iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IV **Month of publication:** April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4258

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Implementation of Online Portal for CV Analysis Using KNN Algorithm

Bhavya Doshi[1], Hemendra Gandhi[2], Nirmal Jayadharan[3], Megharani Patil[4]
[1, 2, 3, 4]*Computer Department, Thakur College of Engineering & Technology, Mumbai, India.*

Abstract: ***The quick development of Information and Communication Technologies (ICTs) has resulted in people turning to the web for seeking a job and to develop the career. A lot of companies use online portals to shortlist candidates for the interview process exploiting the advantages of the internet. These are termed as e-recruitment systems and this automates the process of receiving the CVs of the candidates and short listing the candidates who satisfy the requirements of the company. In this work we propose the application of K-Nearest Neighbors algorithm and Natural Language Processing in automated e-recruitment systems to shortlist the right candidates who satisfy the technical requirements of the company for the interview process.***
Keywords: NLP; CV; machine learning;

I. INTRODUCTION

This project will provide a more effective way to select candidates from a large number of applicants for the interview process of companies. The registered users have to upload their resume in the systems portal and the system will shortlist the candidates based on their technical skills and extracurricular activities. If the candidate satisfies the technical requirements of the company then he would be shortlisted for the next round. The shortlisted candidates would then have to give a technical test based on his major technical skills followed by an aptitude test and personality test. If the candidate successfully clears the entire test process then he would be selected for the interview process in the company. If he fails to clear the tests or does not satisfy the requirements the system will provide suggestions to the candidates on areas he needs to improve. The existing system has many drawbacks like copying someone else's resume or mining it from online. The existing system doesn't even have a proper format which will be overcome by our current system. It'll contain a plagiarism check and if the resume is not updated according to the format specified, it'll be automatically rejected. We will use NLP (Natural Language Processing) to scan the keywords and KNN
algorithm for classification of resumes. We will do text categorization using this algorithm. We will use semantic matching to check whether the qualifications of the candidate would match with requirements of the company.

II. RELATED WORK

The quick development of modern information technology in the past few years has caused an increase in number of people turning to the jobs in web development. Many companies use online knowledge management systems or AI to hire employees. These are called as e-recruitment systems which automates the process of receiving CVs and ranking them accordingly according to their mentioned skills. E-recruitment systems has seen an explosive expansion in the past few years allowing Human Resources (HR) agencies to target a very wide audience at a small cost. Automating the process of analyzing the applicant profiles to determine the ones that fit the positions specifications could lead to an increased efficiency [1]. This project uses automated techniques to identify, extract, and exploit information from CVs to find the most appropriate one for a given post. Our work focuses on CVs analysis [4]. An empirical study is conducted to validate the proposed process and we show that there is an improvement in the extraction phase. It presents a New Fuzzy Expert System (NFES) for student academic performance evaluation based on Fuzzy Logic techniques. It introduces fuzzy logic and illustrates how these principles could be applied by the e-recruitment systems to evaluate students' academic performance. The aim of this NFES methodology is to adaptively adjust the training for each particular student on the basis of his/her achievements in their respective CVs. This means that the NFES will monitor the student's CVs and have the ability to make decision about next step training. Several approaches using fuzzy logic techniques have been mentioned to provide a practical method for evaluating student academic performance and compare the results (performance) with existing statistical method[2].Many approaches can be applied to automate the e-recruitment process combining techniques of NLP and fuzzy logic and machine learning. In this work we have implemented an e-recruitment system that automates the candidate evaluation and recommend them the courses to enhance their CVs.

©IJRASET: All Rights are Reserved

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

III.K-NEAREST NEIGHBOURS

Nearest neighbours is a supervised learning model which is used for classification and regression analysis. But strictly speaking, KNN does not have any learning involved, i.e., there are no parameters we can change to make the performance better nor there is an objective function which we can optimize. This is a major difference from most supervised learning algorithms. KNN is a rule that can be used in production time that can classify an instance based on its neighbours. Computing neighbours and the vector distances does not require class label which is used to make the decision for the classification. For computing its nearest neighbours it has 3 different algorithms and the best one is chosen to implement based on the size of the training data. Brute force, KD tree and Ball tree algorithms are used to determine the distances of each of the vectors from the ideal vector which in our case will be the job description provided by the recruiter. If the distance is less then they are more suited for that job.

A. Algorithm

1) Start.
2) Initialize, define K.
3) Compute the distance between input sample and the training samples.
4) Sort the distance.
5) Take K nearest neighbours.
6) Apply simple majority.
7) End.

The first analysis used a KNN to rank the resumes according to their vector distance from the ideal vector. The KNN was chosen because it is considered one of the best initial classification algorithms and is not so complicated compared to a CNN. The KNN is based on nearest neighbour search. Hence it can be defined as: given a set S of points in a space M and a query point $q \in M$, find k closest points in S to q. Here in our case the query point is our job description vector. The k closest points are vectors eligible for that particular job description and are then ranked according to their distances. More specifically, distance between 2 vectors is the square Euclidean distance.

For vectors v1(w1,x1,y1,z1) and v2(w2,x2,y2,z2)

Squared Euclidean distance = $(w_1-w_2)^2 + (x_1-x_2)^2 + (y_1-y_2)^2 + (z_1-z_2)$

The optimal algorithm that is to be used for finding the nearest neighbours is based on majorly four factors:

a) Size of dataset
b) Data structure used
c) Number of neighbours k requested for a query point
d) Number of query points

KNN algorithm is based on feature similarity that is how closely our out-of-sample features resemble our training set determines how we classify a given data point. KNN might be computationally expensive i.e. the distance calculation cost may be high because of the cost of formation of tree data structures. It has high memory requirement. And due to these features the prediction time might be slow.

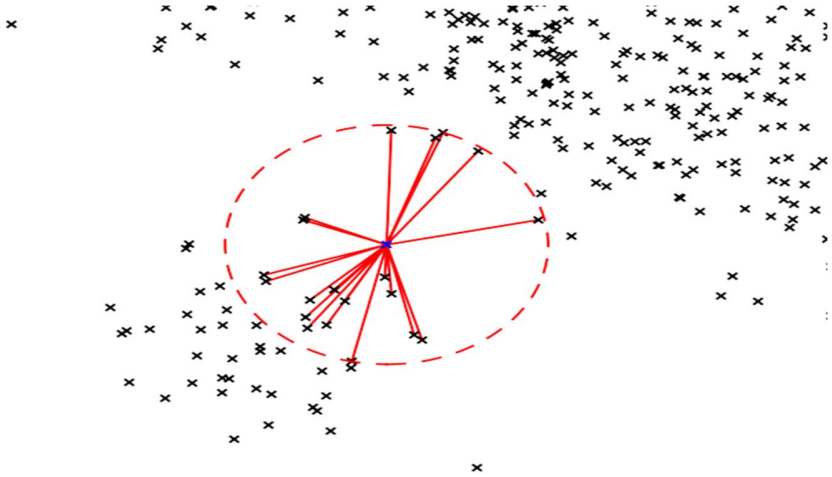


FIG.1: K Nearest Neighbours

©IJRASET: All Rights are Reserved

B. Brute force Approach

The brute force approach algorithm can be used only when the training dataset is small that is N<30. It is the most naive approach for nearest neighbours. This involves brute force computation of distances between two vectors. Brute force query time grows as O[DN] where N is no. of samples and D dimensionality of samples i.e. features extracted. Brute force query time will be the same if any of the data structures are used. It will also be unaffected by the value of k.

C. KD Tree Approach

To overcome the inefficiencies of the brute-force approach, a type of tree-based data structures have been implemented. These structures try to reduce the required number of distance computations by efficiently encoding aggregate distance information for the sample. The basic idea is that if point A is very far from point B, and point B is very close to point C, then we know that points A and C are very far, without having to explicitly calculate their distance, thus if we eliminate B we can also do the same for C. In this way, the overall cost of a nearest neighbours search can be reduced to O[DNlog(N)] or better. This is a very significant improvement over brute-force for large dataset. KD tree approach query time will be large if value of k is also large.

D. Ball Tree Approach

Similarly like KD tree, Ball tree are also tree based data structures which were developed due to inefficiency of KD tree algorithm to work on vectors with higher dimensions. KD trees divides data along Cartesian axes, ball trees divides data in a series of nesting hyper-spheres. This makes ball tree construction more costly than that of the KD tree, but the results here are very efficient on highly structured data, even in very high dimensions. Ball tree query time is given by O[DNlog(N)]. Similar to KD tree ball tree's query time will be large if the value of k is large.

E. Proposed Work

The uploaded will be present in database, from the database we will retrieve the CV using textract which is a library in python. The CVs will be in different format like .doc or .pdf, by parsing the CVs we will convert to same format which will be easy to analyze. This parsing will be done using pattern3 which is a library in python. Each word of the CV will be separated and put in a list and also the count of each word will be stored. Stop words are words which have no use in the project. Hence the stop words will be removed The word embeddings will be made using the list from the preprocessing module. It is an algorithm which will match the company's requirement to the skills mentioned in the CV by the candidate. Based on the skills mentioned in the CV the candidate will be asked to give an online test. The scores are evaluated by the system itself. The top students are shortlisted for the interview process. Non Shortlisted students will be recommended some courses to improve their performance in the future. We are going to use software like Anaconda and programming languages like Python and machine learning. Libraries like textract, gensim, pattern3 etc are used.

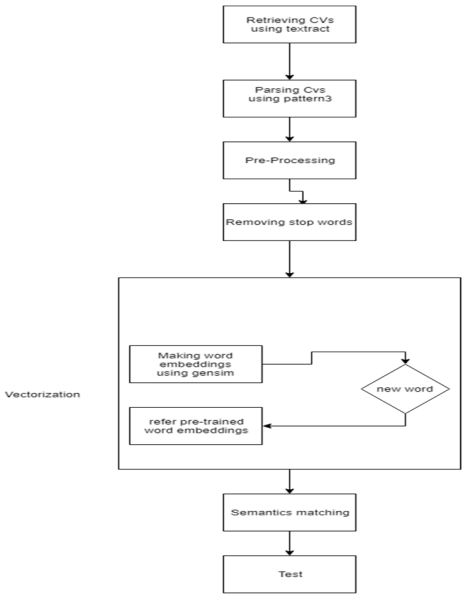
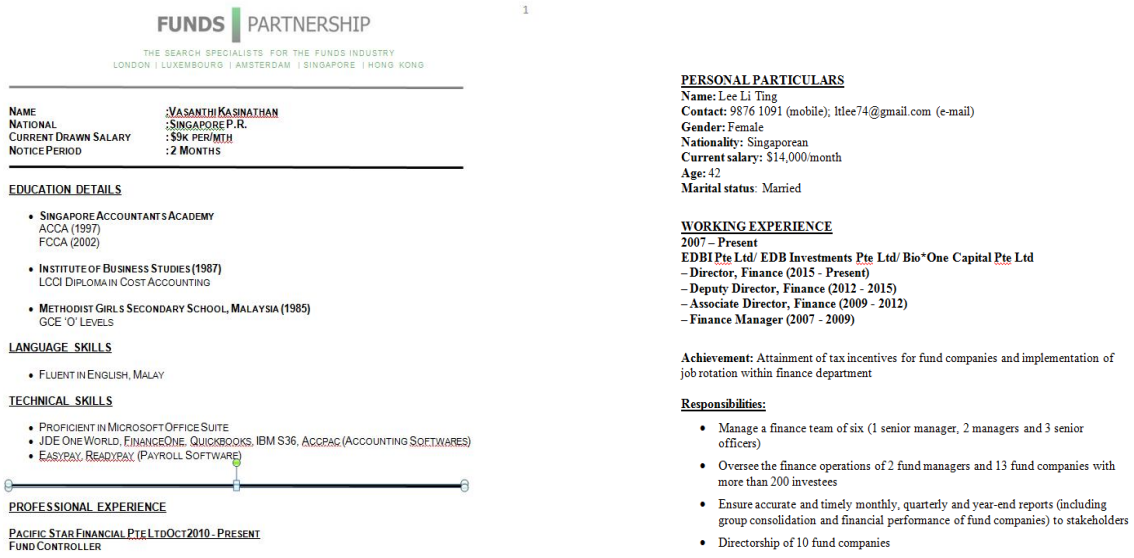


FIG.2: Work flow

©IJRASET: All Rights are Reserved

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

IV. RESULT AND DISCUSSION



FUNDS PARTNERSHIP

THE SEARCH SPECIALISTS FOR THE FUNDS INDUSTRY
LONDON | LUXEMBOURG | AMSTERDAM | SINGAPORE | HONG KONG

NAME :VASANTHI KASINATHAN
NATIONAL :SINGAPORE P.R.
CURRENT DRAWN SALARY :$5K PER MTH
NOTICE PERIOD :2 MONTHS

EDUCATION DETAILS

- SINGAPORE ACCOUNTANTS ACADEMY
  ACCA (1997)
  FCCA (2002)
- INSTITUTE OF BUSINESS STUDIES (1987)
  LCCI DIPLOMA IN COST ACCOUNTING
- METHODIST GIRLS SECONDARY SCHOOL, MALAYSIA (1985)
  GCE 'O' LEVELS

LANGUAGE SKILLS

- FLUENT IN ENGLISH, MALAY

TECHNICAL SKILLS

- PROFICIENT IN MICROSOFT OFFICE SUITE
- JDE ONE WORLD, FINANCEONE, QUICKBOOKS, IBM S36, ACCPAC (ACCOUNTING SOFTWARES)
- EASYPAY, READYPAY (PAYROLL SOFTWARE)

PROFESSIONAL EXPERIENCE

PACIFIC STAR FINANCIAL PTE LTD OCT2010 - PRESENT
FUND CONTROLLER

PERSONAL PARTICULARS
Name: Lee Li Ting
Contact: 9876 1091 (mobile); ltlee74@gmail.com (e-mail)
Gender: Female
Nationality: Singaporean
Current salary: $14,000/month
Age: 42
Marital status: Married

WORKING EXPERIENCE
2007 – Present
EDBI Pte Ltd/ EDB Investments Pte Ltd/ Bio*One Capital Pte Ltd
– Director, Finance (2015 - Present)
– Deputy Director, Finance (2012 - 2015)
– Associate Director, Finance (2009 - 2012)
– Finance Manager (2007 - 2009)

Achievement: Attainment of tax incentives for fund companies and implementation of job rotation within finance department

Responsibilities:

- Manage a finance team of six (1 senior manager, 2 managers and 3 senior officers)
- Oversee the finance operations of 2 fund managers and 13 fund companies with more than 200 investees
- Ensure accurate and timely monthly, quarterly and year-end reports (including group consolidation and financial performance of fund companies) to stakeholders
- Directorship of 10 fund companies

Fig.3: Training Data set

The end product will have the ability to analyze the CVs of every user and give suggestions to them regarding the skills they need to improve in order to be eligible for the companies. The product will be flexible as per the requirements of the users. Accuracy of our project is approximately 85%.

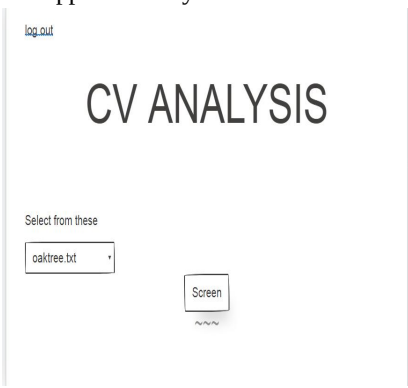


Fig. 4: Output 1



Fig. 5: Output 2

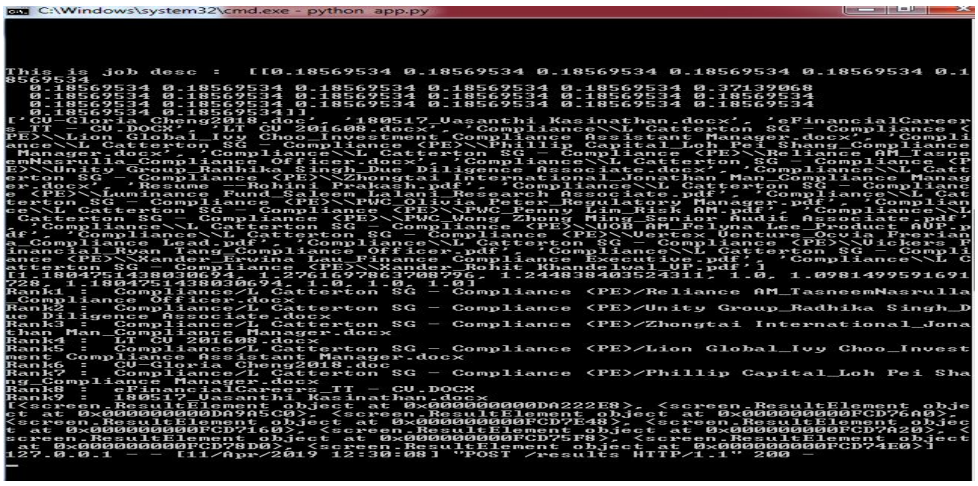


Fig. 6: Output 3

©IJRASET: All Rights are Reserved

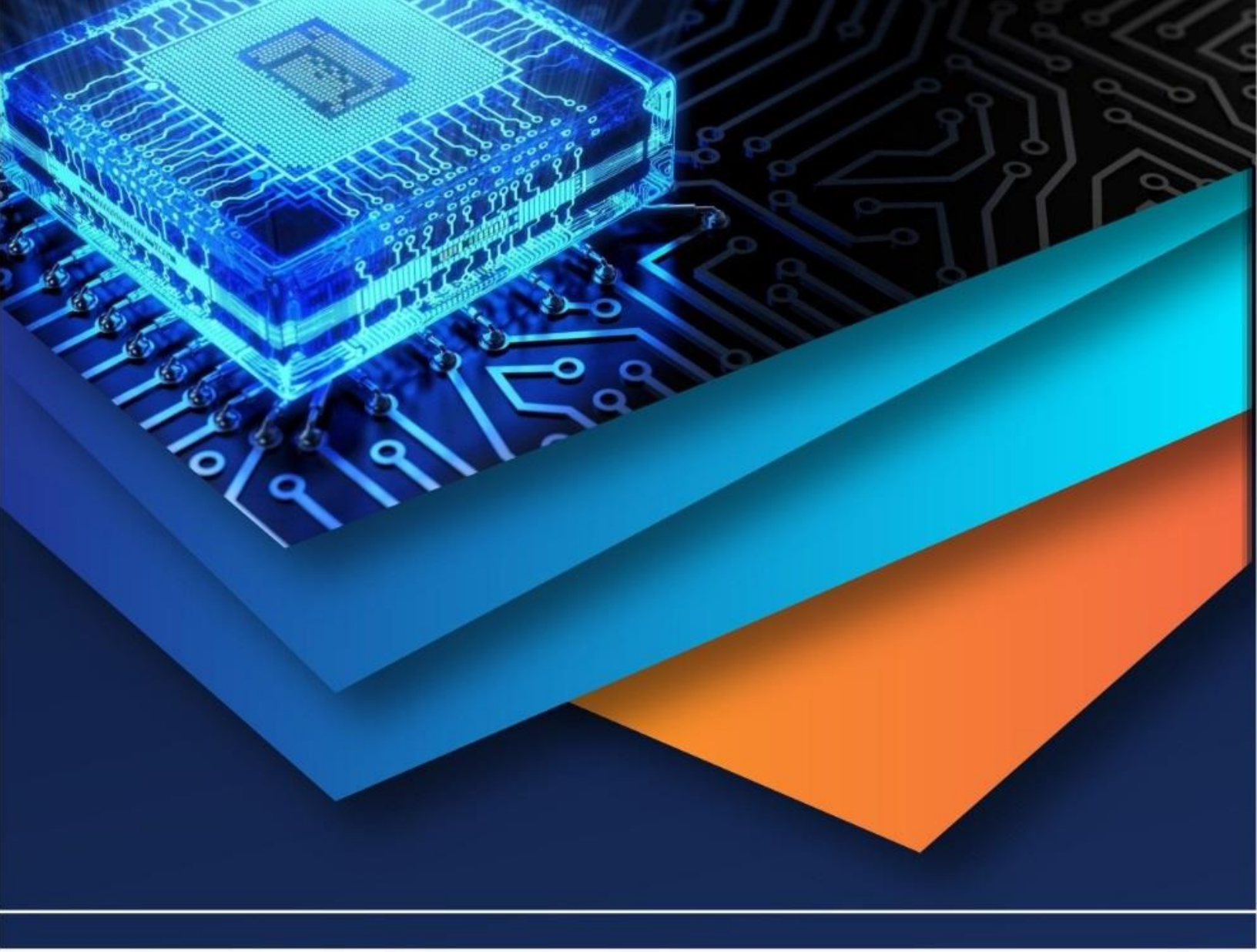
V. CONCLUSION AND FUTURE SCOPE

The end product will have the ability to accept or reject a job applicant based on two factors the requirements of the company should match the skills mentioned in the applicants CV and the other factor is the test evaluation which will based on the skills of the applicant which will ensure that the CVs uploaded by the applicant is genuine i.e. the applicant is really well versed in the skills mentioned in their CVs. The project will use machine learning and natural language processing algorithms like Naïve Bayes classifier which will ensure text categorization and the skills from the CVs of the applicant will be added to the database based on which a test will be given to the applicant. The project also will recommend some courses to non shortlisted applicants based on the test evaluation score. Hence the project will not only benefit the company recruiters but also the applicants as they will be recommend courses if they are lacking any technical skills which will be based on the skills mentioned in the CVs and also the test evaluation score.

REFERENCES

[1] Evanthia Faliagka, Giannis Tzimas ,” Application of Machine Learning Algorithms to an online Recruitment System” ,ICIW 2012.
[2] Ramjeet Singh Yadav, A.K. Soni, “A study of academic performance evaluation using Fuzzy Logic techniques”,March 2014.
[3] M.Mochol, H.Wache, and L.Nixon, “Improving the Accuracy of Job Search with Semantic Techniques”, Business Information Systems,vol. 4439, 2007,pp.301-313.
[4] S.Amdouni and W. Ben Abdessalem Karaa, "Web - based recruiting", Proc.Of International Conference on Computer Systems and Applications (AICCSA) ,2010,pp.1-7.
[5] E.Faliagka, K.Ramantas, A.Tsakalidis, M.Viennas, E.Kafeza and G.Tzimas,”An Integrated e-Recruitment System for CV Ranking based on AHP” ,”Proc.of WEBIST 2011,May. 2011, pp. 147 -150.
[6] Available at https://www.coursera.org/learn/machine-learning, accessed on (25th September 2018)
[7] Available at https://www.udemy.com/python-for-data-science-and-machine-learning-bootcamp, accessed on (4th October 2018)
[8] Available at https://towardsdatascience.com/build-a-handwritten-text-recognition-system-using-tensorflow-2326a3487cd5, accessed on (20th September 2018)

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)