields of information. Experimental results carried out on a large number of resumes show that the proposed system can handle a large variety of resumes in different document formats with a precision of 91% and a recall of 88%.  *Figure 5 Precision and recall plot for train dataset (♦) and test datasets (▲).*

Paper - 7

Paper Title	A Machine Learning approach for automation of Resume Recommendation System
Author	Pradeep Kumar Roy, Sarabjeet Singh Chowdhary, Rocky Bhatia
Publisher	International Conference on Computational Intelligence and Data Science, Elsevier
Year	2019
Summary	The System produced in this paper, works with a large number of resumes first for classifying the right categories, then as per the job description top candidates would be ranked using Content-based recommendation using cosine similarity and KNN to identify the CV's that are nearest to the provided job description. The classification was done using four different models and their accuracy score was recorded. 1) Random Forest 2) Multinomial Naive Bayes 3) Logistic Regression 4) Linear SVC Results using the different classifiers **Classifier** / **Accuracy** Random Forest / 0.3899 Multinomial Naive Bayes / 0.4439 Logistic Regression / 0.6240 **Linear Support Vector Machine Classifier** / **0.7853**

Paper - 8

Paper Title	An Automatic Online Recruitment System based on Exploiting Multiple Semantic Resources and Concept-relatedness Measures
Author	Aseel B. Kmail, Mohammed Maree, Mohammed Belkhatir, Saadat M. Alhashmi
Publisher	IEEE 27th International Conference on Tools with Artificial Intelligence
Year	2015
Summary	This paper focuses on an automatic online recruitment system that employs multiple semantic resources to highlight the semantic contents of resumes and job posts. Additionally, it utilizes statistical concept-relatedness measures to further enrich the highlighted contents with relevant concepts that were not initially recognized by the used semantic resources. The system has been instantiated and validated in a precision-recall based empirical framework. The semantics-based system used is EXPERT which constructs ontology documents that describe both job posts and resumes based on the concept linking approach, and then ontology documents of job posts are mapped to ontology documents of resumes. The comparison of the system is given below :

	TABLE VI. PRECISION/ RECALL RESULTS System / Indicator: Our system: EXPERT P: 0.91: 0.89 R: 0.88: 0.93 F-measure: 0.89: 0.87 They have assigned different weights for optional and obligatory requirements, and then used these weights in computing the relevance scores between job posts and resumes.

Paper - 9

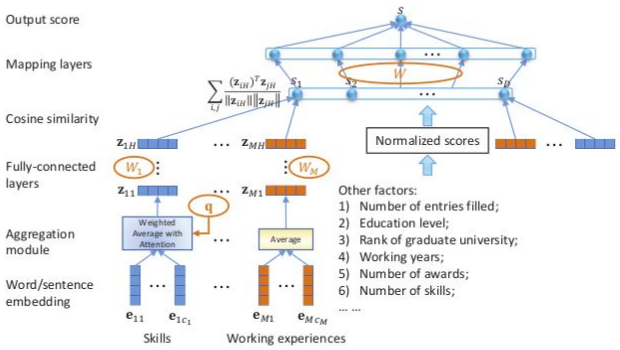
Paper Title	Smart Talents Recruiter – Resume Ranking and Recommendation System
Author	Ashif Mohamed, Wickram Bagawathinathan, Usama Iqbal
Publisher	IEEE
Year	2018
Summary	Smart Applicant Ranker is a candidate recommendation tool designed to supervise recruiters while they input their job requirements into the system. This system is designed using Ontology where they compare the resume models with the given job requirements to match the best comparable candidates. Two ranking algorithms are underlined in this system which will be invoked to assign a ranking point to the recommended candidates against the other candidates on the recommendation pool. This system will be kept in a Semantic Web approach that provides IT recruitment firms to seek experts in an efficient way. The ontology based web application is implemented using J2EE technologies running with Apache Tomcat server. In order to handle the business logics and the client calls to the server, Model View Controller pattern is used while MySQL database with JDBC interface is used to process simple user manipulations. For creating and manipulating Ontologies, OWL API is used via Apache Jena. The mode works has 3 main modules: A) Information Extraction B) Candidate search C) Candidate Ranking Algorithms. The similarity of skills is matched by the given formula : $\text{SimSkill}(J_i, R) = \begin{cases} 1, & J_i \in R \\ \max(\text{Skill}(J_i, R)), & J_i \notin R \end{cases}$ The performance of the candidate ranking module of the system is evaluated by the number of correctly ranked resumes with regard to the total number of resumes used for testing. In order to find out whether a resume is ranked correctly or not, the ranking assigned by the system is compared with the manual ranking given for that particular resume. If the ranking difference is not more than five, either positive or negative, the ranking given by the system is considered as correct

Table 2: Resume matching compatibility results

Resume No	Algorithm I	Algorithm II			Relative Score (RS)	SAR Ranking	Manual Ranking
		SWE	*SK*	*(SWE+SK)÷2*			
Resume 1	0.245	0.375	0.33	0.3525	0.29875	6	6
Resume 2	0.319	0.25	0.20	0.225	0.272	7	7
Resume 3	0.690	0.875	1.00	0.9375	0.81375	1	1
Resume 4	0.750	0.375	0.50	0.4375	0.59375	5	4
Resume 5	0.293	0.125	0.00	0.0625	0.17775	8	8
Resume 6	0.634	0.625	0.50	0.5625	0.59825	4	5
Resume 7	0.746	0.75	0.90	0.825	0.7855	2	2
Resume 8	0.789	0.75	0.66	0.705	0.747	3	3
Resume 9	0.165	0.125	0.14	0.1325	0.14875	9	9
Resume 10	0.075	0.25	0.00	0.125	0.1	10	10

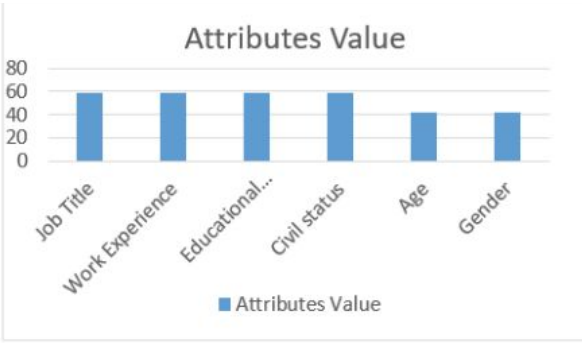
The results show that the system is useful in real-world online recruitment and ranking of candidate resumes, and has a better recommendation precision and efficiency than current existing systems.

Paper - 10

Paper Title	A Learning-based Framework for Automatic Resume Quality Assessment (RQA)
Author	Yong Luo, Huaizheng Zhang, Yongjie Wang, Yonggang Wen, Xinwen Zhang
Publisher	IEEE, International Conference on Data Mining
Year	2018
Summary	This paper throws light on the fact that from the talent perspective, many recruiters may want to know whether a resume is good enough or not. Therefore, the tool was developed to assess the quality of each resume automatically. Although there exist some resume quality assessment (RQA) websites (e.g., http://rezscore.com/), their underlying assessment schemes or algorithms are unknown and there is no public dataset for model training and evaluation. To tackle these issues, the authors had built a dataset and developed a general model for the same. The diagram of the system is given below :  From the system designed, following conclusions can be drawn that : 1) Learning adaptive weights using the attention scheme to aggregate multiple

	embeddings is superior to the simple average in general. 2) Either using the designed pair/triplet-based loss or adding a regularization term to utilize unlabeled data can improve the performance, it seems that the model based on triplet loss achieves the best performance overall. TABLE I A COMPARISON OF OUR MODELS WITH THE OTHER APPROACH IN TERMS OF F1-MEASURE. Methods / F1-measure L2 / 0.459±0.022 Contrastive / 0.500±0.054 Triplet / 0.541±0.051 MR / 0.492±0.109 Rezscore / 0.341

Paper - 11

Paper Title	Feature Selection for Job Matching Application using Profile Matching Model
Author	Leah G. Rodriguez, Enrico P. Chavez
Publisher	IEEE, 4th International Conference on Computer and Communication Systems
Year	2019
Summary	This paper aims to extract the relevant information from resumes and analyze it based on the different attributes. With the identification of the attributes, the proposed system is directed to adopt a clustering algorithm to match the profile of the job seekers against the requirements of the job posted by the prospect employers. Computing similarity scores between two profiles was the important task. For the similarity score, the values of common attributes in both profiles are extracted and their similarity scores are then computed and compared. Then, the obtained similarity scores are tuned in order to have more realistic scores that take into consideration the importance assigned to each attribute. By doing so, the new similarity value will tend to increase or decrease depending on the importance of each attribute. This tuning is an attribute based operation that outputs a new similarity score to each attribute by applying a weight to the computed similarity scores. The below graph shows the ranking of the attributes :  Figure 2. Ranking of identified attributes for profile matching.

Paper - 12

Paper Title	A Research of Job Recommendation System Based on Collaborative Filtering
Author	Yingya Zhang, Cheng Yang, Zhixiang Niu
Publisher	IEEE, 7th International Symposium on Computational Intelligence and Design
Year	2014
Summary	This paper contrasts between user-based and item-based collaborative filtering algorithms to choose a better performed one. They take background information including students' resumes and details of recruiting information into consideration, bring weights of co-apply users (the users who had applied the candidate jobs) and weights of student used-liked jobs into the recommendation algorithm. It also takes into consideration four Methods of Similarity Calculation (i) Cosine Similarity (ii) Tanimoto Coefficient (iii) Log Likelihood (iv) The City Block Distance. The accuracy for both the filtering methods is given as follows:

TABLE I. PRECISION AND RECALL OF DIFFERENT RECOMMEDERS

Recommender(r_num=3)	Similarity	Precision	Recall
User-Based CF (n=10)	Log likelihood	62.82%	53.85%
	City Block	83.33%	56.41%
	Tanimoto	65.38%	53.85%
Item-Based CF	Log likelihood	58.33%	58.33%
	City Block	0.00%	0.00%
	Tanimoto	41.67%	41.67%

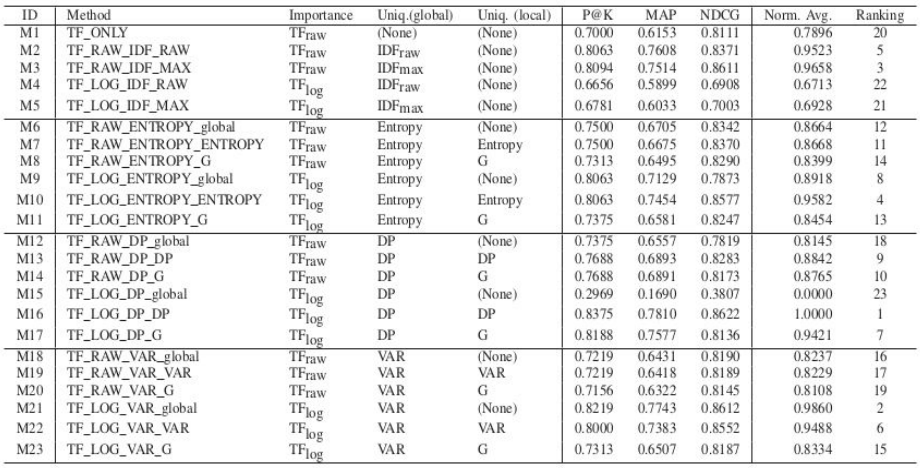
Paper - 13

Paper Title	Dynamic User Profile-Based Job Recommender System
Author	Wenxing Hong, Siting Zheng, Huan Wang
Publisher	IEEE. 8th International Conference on Computer Science & Education
Year	2013
Summary	This paper challenges the traditional job applicant system that takes the personal information and job intention of an applicant, and uses it to generate the recommendation result by employing the recommendation algorithms. It stated the shortcomings such as, the personal information and job intention may not be true because of the job applicant's cognitive deviation. Further, the job applicant does not

	update his/her personal information in general after entering the information on the recruiting website for the first time. Considering these situations they employed the dynamic recommendation in a job recommender system. It uses a threefold method : • Based on the basic features of jobs applied by an applicant which indicate his/her preference, the basic features of this applicant are updated automatically and at regular intervals. • From the perspective of dimensionality, they used the extracted feature for feature selection to extend the number of features. Along with the increasing number of applied jobs, the number of extended features will become greater and they will change. • According to the characteristics of dynamic user profiles, they used a hybrid recommendation algorithm, i.e. user based collaborative filtering algorithm, for improving the accuracy and effectiveness of the recommendation results.

Paper - 14

Paper Title	Quantifying Skill Relevance to Job Titles
Author	Wenjun Zhou, Yun Zhu, Faizan Javed, Mahmudur Rahman, Janani Balaji, Matt McNair
Publisher	IEEE, International Conference on Big Data
Year	2016
Summary	In this study, the goal was to profile job titles by effectively quantifying the relevance of skills. It started with using a naive, frequency-based skill ranking approach, which resulted in the most generic skills ranked on the top and hence they adopted a number of alternative metrics and compared their performances on a number of job titles. They adapted information theoretic metrics and measurements of variation to assess the (un)certainty of a skill to a title, to adjust for the frequency of very commonly required skills. The basic idea was to leverage the dispersion of a skill term across different job titles. The intuition was that the more titles that require a skill (i.e., the skill is “dispersed”), the less unique the skill is to any title. On the contrary, if a skill was required by just a few titles, it was quite unique to those titles. The compared results are as follows :

	METHODS IMPLEMENTED AND OVERALL PERFORMANCE (ACROSS ALL TITLES)  While the TF-IDF measure only considers the skill-title relationship like document-term relations, they have further considered the variation within a given job title, where the variation among job ads were considered. By collecting and comparing with expert ranked skills for a random set of job titles, our experiments showed that the (un)certainty measures did help improve skill rankings, especially when they used the DP for both global and local uniqueness measures. They also found that the performance of all such measures vary greatly among different titles, and deduplicating similar ads before computing relevance scores has consistently helped improve the performance.

ID	Method	Importance	Uniq.(global)	Uniq. (local)	P@K	MAP	NDCG	Norm. Avg.	Ranking
M1	TF_ONLY	TF_{raw}	(None)	(None)	0.7000	0.6153	0.8111	0.7896	20
M2	TF_RAW_IDF_RAW	TF_{raw}	IDF_{raw}	(None)	0.8063	0.7608	0.8371	0.9523	5
M3	TF_RAW_IDF_MAX	TF_{raw}	IDF_{max}	(None)	0.8094	0.7514	0.8611	0.9658	3
M4	TF_LOG_IDF_RAW	TF_{log}	IDF_{raw}	(None)	0.6656	0.5899	0.6908	0.6713	22
M5	TF_LOG_IDF_MAX	TF_{log}	IDF_{max}	(None)	0.6781	0.6033	0.7003	0.6928	21
M6	TF_RAW_ENTROPY_global	TF_{raw}	Entropy	(None)	0.7500	0.6705	0.8342	0.8664	12
M7	TF_RAW_ENTROPY_ENTROPY	TF_{raw}	Entropy	Entropy	0.7500	0.6675	0.8370	0.8668	11
M8	TF_RAW_ENTROPY_G	TF_{raw}	Entropy	G	0.7313	0.6495	0.8290	0.8399	14
M9	TF_LOG_ENTROPY_global	TF_{log}	Entropy	(None)	0.8063	0.7129	0.7873	0.8918	8
M10	TF_LOG_ENTROPY_ENTROPY	TF_{log}	Entropy	Entropy	0.8063	0.7454	0.8577	0.9582	4
M11	TF_LOG_ENTROPY_G	TF_{log}	Entropy	G	0.7375	0.6581	0.8247	0.8454	13
M12	TF_RAW_DP_global	TF_{raw}	DP	(None)	0.7375	0.6557	0.7819	0.8145	18
M13	TF_RAW_DP_DP	TF_{raw}	DP	DP	0.7688	0.6893	0.8283	0.8842	9
M14	TF_RAW_DP_G	TF_{raw}	DP	G	0.7688	0.6891	0.8173	0.8765	10
M15	TF_LOG_DP_global	TF_{log}	DP	(None)	0.2969	0.1690	0.3807	0.0000	23
M16	TF_LOG_DP_DP	TF_{log}	DP	DP	0.8375	0.7810	0.8622	1.0000	1
M17	TF_LOG_DP_G	TF_{log}	DP	G	0.8188	0.7577	0.8136	0.9421	7
M18	TF_RAW_VAR_global	TF_{raw}	VAR	(None)	0.7219	0.6431	0.8190	0.8237	16
M19	TF_RAW_VAR_VAR	TF_{raw}	VAR	VAR	0.7219	0.6418	0.8189	0.8229	17
M20	TF_RAW_VAR_G	TF_{raw}	VAR	G	0.7156	0.6322	0.8145	0.8108	19
M21	TF_LOG_VAR_global	TF_{log}	VAR	(None)	0.8219	0.7743	0.8612	0.9860	2
M22	TF_LOG_VAR_VAR	TF_{log}	VAR	VAR	0.8000	0.7383	0.8552	0.9488	6
M23	TF_LOG_VAR_G	TF_{log}	VAR	G	0.7313	0.6507	0.8187	0.8334	15

Paper - 15

Paper Title	Skills and the graduate recruitment process: Evidence from two discrete choice experiments
Author	Martin Humburg, Rolf van der Velden
Publisher	Elsevier
Year	2015
Summary	In this study the authors elicit employers' preferences for a variety of CV attributes and types of skills when recruiting university graduates. Using two discrete choice experiments, they simulate the two common steps of the graduate recruitment process: (1) the selection of suitable candidates for job interviews based on CVs, and (2) the hiring of graduates based on observed skills. In line with the preferences in the first step, employers' actual hiring decision is mostly influenced by graduates' level of professional expertise and interpersonal skills. Other types of skills also play a role in the hiring decision but are less important, and can therefore not easily compensate for a lack of occupation specific human capital and interpersonal skills. From the results they concluded that there was a large impact of interpersonal skills on graduates' chances to get hired is in line with earlier studies emphasizing the increasing

	importance of communication in today's work-life in general, and especially for team productivity. Other types of skills and attributes also play a role in the recruitment process but are less important and can therefore not easily compensate for a lack of more specific human capital and interpersonal skills. The large standard deviations of the estimated mean coefficients imply that there is not one graduate profile which all employers prefer. Rather, employers' demand for skills varies substantially. Some employers may not want to recruit the graduates with the highest skill levels because the job does not require them and they fear that graduates will get bored too quickly. Other employers, and the in-depth interviews confirm this, may not have a strong preference for graduates with high professional expertise because they have the internal training facilities to teach them the occupation specific knowledge they need. The same employers may therefore put more emphasis on other, more transversal types of skills such as general academic skills because they are an important ingredient for further professional growth. Employers' willingness to pay for skills. MeanWTP SD *Professional expertise* High 14.9% 28.5% Average Ref. Low −35.9% 38.2% *General academic skills* High 9.0% 19.3% Average Ref. Low −26.5% 33.5% *Innovative/creative skills* High 11.5% 19.4% Average Ref. Low −30.7% 34.6% *Strategic/organizational skills* High 11.1% 20.4% Average Ref. Low −25.8% 25.3% *Interpersonal skills* High 12.4% 24.8% Average Ref. Low −39.1% 39.7% *Commercial/entrepreneurial skills* High 7.3% 33.4% Average Ref. Low −32.8% 32.5%

Paper - 16

Paper Title	Resume Parser
Author	Aneesha T Ibrahim, Annette J K, Geethika S, Archana Naik
Publisher	International Journal of Innovations in Engineering and Technology (IJIET)
Year	2018

Summary	This paper discusses developing a parsing application, for resumes received in multiple formats which include .docx, .doc, and .pdf. This application reduces the time and manual effort of searching through the multiple resumes for choosing the suitable resumes. The technique involved is known as resume parsing. Other names include, resume extraction, CV parsing, CV extraction, which allows the automated storage and analysis of resume information. Multiple resumes are uploaded into parsing software and the information is extracted so that it can be sorted and searched. Resume parsers first analyzes a resume, and then extracts the desired information. After the resume has been analyzed, a recruiter can specify the job skills required and get a list of relevant resumes as the output. Some parsers provide semantic search, which adds context to the search terms and tries to understand the intent in order to make the results more reliable and comprehensive.

Paper - 17

Paper Title	Challenge: Processing Web Texts for Classifying Job Offers
Author	Flora Amato, Roberto Boselli, Mirko Cesarinit, Fabio Mercoriot, Mario Mezzanzanica, Vincenzo Moscato, Fabio Persia and Antonio Picariello
Publisher	IEEE 9th International Conference on Semantic Computing
Year	2015
Summary	The contribution of this work goes towards the direction by classifying Web job offers onto the categories of a well established classifier. To this end, they have proposed to apply several (and different) text classification techniques to classify a real dataset of Web job offers. Furthermore, the effectiveness of each approach is evaluated by comparing classification results against a gold classification manually performed by domain experts. Two machine learning classifiers were used to perform the text classification purposes: the LinearSVC (an implementation of Support Vector Machine Classification using a linear kernel) and the Perceptron classifier, both built using the Scikit-Learn framework. The results are given below:-

	LDA	Rules	Linear SVC	Percept.
Accuracy	.5 (.51)	.444 (.469)	**.556 (.633)**	.483 (.543)
Avg. Precision	**.507 (.587)**	.353 (.432)	.259 (.576)	.284 (.503)
Avg. Recall	**.502** (.538)	.354 (.432)	.272 (**.576**)	.254 (.503)
Avg. FScore	**.471** (.532)	.328 (.462)	.26 (**.607**)	.263 (.564)
Precision Std.Dev.	.41 (.305)	.378 (.328)	**.34 (.239)**	.364 (.243)
Recall Std. Dev.	.391 (.274)	.348 (.224)	.358 (.224)	**.332 (.195)**
FScore Std. Dev.	.366 (.253)	**.326** (.257)	.337 (.216)	.336 (**.201**)

TABLE I
TEXT CLASSIFICATION TECHNIQUES SCORES.

Paper - 18

Paper Title	ResuMatcher: A personalized resume-job matching system
Author	Shiqiang Guo, Folami Alamudun, Tracy Hammond
Publisher	Elsevier
Year	2016
Summary	The proposed model, intelligently extracts the qualifications and experience of a job seeker directly from his/her resume, and relevant information about the qualifications and experience requirements of job postings. Using a novel statistical similarity index, ResuMatcher returns results that are more relevant to the job seekers experience, academic, and technical qualifications, with minimal active user input. The system comprises of the following main components :- Job Data Processor, Search Interface, and the Resume Matcher. 1) The Job Data Processor executes a daily batch job of crawling the web for different job postings and building the Jobs Model. 2) The Search Interface provides an interactive front-end from which it accepts the resume of users and builds a Resume Model. 3) The Resume Matcher receives as input, a resume object from the Search Interface, and queries the job database using a novel similarity method to retrieve the most relevant jobs. The similarity between a resume object and a job object is calculated as a weighted sum of the computed features. The comparison is given below :-

Comparison with www.indeed.com keyword search.

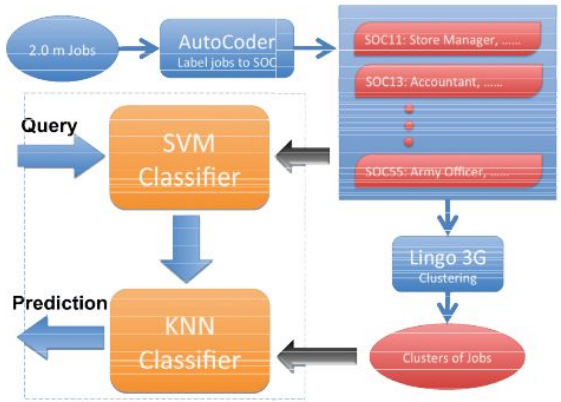
	Precision@k		DCG	
k	Indeed	RésuMatcher	Indeed	RésuMatcher
5	0.84	0.87	23.87	32.97
10	0.72	0.86	37.02	45.57
20	0.645	0.768	58.70	66.70

Paper - 19

Paper Title	Resume Information Extraction With A Novel Text Block Segmentation Algorithm
Author	Shicheng Zu and Xiulai Wang
Publisher	International Journal on Natural Language Computing (IJNLC)
Year	2019
Summary	In this study, they have proposed an end-to-end pipeline for resume parsing based on neural networks-based classifiers and distributed embeddings. This pipeline leverages the position-wise line information and integrated meanings of each text block. The

	coordinated line classification by both line type classifier and line label classifier effectively segment a resume into predefined text blocks. The proposed pipeline joins the text block segmentation with the identification of resume facts in which various sequence labelling classifiers perform named entity recognition within labelled text blocks. Comparative evaluation of four sequence labelling classifiers confirmed BLSTM-CNNs-CRF's superiority in named entity recognition tasks.

Paper - 20

Paper Title	Carotene: A Job Title Classification System for the Online Recruitment Domain
Author	Faizan Javed, Qinlong Luo, Matt McNair, Ferosh Jacob, Meng Zhao, Tae Seung Kang
Publisher	IEEE First International Conference on Big Data Computing Service and Applications
Year	2015
Summary	In this paper they present Carotene, a machine learning-based semi-supervised job title classification system. Carotene leverages a varied collection of classification and clustering tools and techniques to tackle the challenges of designing a scalable classification system for a large taxonomy of job categories. Carotene's classification component is composed of a hierarchical coarse and fine-level classifier cascade where a fine level classifier utilizes job title datasets for every SOC (Standard Occupational Classification) major depending on the classification results of the coarse-level classifier.  Fig. 2. Architecture of Carotene

Paper - 21

Paper Title	Applicability of Naïve Bayes Model for Automatic Resume Classification
Author	Patrick Nyanumba Mwaro, Dr. Kennedy Ogada, Prof. Wilson Cheruiyot
Publisher	International Journal of Computer Applications Technology and Research
Year	2020
Summary	In the research, the Bayesian model was trained using a dataset collected through extracting information from advertisements and also through interviews with few selected experts. The dataset was divided into five subsets. This was done because the dataset was small and the major objective of this research was to improve predictive accuracy of Naïve Bayes classifier by combining four homogeneous Naïve Bayes models developed from the four data subsets. The base classifiers were then combined to develop Ensemble Naïve Bayes Classifier(ENBC). Both the original Naïve Bayes and Ensemble Naïve Bayes Classifier were used to classify resume data and their accuracies were recorded and compared. It was noted that the predictive accuracy of Ensemble Naïve Bayes Classifier was better than the original Naïve Bayes Classifier.

Table 9. Models Performance Measures

Model	Accuracy	Precision	Recall
NBC-1	88.8889%	0.9375	0.8824
NBC-2	87.0370%	0.7727	1.0000
NBC-3	92.5926%	0.8889	0.9412
NBC-4	90.7407%	0.9333	0.8235

Table 10. Ensemble Performance Measures

Model	Accuracy	Precision	Recall
ENBC	94.4444%	0.9412	0.9412

Paper - 22

Paper Title	Automated Tool For Resume Classification Using Semantic Analysis
Author	Suhas Tangadle Gopalakrishna and Vijayaraghavan Varadharajan
Publisher	International Journal of Artificial Intelligence and Applications (IJAIA)
Year	2019
Summary	This paper discusses the design and implementation of a resume classifier application which employs an ensemble learning based voting classifier to classify a profile of a candidate into a suitable domain based on his interest, work-experience and expertise

mentioned by the candidate in the profile. The model employs topic modelling techniques to introduce a new domain to the list of domains upon failing to achieve the threshold value of confidence for the classification of the candidate profile. The Stack-Overflow REST APIs are called for the profiles which fail on the confidence threshold test set in the application. The topics returned by the APIs are subjected to topic modelling to obtain a new domain, on which the voting classifier is retrained after a fixed interval to improve the accuracy of the model.

Association Rule Mining algorithm is done on the data dump in order to obtain the related topics for a given particular topic based on the responses from the users for the question related to the topic. The response from the Stack-Overflow APIs is then subjected to topic modelling to obtain a suitable specialisation name for the resume in consideration. The new specialisation is further added to the list of domains on which the ML model was trained initially. The ML model is then retrained with the additional new domain added to the list to improve the accuracy of the prediction. In this way, they are able to eliminate the dependency of initial training set for accurate prediction of the profile into a suitable domain.

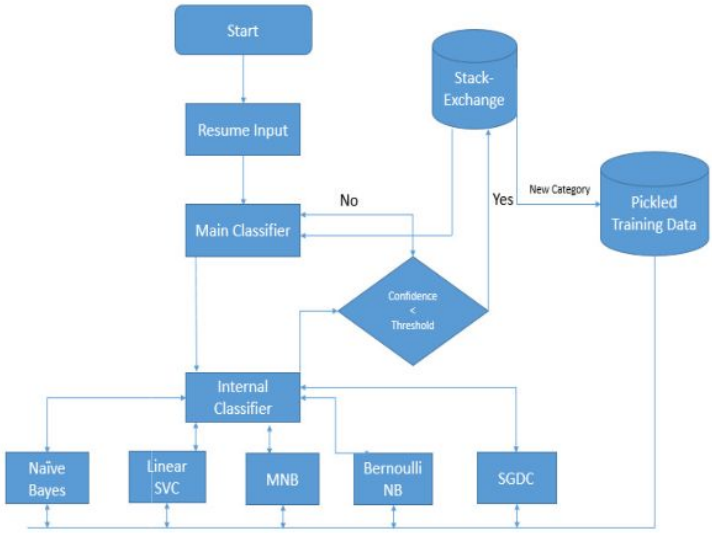


Figure 9: Flow chart of classification module

The accuracies of each of the classifiers is given below :-

Table 2: Efficiency of individual classifiers

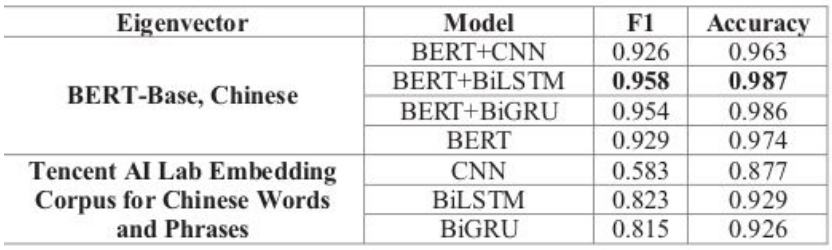
Name of the Classifier	**Efficiency of prediction in %**
Naïve Bayes	79
Linear SVC	83
MNB	91
Bernoulli NB	89
Logistic Regression	81
K- nearest neighbours	80

Paper - 23

Paper Title	An Integrated System for Occupational Category Classification based on Resume and Job Matching
Author	Disha Lamba, Shivam Goyal, V. Chitresh, Neha Gupta
Publisher	International Conference on Innovative Computing and Communication (ICICC)
Year	2020
Summary	The Smart Resume Selector (SRS) model proposed in this paper first segments a resume to extract vital information. The resume is then converted into tokens which are compared with the defined array of information and skill set as per the company's requirement. An overall score is calculated for each resume and its area of expertise out of the basic 5 occupational categories. The work is as follows :- ● Automatic occupational category classification of resumes ● Section-based scoring ● NLP independent therefore can adapt to new domains ● Sorting resumes based on category and score

Paper - 24

Paper Title	The Resume Corpus: A Large Dataset for Research in Information Extraction Systems
Author	Yanyuan Su, Jian Zhang Jianhao Lu
Publisher	International Conference on Computational Intelligence and Security
Year	2019
Summary	The paper aims to publish a corpus of Chinese resumes for information extraction research. In order to evaluate the potential of this corpus, they perform a series of experiments to train and evaluate several neural network models on it. The models contain Convolutional Neural Network (CNN), Bi-directional Long Short-Term Memory (BiLSTM), Bi-directional Gated Recurrent Unit (BiGRU), and Bi-directional Encoder Representation from Transformers (BERT). They first take a partition of data from the corpus, then manually label some tags on this part as a labelled dataset. Finally, use the cross-validation method to evaluate the potential of the corpus. They suppose that a corpus is useful if it can always get a stable result over different types of neural network models. After using the corpus they have compared the accuracies :-

Eigenvector	Model	F1	Accuracy
BERT-Base, Chinese	BERT+CNN	0.926	0.963
	BERT+BiLSTM	**0.958**	**0.987**
	BERT+BiGRU	0.954	0.986
	BERT	0.929	0.974
Tencent AI Lab Embedding Corpus for Chinese Words and Phrases	CNN	0.583	0.877
	BiLSTM	0.823	0.929
	BiGRU	0.815	0.926

Paper - 25

Paper Title	Named Entity Recognition using Hidden Markov Model (HMM)
Author	Sudha Morwal, Nusrat Jahan and Deepti Chopra
Publisher	International Journal on Natural Language Computing (IJNLC)
Year	2012
Summary	In this paper they have described the Hidden Markov Model (HMM) based approach of machine learning in detail to identify the named entities. The main idea behind the use of HMM model for building NER systems is that it is language independent and we can apply this system for any language domain. In our NER system the states are not fixed means it is dynamic in nature one can use it according to their interest. The corpus used by our NER system is also not domain specific. The Viterbi algorithm is implemented to find the most likely tag sequence in the state space of the possible tag distribution based on the state transition probabilities. The Viterbi algorithm allows us to find the optimal tags in linear time. The idea behind the algorithm is that of all the state sequences, only the most probable of these sequences need to be considered.

Paper - 26

Paper Title	A System for Detecting Professional Skills from Resumes Written in Natural Language
Author	Emil St. Chifu, Viorica Rozina Chifu, Iulia Popa, Ioan Salomie
Publisher	IEEE
Year	2017
Summary	In this paper, they presented an ontology based system capable of detecting professional skills (as noun phrases) from CVs or resumes written in natural language. The system uses a domain ontology of skills with more than 13,000 concepts, but an important aspect of our work is the ability to detect even new skills, i.e. skills that are not present in the ontology. The educational and personal information is extracted by

	using a Hidden Markov model and Support Vector Machines.

Paper - 27

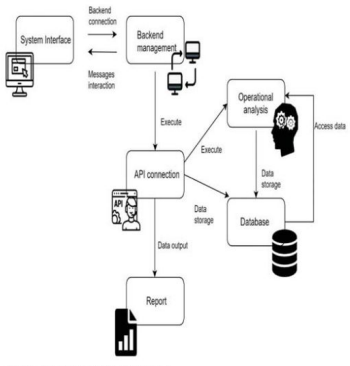
Paper Title	Language Modelling for Collaborative Filtering: Application to Job Applicant Matching
Author	Thomas Schmitt, François Gonnard, Philippe Caillou, Michele Sebag
Publisher	International Conference on Tools with Artificial Intelligence, IEEE
Year	2017
Summary	This paper addresses a collaborative retrieval problem, the recommendation of job ads to applicants. Specifically, two proprietary databases are considered. The first one focuses on the context of unskilled low-paid jobs/applicants; the second one focuses on highly qualified jobs/applicants. Each database includes the job ads and applicant resumes together with the collaborative filtering data recording the applicant clicks on job ads. The proposed approach, called LAJAM , focuses on the semi-cold start recommendation problem of recommending new job ads to known applicants. This setting is relevant to the temporary job sector, of increasing importance for current job markets. LAJAM learns a continuous language model on the job ad space, trained to comply with the collaborative filtering metrics. This language model, implemented as a neural net, can flexibly take into account heterogeneous additional information, e.g. related to the posting time and geolocation of the job ads.

Paper - 28

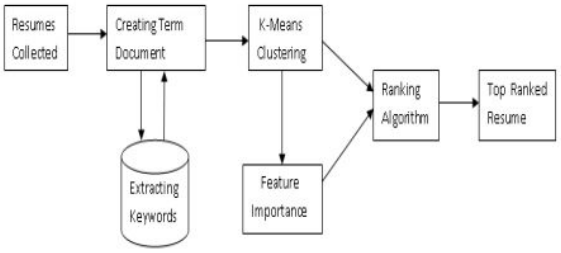
Paper Title	A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm.
Author	Gerard Deepak, Varun Teja & A. Santhanavijayan
Publisher	Journal of Discrete Mathematical Sciences and Cryptography
Year	2020
Summary	In this paper, contemporary Natural Language Processing techniques have been leveraged to demonstrate the capability of data-driven HR towards significant improvement in the quality and speed of the whole recruiting process. Firstly, by using NLP, a resume parser has been implemented to analyze the most crucial recruitment parameters. Thereafter, the ability to display a pie chart for a candidate has been employed in the algorithmic structure of the parser to prepare a powerful tool for the resume matching based on job criteria. To determine the efficacy and accuracy of the proposed resume ranker, an enhanced rival modern optimizer, i.eFirefly ranking algorithm is applied to accelerate the speed of ranking algorithm. The accuracy of the model to extract information is given below :-

	Performance Measures of Named Entity Recognition | Experimentation Entities	Recall %	Precision %	F-measure %	Accuracy %
---	---	---	---	---
Name	92.47	95.64	94.01	94.06
Qualification	92.81	95.77	94.32	94.3
Skillset	93.21	94.63	93.89	93.92
Mobile number	91.83	96.91	94.43	94.37
Email	91.47	93.82	92.69	92.645
Work Experience	93.33	95.87	94.57	94.6

Paper - 29

Paper Title	A Resume Evaluation System Based on Text Mining
Author	Yi-Chi Chou, Chun-Yen Chao, Han-Yen Yu
Publisher	IEEE
Year	2019
Summary	This study developed an AI-based interviewing system to reduce the loss of talent caused by the emotional reactions and subjectivity of interviewers when viewing resumes. The designed system performs the function of resume assessment and explores the personality traits of candidates by classifying them into four dimensions of soft power, namely dominance, influence, steadiness, and compliance (DISC) after assessing the submitted electronic resumes. This system also assesses three dimensions of competence, namely education and experience, skills, and personality traits, which are indicated by the information contained in a resume. Finally, the designed system quantifies the aforementioned DISC data and three competency dimensions by scoring each resume. The results are then compiled into a report that contains the personal analysis, ranking, and distribution forecast for the candidate in question.  Fig. 2. Backend architecture diagram

Paper - 30

Paper Title	Cluster based Ranking Index for Enhancing Recruitment Process using Text Mining and Machine Learning
Author	Mayuri Verma
Publisher	International Journal of Computer Applications
Year	2017
Summary	This paper presents an effective approach for extracting relevant words from the resumes using Term Document Matrix. The role of the candidate, various skills, familiarity with various frameworks, experienced skills and operating systems have been considered. A clustering methodology has been used to find the similar resumes. The importance of each word has been calculated according to the cluster. The appropriate rank of the resumes have been calculated. The experimental results show that Cluster Based Ranking gives the potentially best candidate for a particular job profile. The weighted importance in calculating the ranks is the very first effort in itself. The architecture of the model is shown below :-  **Figure 1: Proposed Model**

Summarization

From the above research papers, here's an excerpt :

SNo.	Paper	Algorithm/Model	Challenges	Drawbacks
1	Paper - 1	Item-based, user-based recommendation system User-similarity, major similarity	Due to the limited data, it is difficult to calculate the similarity of major just by the name of major	It is a static based algorithm.
2.	Paper - 2	NLP after parsing the resumes to read the skills and experience (N-gram and Tokenization). Tf-Idf for skill set matching.	Finding and hiring the right talent from a wide and heterogeneous range of candidates remains one of the most important and challenging tasks.	The main drawback of this approach is the huge run time complexity of the matching process. Also a large fraction of the produced results suffer from low precision since the information extraction process passes through two loosely coupled stages,
3	Paper - 3	Cosine Similarity, NLP, Regex matching	This resume is raw and unstructured data and it is a challenge to extract relevant important data.	There were issues while calculating the number of years of relevant experience for any prospective job candidate.
4	Paper - 4	N-Gram Algorithm, Jaro-Winkler Algorithm, Regex Matching, Knowledge Based Expert System	Using different classes to store the information of the candidates.	Wrong/incomplete information might be provided to the system.
5	Paper - 5	NLP, Spacy Pipeline	To reduce the time complexity of comparing and matching the candidate's profile and job posted.	It converts the input data first into JSON format to be passed to the NLP pipeline for matching.
6	Paper - 6	A mix of NLP techniques and heuristics were used to build information extraction modules to aid extraction of useful information from resumes. The knowledge base was created using reference resumes and the system was tested on a large number of resumes which was not part of the reference resumes.	Automatic extraction of information from resumes with high precision and recall is not an easy task essentially because of the non-standardization of resume structure. In spite of constituting a restricted domain, resumes can be written in multitude of formats (e.g. structured tables or plain texts) and in different file types (e.g txt, .pdf, .doc(x) etc.)	Extracting some information using only HR-XML
7	Paper - 7	It uses Content-based collaborative recommendation using cosine similarity and KNN for identifying the CV's closest to the job profile.	Different structure and format of every CV. Mapping the CV to the right job description	i) Model takes CVs in CSV format. ii) Generation of a summary using genism library might cause loss of important

				information due to compression of the text.
8	Paper - 8	Tf-idf weighting, Semantic resources, Semantic Networks, Semantic Network Enrichment	Approaches based on keyword matching ignore the semantics of the job post and resume contents; and consequently a large portion of the matching results is irrelevant. The more recent semantics-based models are influenced by the limitations of the used semantic resources, namely the incompleteness of the knowledge captured by such resources and their limited domain coverage.	Drawbacks associated with the limited domain coverage and semantic knowledge incompleteness problems
9	Paper - 9	Ontology models, Ontology Language (OWL), Natural Language Processing, Ranking Algorithm, Cosine similarity	Currently, various job portals utilize a combination of distinct algorithms so as to rank applicant profiles. The ranking policies are still inefficient, considering the way that highly impactful and conceivable factors that would describe an individual are not considered.	The incomplete analysis of the resume analysis.
10	Paper - 10	Multi-layer neural network, Cosine Similarity	Since there is no public algorithm for RQA, we submit our labeled resumes to a website (http://rezscore.com/), which can assign a grade for each resume.	Lacking a larger corpus that includes job-post information and identify more useful features for RQA
11	Paper - 11	Feature selection, Cosine Similarity, Weighted Similarity for ranking.	One of the most challenging tasks of this type of job matching was that there was a bulk of information to coordinate against and it was in free form.	Lack of training data set from job seekers and company, need to conduct tests of the clustering model to verify reliability and performance of the job matching system
12	Paper - 12	Collaborative Filtering, Cosine Similarity, Tanimoto Coefficient, Log Likelihood, The city block distance.	Collaborative Filtering approaches often suffer from three problems: cold start, scalability and sparsity.	It's a comparison between the two kinds of recommendation systems so it doesn't rank the candidates, it just compares based on the accuracy.
13	Paper - 13	User based collaborative filtering, TF-IDF, Feature selection using Information Gain	Given a job applicant, his/her user profile is updated and extended dynamically, and then a hybrid recommendation algorithm is employed to generate the results and achieve the dynamic recommendation.	The context formed in the peak season and the off season has an influence on the job desire of a job applicant.
14	Paper - 14	TF-IDF, NLP, min-max normalization	Frequency-based skill weighting resulted in the most generic skills ranked on the top, yet they were not the most relevant, we first considered TF-IDF adjustments	Weighted versions of the metrics needs to be considered

15	Paper - 15	Econometric model, Experimental Survey	Problems of unob-served heterogeneity that often hampers conclusions based on cross-sectional data	Hypothetical bias – the divergence of stated and revealed preferences – cannot be entirely excluded.
16	Paper - 16	NLP for parsing the document and Regex Matching to extract specific information like Location, Phone Number etc.	The ambiguity problem caused by the polysemy of the phrases, in expressing the skills in the text of a resume	There is a limitation in the number of resumes that can be processed at a time. Further work involves overcoming that barrier and making it a highly efficient parser. Also the system needs to be made more large-scale.
17	Paper - 17	NLP pipeline, Linear SVC, Perceptron, Levenshtein metric, Latent Dirichlet Allocation	The problem of incorrect classification is related to the presence of categories having a similar lexicon and a different number of training samples. About 30% of the offer titles do not carry enough information to identify the occupation.	The paper does not extract other relevant information like, the required skills, contract types, business sectors, education levels, etc. Also, it does not involve multi-classification.
18	Paper - 18	NLP, Cosine Similarity	A significant amount of ambiguity exists between domain specific words and their respective interpretation. Resumes both contain richer and more complex words that cannot be described simply by keywords.	Firstly, they only consider the shortest path between concept pairs. When faced with more complex structures, such as multiple taxonomic inheritance, the accuracy similarity measures is significantly reduced. Another limitation of the path-based approaches is the assumption that all links in the taxonomy have uniform distance.
19	Paper - 19	NLP Pipeline, Text-CNN, Named Entity Recognition (NER), TF-IDF, Cosine Similarity	The uncertainties associated with the resume layouts pose a significant challenge to an efficient resume parsing and reduce the accuracy for subsequent candidate recommendation. Also, the challenge in classifying the job description is to determine the beginning and end of the description.	Does not incorporate the ontology of each person to develop the talent recommendation system.
20	Paper - 20	KNN, SVM Classifier, Lingo3G, TF-IDF	The class imbalance problem that exists in the dataset. Also, hierarchical classifiers suffer from error propagation where the final classification decision relies on the accuracy of a series of previous decisions in the hierarchy and thus is more likely to make a misclassification.	The architecture to support classification at the O*NET code level and more than one coarse-level SVM classifications at the kNN level.
21	Paper - 21	Naive Bayes Classifier, Feature	The biggest challenge was in data	Using homogenous base

		Extraction	management, to support the construction and maintenance of machine learning models over large data which was multidimensional and Evolving. Also the Bayesian network classifier was error prone in high confidence labels.	classifiers to investigate the overall effects on the predictive accuracy of the model.
22	Paper - 22	NLP Pipeline, Naive Bayes, Linear SVC, Multinomial NB, Bernoulli NB, Logistic Regression, KNN, Named Entity Recognition (NER), Association Rule Mining, Latent Dirichlet Allocation (LDA)	The most important challenge faced was for profiles which do not fall under any of the given domains, the classifier arbitrarily classifies the profile into any domain.	The retraining of the individual classifiers after a fixed interval to incorporate the classification of new domains returned by the Stack-Overflow API trained on Association Rule Mining of user question and answer dump, has caused performance issues on the model.
23	Paper - 23	PDFMiner, Word Count	The problem with NLP is its domain is fixed, it can work on a range of data only. In the case of adaptive nature, it fails to show prominent results. One of the other drawbacks is that precision is low in most results.	Job requirements are put manually into the system, which doesn't make the model dynamic.
24	Paper - 24	NLP, CNN, BiLSTM, BiGRU, BERT	There is a big challenge to extract useful information from massive amounts of data.	There is a large difference in the F1-score and accuracy of CNN based on word embedding. It is probably due to the limited number of neural network layers of standard CNN model.
25	Paper - 25	NLP, Named Entity Recognition (NER), Hidden Markov Model (HMM), Viterbi	Large number of ambiguity exists in Indian names and this makes the recognition a very difficult task for Indian language. Also, there is a lack of standardization and spelling.	The NER-HMM model works only for 22 Indian Languages but there is a scope of expansion.
26	Paper - 26	Named Entity Recognition (NER), Named Entity Normalization (NEN), NLP, Hidden Markov Model, Support Vector Machine	The ambiguity problem caused by the polysemy of the phrases that express skills in the text of a resume. For the skill insertion process an issue occurs, i.e., finding the right place in the ontology for the new skill (finding the right taxonomic parent for the new skill.)	Searching Wikipedia for every phrase that expresses a new skill causes performance issues.
27	Paper - 27	TF-IDF, Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Item-based Recommendation System	JAM systems involve two issues: finding representation(s) for resumes and/or job ads, and a distance on top of these representations, amenable to predict whether a (resume, job ad) pair is well-suited to each other.	The deep-LAJAM is always dominated by shallow LAJAM, this is because of the lack of regularization for the deep architecture. Also, a limitation of the approach is

				that deep-LAJAM currently lags behind shallow-LAJAM .
28	Paper - 28	NLP, Regex Matching, Firefly Algorithm, Named Entity Recognition (NER)	The extraction of texts from the document and the order and form of information in it is of primary concern in parsing resumes.	The size of the resume corpus used should have been large for better accuracy.
29	Paper - 29	Jieba, PDFMiner	To be able to create a real-time report and store the huge dataset.	It does most of the work using predefined API calls and may falter at response time of the API.
30	Paper - 30	K-means, Frequency Count, Cluster Based Ranking	Using an unsupervised learning algorithm and being able to categorize the resumes for the dataset.	Improving the productivity in the recruitment process by enhancing the model accuracy and time Evaluation should be done on large dataset.

Significance of Work

Through this project we found ways of using and storing unstructured data and how feature extraction works. The huge format of resumes and being able to parse them was indeed a challenging task. It thus helped us delve deeper into feature extraction and natural language processing. This project has helped us develop a new perspective for commonly used machine learning algorithms and to extract more than just predictions. The natural language processing toolkit (NLTK) and KNN algorithms, have been helpful in carrying out the results. Also, working with huge unstructured data, with a variety of forms was a challenge, especially to extract correct information from it. This project thus optimizes the conventional resume matching methods which parse the entire set of resumes and do the matching between the job description and the resumes. In this project a resume once fetched into the system is parsed and its job profile category is determined via the machine learning model which is then used to match with a Job Description looking for similar profile people.

References

[1] “A System for Screening Candidates for Recruitment”, Amit Singh, Rose Catherine, Karthik Visweswariah, Vijil Chenthamarakshan, Nanda Kambhatla, October, 2010, Toronto, Ontario, Canada.

[2] ”Matching People and Jobs: A Bilateral Recommendation Approach”, Jochen Malinowski, Tobias Keim, 39th Hawaii International Conference on System Sciences - 2006

[3] “Cluster based Ranking Index for Enhancing Recruitment Process using Text Mining and Machine Learning”, Mayuri Verma, International Journal of Computer Applications - 2017.

[4] "Automatic extraction of usable information from unstructured resumes to aid search", Kopparapu, Sunil Kumar, Progress in Informatics and Computing (PIC), 2010 IEEE International Conference on. Vol. 1. IEEE.

[5] “Resume Parser: Semi-structured Chinese Document Analysis”, Zhang Chuang, Wu Ming, Li Chun Guang, Xiao Bo, WRI World Congress on Computer Science and Information Engineering, April 2009.

[6] “Towards an Information Extraction System Based on Ontology to Match Resumes and Jobs”, Celik Duygu, Karakas Askyn, Bal Gulsen, Gultunca Cem, IEEE 37th Annual Workshops on Computer Software and Applications Conference Workshops, July 2013.

[7] "Algorithm AS 136: A k-means clustering algorithm", Hartigan, John A., and Manchek A. Wong. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108

[8] “Collaborative Filtering based on Subsequence Matching: A New Approach”, Alejandro Bellogín, Pablo Sánchez, Informatics and Computer Science Intelligent Systems Applications, August 2017.

[9] “A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm”, Gerard Deepak, Varun Teja & A. Santhanavijayan, Journal of Discrete Mathematical Sciences and Cryptography, April 2020.

[10] “End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT”, Vedant Bhatia, Prateek Rawat, Ajit Kumar, Rajiv Ratn Shah, IEEE 2009.

[11] “Resume: A Robust Framework for Professional Profile Learning & Evaluation”, Clara Gainon de Forsan de Gabriac, Amina Djelloul, Constance Scherer, Vincent Guigue, Patrick Gallinari, Journal, 2000.

[12] “Resume Information Extraction With A Novel Text Block Segmentation Algorithm”, Shicheng Zu and Xiulai Wang, International Journal on Natural Language Computing (IJNLC) Vol.8, No.5, October 2019.

[13] “Learning to Rank Resumes”, Sangameshwar Patil, Girish K. Palshikar, Rajiv Srivastava, Indrajit Das, IEEE, 2012.

[14] “The Resume Corpus: A Large Dataset for Research in Information Extraction Systems”, 15th International Conference on Computational Intelligence and Security (CIS), 2019.