iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** X **Month of publication:** October 2022

DOI: https://doi.org/10.22214/ijraset.2022.47201

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: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

Job Role and Personality Prediction Using CV and Text Analysis

Muskan Goyal[1], Shrey Shah[2], Aakash Sangani[3], Bhoomika Valani[4], Neha Ram[5]
[1, 2, 3, 4, 5]*Department of Information Technology, Dwarkadas Jamnadas Sanghvi College Of Engineering, Mumbai, India*

Abstract: ***As With the rapid change in technology, there has been an equally rapid change in the employment trends. Along with the rate of increasing literacy there's a high number of potential candidates for any job role. Thus, whenever a job posting is made public it tends to attract a lot of attention thus resulting in many responses. This creates a huge workload on the HR department of any company responsible for the hiring process. Our project aims to help solve this problem by proposing the following solution. The HR department can easily filter out candidates based on their performance in the tests set and the cv ranking in accordance with the job posted. The user will be classified into one of the 16 MTBI personality types. This personality classification of the user will be done on the basis of the data found on their social media handles and hence the MTBI personality types of the user will be given according to the probabilities. With this tool the HR can easily shortlist the more suitable candidates easing their selection process and avoid the large amount of time they would have otherwise taken going through the tedious process of reviewing each candidate's resume and then interviewing them to gauge their personality. The HR department will be able to post any job openings they have on our website where the candidates will apply for them. The HR can also choose to select or reject candidates through our system and the respective message would be sent to the candidate.***
Keywords: Jobs, Machine Learning, Personality, Twitter, Sentiment Analysis, Natural Language Processing.

I. INTRODUCTION

Here we'll discuss the objective of this project – The aims that were kept in mind while designing this project. We also discuss the major challenges we had faced with respect to the design of the projects and a brief overview of this report – what the user will have understood having read this report.

A. Motivation/Objective

The main motivation behind the project is to automate and ease the hiring process reducing the heavy workload the HR department faces during the hiring drive. It is also helpful for the candidate as it can act as a streamlined interface between candidate and HR. The main purpose of our project is to develop a portal which can be mainstream for both the candidate and HR and automate and optimize the entire hiring process. The recruiter will be able to post the job requirement and job description and the candidate can view the job details and choose to apply for it by giving an optional test and providing their details.

B. Major Challenges

1) Finding an accurate resume dataset to prevent overfitting for the machine learning model.
2) Predicting social media using a natural language process like BERT can be time consuming.
3) Finding ways to predict more than one job role for a given CV by a candidate.

II. LITERATURE REVIEW

A. Existing Systems

1) *Using CV Ranking Policy:* It is used to shortlist candidates by sorting them such that the ones having desirable features such as a certain skillset are ranked higher than those not.
2) *Myers Briggs Psychometric Analysis:* It's a decentralized ride sharing application with virtually no middleman fees, and trustless smart contracts to manage liability for the workers as well as the riders, all done autonomously. It uses the IPFS protocol to store all the user data so that it is decentralized and thus prevents any external sources from accessing the private user information.
3) *OCEAN Model (also known as Five Factor Model):* It has been used to predict the personality of the candidate which includes Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [2,7].

4) *Rorschach Test:* This is the famous test where inkblots are shown and how the user perceives them, they are judged and their personality is determined [5].
5) *MBTI Test:* This test determines how people make decisions in various circumstances and judges them accordingly [5].
6) *Using social Media:* In the evolving times, as people become more and more connected globally with the help of social media it only makes sense to judge them using the same [6, 7 ,8].
7) Using Mobile Metadata: Device metadata such as duration of voice calls, no. of messages sent, duration of Bluetooth usage, and many more have been used to group people into clusters hence, making it easier for applications taking this data to analyse their user base better and serve them better [9].

B. Existing Algorithms/Approaches

1) *Using Analytical Hierarchy Process (AHP):* It is the same as ranking CVs by using various factors for selecting an individual candidate and ranking them accordingly [1].
2) *Using Recommendation System:* The It usually uses a number of filters to get to the result. There are two types namely content based and collaborative based recommender each having its own set of advantages and drawbacks [2,8,9,10].
3) *Using ML in social media and Smartphones:* In an esteemed review paper, it was suggested to use a flexible Machine Learning algorithm as unregularized linear models will quickly reach their limits due to factors such as multicollinearity, it also helps to use interpretable models so as to detect the causes behind any bias in the model or to keep it as flexible as possible keeping future feature engineering ramifications in mind. Keeping flexibility in mind, the machine learning approach at the personality detection problem usually comes under:

a) Data Sources

- Smartphones
- Social Media

b) Data Preparation techniques used:

- Data Cleaning
- Feature Engineering

c) *Data Ethics to keep in mind while handling such sensitive data:*

- Privacy protection
- Open Data, Materials, and Code [10]
- Researcher Degrees of Freedom [11]

Techniques such as nested resampling along with model optimization techniques are used to bring the machine learning model to a decent level of accuracy. Resampling methods like K-fold Cross validation and optimization techniques such as hyperparameter tuning, dimensionality reduction, and variable selection are used to find a high correlation to predict personality of a person better. During evaluation of a model, one has to keep in mind to not remove features indicating a real-world bias unless and until it reflects in the outcome of the models' predictions making it susceptible to real world bias. For e.g., the person's gender is to be taken into consideration unless and until the model doesn't adopt any real-world biases from the data.
We should be careful with the bias we use to prevent any form of overfitting and should not be unidimensional.

4) *Using ML in social media and Smartphones:* Many NLP models such as BERT, RoBERTa are used to evaluate the personality type and they also are used after combining various sources of social media sites to get text required to predict.
5) *Using TF-IDF Algorithm:* It calculates the number of words present in that sentence then divides it by the number of times it is present in the entire document to determine its weightage in determining the class it belongs to, thus the predictability.
6) *Using Boyer Moore Algorithm:* The relevant words are checked using Boyer Moore Algorithm. Various algorithms such as Rabin Karp, Knutt Morris Pratt have been studied and they have high time complexity. The Boyer-Moore Algorithm is used wherein pattern matching is done from right rather than left [4].
7) *Using Logistic Regression:* It is similar to Linear Regression; however, it only gives a binary input either 1 or 0 / yes or no.
8) *Using Naive Bayes:* It is used in calculating conditional probability for both event A and event B to occur. In code it is usually computing logistical probability and hence has to be normalised for further use.
9) *Using kNN:* It is mainly used to solve classification problems but can also solve regression problems.
10) *Using SVM:* It is a model used for regression in ML. It classifies points by finding a hyperplane in a multidimensional space

11) *Using Random Forest:* It is a collection of many decision trees helping achieve better accuracy in some cases.
12) *By using Decision Tree:* Here the nodes are treated as the results and the branches are traversed based on the probability [13].
13) *Gradient Boosting:* It is a method based on regression. It produces a collection of weak predictive models. [8]. To predict personality from Facebook posts and statuses, we use ML methods such as Linear Regression and Support Vector Regression. However, these methods have very low accuracy. LIWC and DLA is used for the textual extraction. [13].
14) *Using XGBoost Classifier:* It is an improvement on the gradient boosted decision trees for higher efficiency and accuracy. It usually outperforms all the other algorithms in terms of accuracy in the right context. This used as a main classifier along with the other classification methods mentioned above gives the best accuracy so far.
15) *Long Short-Term Memory (LSTM):* They are a form of Recurrent neural Networks. They have a vast number of layers and are used to analyse the text data present.

a) Convolution Layer-We use CNNs convolution neural networks for performing feature extraction. It usually consists of a convolution and max pooling layer.

b) Classification- For this we use the perceptron learning network which consists of two parts, one being the fully connected layer which has a size of 200 and a SoftMax layer of size 2 to determine if or not it falls into that category.

16) *Using Neuro Fuzzy Systems:* Fuzzy systems are used when each object is given a certain degree of association and is then evaluated with the fuzzy rule set. They are also known as fuzzy neural networks. [14].
17) *Using LIWC tool for Text Analysis:* Linguistic Inquiry and Word Count helps to analyse if the text belongs to one of the 80 linguistic, psychological and topical categories which massively affect your personality. LIWC can be used for sentiment analysis [15].
18) *Using DLA for Text Analysis:* Known as Diffusion Limited Aggregation. It is a clustering technique using Brownian motion cluster by particles walking randomly. [15].
19) *Using Word Cloud Method:* It is a visual method for displaying the more frequent/ important words.

C. Literature Related to Technologies/ Tools/ Frameworks

For analysing the CV, ML Techniques have been used. Even Career mapper is used which checks the completeness of the profile.
The approach using the TF-IDF algorithm is performed by MSSQL SSM. The hardware requirements are Pentium 4 Processor and RAM 512 MB or above.[4] The system will run on any operating system having a web browser. – The websites are usually designed using HTML and CSS while Databases used are MySQL and PHP, we perform validation using JS, we would need modern browsers such as Internet Explorer later than 7.0, Firefox later than 3.5 and Google chrome versions later than 1.0.
For parsing CVs to extract useful information NLP libraries of python are used such as pyresparser, NLTK, TextBlob. A tool used for handling massive amounts of textual data and performing advance NLP is SpaCY.
One of the tech stacks used in a previous project is Python in Django Framework along with both machine learning and deep learning algorithms by integrating the TensorFlow Library.

D. Observation on Existing Work

While Random Forest forecasts maximum accuracy of 71 percent, SVM gives an accuracy of just 51.8 percent overall. While the results of random forest are barely crossing minimum accuracy threshold, we can use deep learning to achieve an accuracy of 80 percent overall. All the systems considered have not taken the personality of the candidate into account and use comparative strategies to compare the various resumes. The proposed system aims to solve the problem by analysing the personality through social media and provide standalone results of predicting the job role. The graphs shown in personality helps the HR to understand the probability of each personality type (out of 16) of the candidate.

III.PROPOSED METHODOLOGY AND APPROACH

Our problem definition is as follows: In this global pandemic, the job market has become abundant with human resources making it easier to find skilful people. Human resources are not just meant to fill skill gaps in an organisation but they're also meant to fit the team that they work with for a long term and fulfilling role. These unspoken qualities that recruiters look for to find the perfect fit for an organisation were already tough to evaluate in-person, making it even more challenging in today's work-from-home era. It is not feasible to judge a person's personality within a few minutes of the interview. Hence, social media of applicants could be used to evaluate the personality of the candidate to see if they'd be a right fit for the organisation or not. Along with their social media, CVs could indicate that the skill requirement is met.

A. Note on the Scope of Implementation

Personality of person plays a vital role in both personal as well as professional life.

Our model can be used by:

1) Personality Prediction using CV analysis and text will prove to be a boon to the HR departments as the candidates will be pre shortlisted according to their personalities thus providing more accurate selection results.
2) Apart from the HR department, college clubs and various admission institutes, business schools and universities can use this model on student essays/ applications / SOPs to simplify the shortlisting process.
3) It can also be used for matching profiles of like minded people or people suited for the particular jobs etc.
4) Various data driven insights about people can help us formulate a trend in hiring strategy which can further boost greater profits in the industry.

B. Some Assumptions and Constraints

1) The CVs the candidates have uploaded are completely factual and they haven't lied on their resumes.
2) The candidate is an active user of social media.
3) The candidate is used to using online job portals.

C. Proposed Approach to build Job Role and Personality Prediction using CV and Text Analysis

1) Feature of Proposed System:
a) Login system for both company representatives and candidates where they can store information about themselves (about, educational qualifications, etc.)
b) Company representatives/ HR will be able to make job posting publics where they can ask for the candidate's CV, social media and can also ask each candidate to give an optional test
c) Candidates can apply for jobs view their status after they have applied
d) Company representatives can view the response and choose to accept or reject the candidate

2) Tools used for data collection, size of sample and limitations:
a) We pre-processed Kaggle datasets to achieve clean resume texts and their job roles to train the machine learning model used.
b) We also collected data from surveys regarding the need for proposed system
c) We retrieve data from Twitter through the tweepy API which is later used in BERT

D. Benefits of Proposed Implementation

1) It will provide HR with a burden reddening task of assessing all the candidates and their resumes individually
2) It will give the candidate an open interface between the HR and them
3) It will help the candidate assess what job roles are suited for them
4) It will help HR assess their aptitude for the same

IV. PROJECT RESOURCES

A. Hardware Requirements

1) *Processor:* Pentium Processor and later
2) *RAM:* More than 512 MB

B. Software Requirements

We would need modern browsers such as Internet Explorer later than 7.0, Firefox later than 3.5 and Google chrome versions later than 1.0.

C. Operating Requirements

The project is functional on Windows, Linux and Macintosh

V. SYSTEM ARCHITECTURE

Principal components of our project are as follows:

A. On HR Side

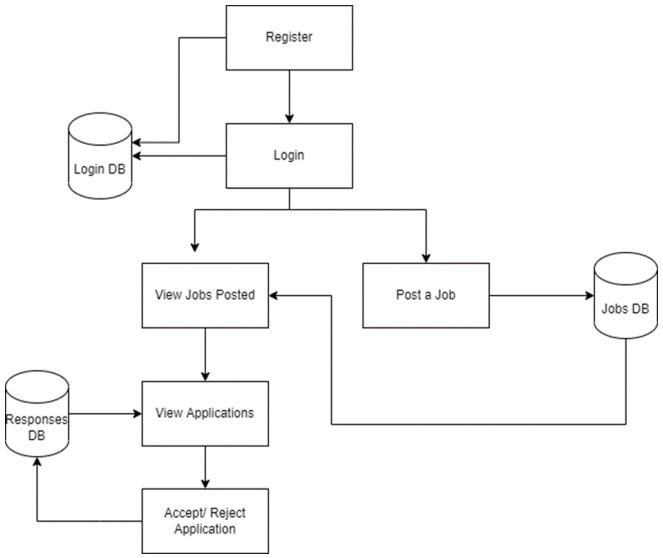


Fig. 1 HR Side

- Login and Registration: The HR admin will be able to register their company and then login
- The HR admin will be provided with an option to post a job as soon as they login on the navigation bar after which they will be asked the job details and to fill the questions and options for the optional tests
- After this the job posting will be made public
- They can view their response for each job posted on their dashboard and can accept or reject candidates respectively

B. On Candidate Side

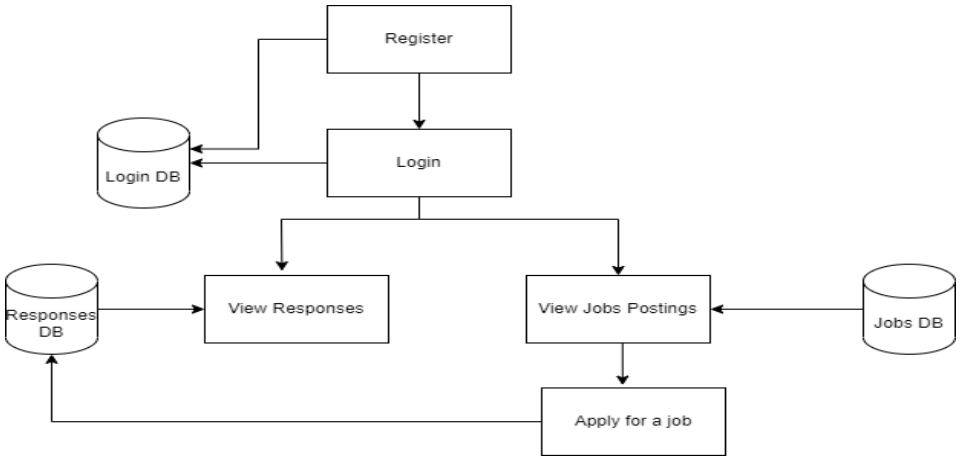


Fig. 2 Candidate Side

- Login and Registration: The candidate will be able to register with their educational qualifications and then login.
- They will be directed to a dashboard of publicly posted jobs where they can read through and apply if interested.
- After applying they can view the jobs they have applied for and their status

C. Description of Datasets

- Updated Resume Dataset: Pre-processed to give us clean resume text with their respective job roles
- Personality Dataset: Pre-processed to assess one of the 16 personality types from clean text

VI. IMPLEMENTATION

A. Working of System

1) *Login and Registration:* Both the candidate and the company representatives are asked to register and login. While registering they are each asked their respective details and are allowed to set a password using their email as their username. The candidate is asked their educational qualifications while the company is asked about their company
2) *Navigation Bar:* The Navigation bar for both the candidates and HR is different. While on the HR side it displays the dashboard and an option to find candidates, on the candidate side it provides them with an option to apply for jobs along with the dashboard.
3) *Dashboard:* This too is different for both HR and Candidate. For the HR, it displays all the responses for the jobs they posted while on the candidate side it displays all the jobs that specific candidate has applied for
4) *Response:* This is the same for both HR and Candidate. It has multiple parts. It displays the Aptitude score (score on test) along with the personality. It then displays the best suited job roles along with a competitive analysis of the score of the best candidate

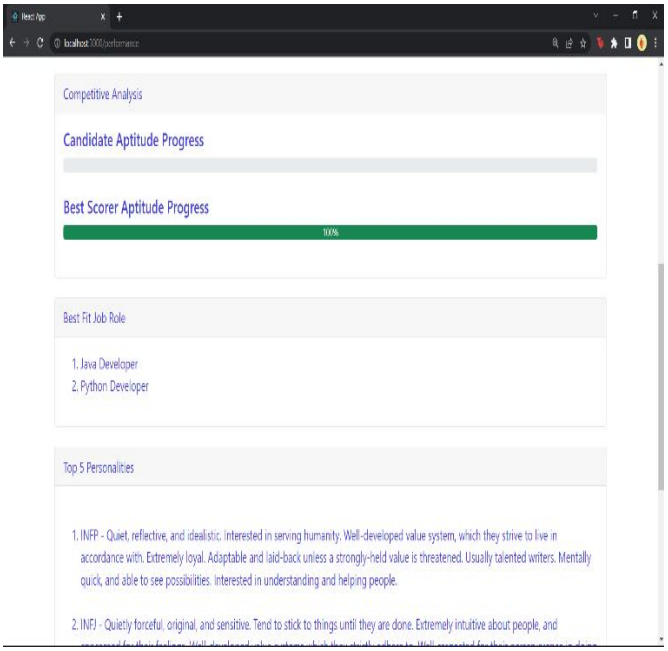
B. Algorithms used (Modular Description)

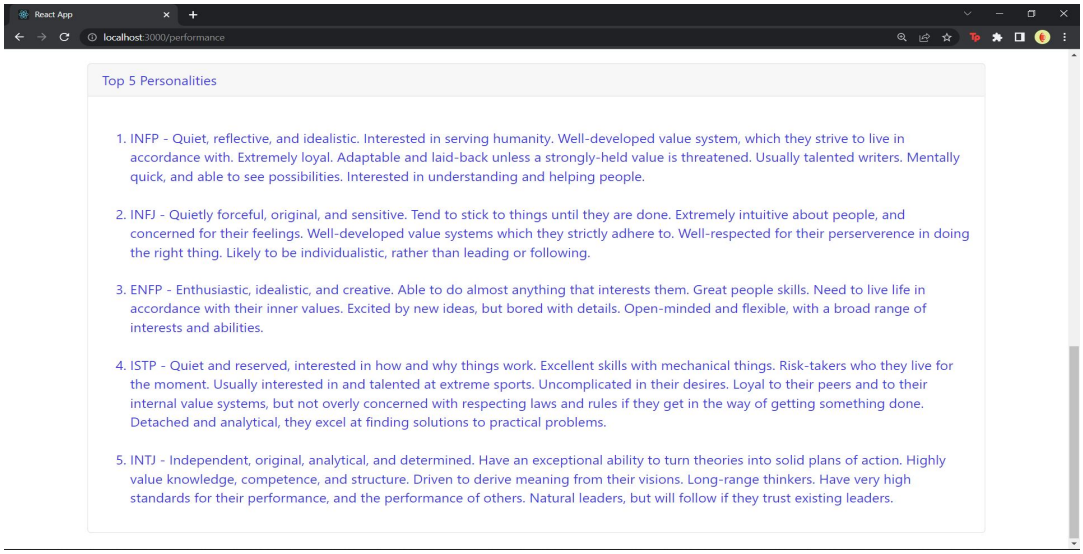
1) *Login and Registration:* Hashing techniques provided by Django are used to store the data in the database.
2) *Response:* In the social media section BERT is used using transformers to perform natural language processing to assess the probabilities of 16 personality types and display those in a pie chart using React charts. For assessing the best suited job roles, we use TF – IDF vectorizer along with Multinomial DB to produce the probabilities of different job roles. After this linear regression is performed to understand the threshold and recommend jobs above the threshold.

C. Tools Used For Our Implementation

1) *Front End:* React
2) *Back-end:* Django
3) *Database:* MySQL
4) *APIs:* Django Rest Framework, Tweepy
5) *Models:* BERT, TF – IDF Vectorizer with Multinomial Naïve Bayes, Linear Regression

VII. RESULT





The Result of the Application is shown below. The systems give an output of the best suited job for the candidate according to their resume and the probability of the candidate being a particular personality type out of the 16 MTBI personality types. The top 5 probabilities are given as list and their descriptions are given on top of the bar graph.

VIII. CONCLUSIONS

The employment patterns have changed at an equal rate as that of technological change. There are many potential candidates for every work position, and literacy rates are rising. As a result, anytime a job posting is made public, it tends to garner a lot of interest and, consequently, generate a lot of answers. This significantly increases the workload for any company's HR department, which is in charge of the hiring procedure. Our initiative intends to assist in the problem-solving process by analyzing candidates' social media handles in addition to their resumes to anticipate not only their job roles but also their personalities.

With the use of this technology, HR can speed up the selection process and save a significant amount of time by eliminating the need to read through each applicant's resume carefully before conducting an interview to determine their personalities. Any job openings that the HR department has can be posted on our website, where applicants can apply for them. Through our system, the HR can also decide to accept or reject applications, and the appropriate notification will be forwarded to the applicant.

This document comprehensively covers the various existing systems, their features. We then proceed to propose our solution with the help of use case diagrams etc. and then attach the photos for the actual interface. We also perform testing for the same and observe the results.

IX. FUTURE SCOPE

This project can be further extended to add more features such as enabling the company to specify skills required by entering keywords and only parsing the candidates that have the same It can further be optimized to increase the speed of BERT to improve the efficiency of the entire system and can assess social media in real time. Further it can be streamlined into the recruiting process where it becomes a universal tool used by all companies and candidates for the job seeking process.

X. ACKNOWLEDGMENT

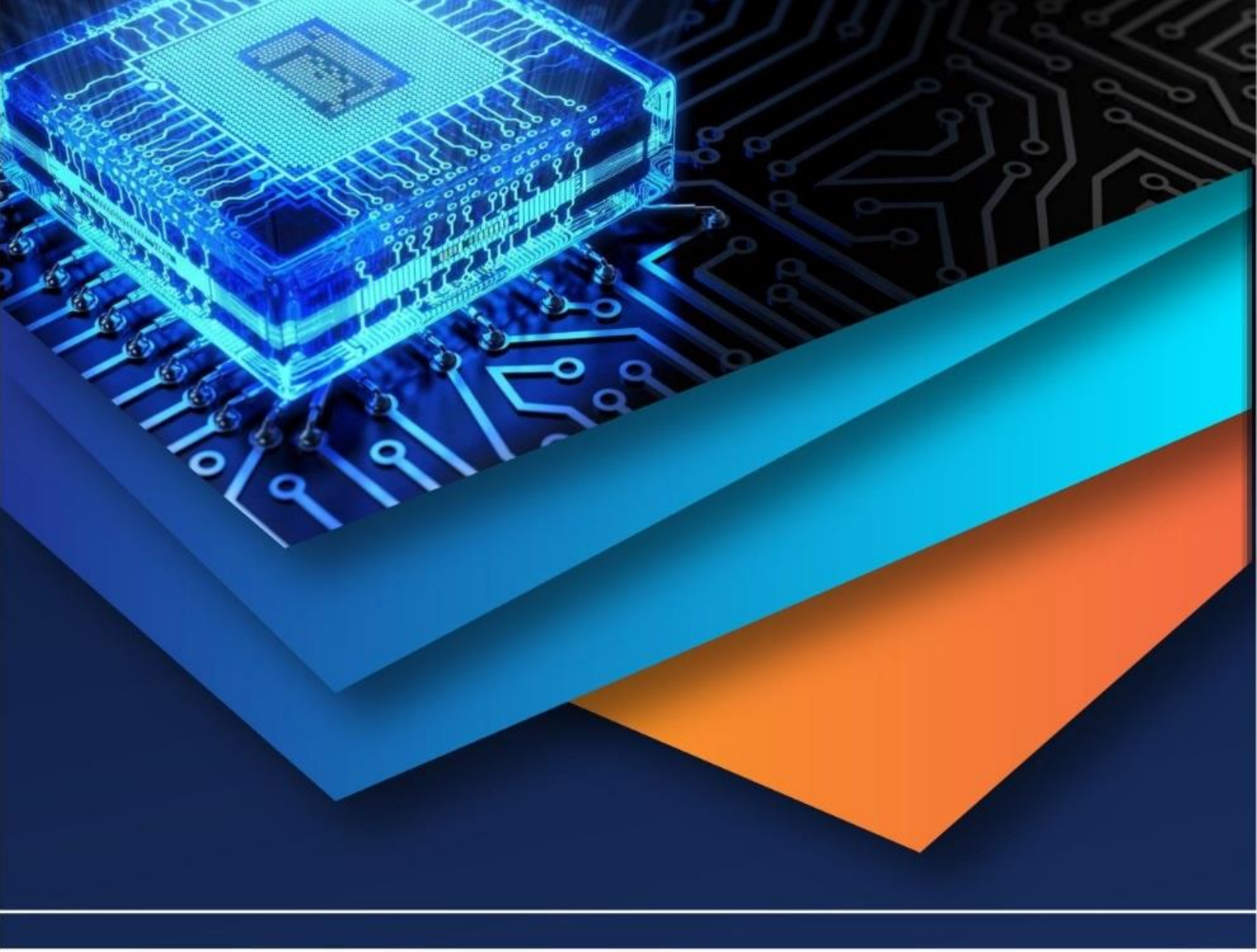
We would like to express our heartfelt gratitude to Prof. Neha Agarwal for guiding us and mentoring us on each step of the way throughout the development of this project. Her inputs were very insightful and greatly helped us in making our project more efficient. We would also like to take this opportunity to thank all the faculty members including the HOD, Dr. Vinaya Sawant in the department who were always approachable for any queries we had throughout the course of this project.

Lastly, we would like to thank Dr. Hari Vasudevan for incorporating this subject into the syllabus and encouraging us to think outside the box and truly innovate and make some meaningful contribution in technology. Similarly, we express sincere gratitude to all the staff in the college who provided us with necessary resources and enriched our project with valuable suggestions.

We express our sincere and heartfelt gratitude to all individuals without whose help it would not have been possible for us to complete.

REFERENCES

[1] FALIAGKA, EVANTHIA ET AL. “AN INTEGRATED E-RECRUITMENT SYSTEM FOR AUTOMATED PERSONALITY MINING AND APPLICANT RANKING.” INTERNET RES. 22 (2012): 551-568.
[2] LINDEN, DIMITRI VAN DER ET AL. “THE GENERAL FACTOR OF PERSONALITY: A META-ANALYSIS OF BIG FIVE INTERCORRELATIONS AND A CRITERION-RELATED VALIDITY STUDY.” JOURNAL OF RESEARCH IN PERSONALITY 44 (2010): 315-327.
[3] ROUT, JAYA K. ET AL. “PERSONALITY EVALUATION AND CV ANALYSIS USING MACHINE LEARNING ALGORITHM.” INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING (2019): N. PAG.
[4] NARWADE, RUTUJA, ET AL. “PERSONALITY PREDICTION WITH CV ANALYSIS.” INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING AND TECHNOLOGY (IRJET), VOL. 09, NO. 04, APR. 2022, PP. 3220–3225.
[5] ATHARVA KULKARNI, TANUJ SHANKARWAR, SIDDHARTH THORAT, 2021, PERSONALITY PREDICTION VIA CV ANALYSIS USING MACHINE LEARNING, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) VOLUME 10, ISSUE 09 (SEPTEMBER 2021),
[6] THE INTERNATIONAL JOURNAL OF FUZZY LOGIC AND INTELLIGENT SYSTEMS (IJFIS.ORG)
[7] CHRISTIAN, H., SUHARTONO, D., CHOWANDA, A. ET AL. TEXT BASED PERSONALITY PREDICTION FROM MULTIPLE SOCIAL MEDIA DATA SOURCES USING PRE-TRAINED LANGUAGE MODEL AND MODEL AVERAGING. J BIG DATA 8, 68 (2021). HTTPS://DOI.ORG/10.1186/S40537-021-00459-1
[8] TADESSE, MICHAEL M. ET AL. “PERSONALITY PREDICTIONS BASED ON USER BEHAVIOR ON THE FACEBOOK SOCIAL MEDIA PLATFORM.” IEEE ACCESS 6 (2018): 61959-61969.
[9] STACHL, C., PARGENT, F., HILBERT, S., HARARI, G. M., SCHOEDEL, R., VAID, S., GOSLING, S. D., AND BÜHNER, M. (2020) PERSONALITY RESEARCH AND ASSESSMENT IN THE ERA OF MACHINE LEARNING. EUR. J. PERS., 34: 613– 631. ONLINE "HTTPS://DOI.ORG/10.1002/PER.2257."
[10] HTTPS://SUBJECTGUIDES.YORK.AC.UK/OPENRESEARCH/DATA_CODE
[11] HTTPS://RESEARCHER-HELP.PROLIFIC.CO/HC/EN-GB/ARTICLES/360009501593-RESEARCHER-DEGREES-OF-FREEDOM
[12] HTTPS://WWW.GEEKSFORGEEKS.ORG/OVERVIEW-OF-PERSONALITY-PREDICTION-PROJECT-USING-ML/
[13] THAHIRA M, MUBEENA A K, 2021, COMPARATIVE STUDY OF PERSONALITY PREDICTION FROM SOCIAL MEDIA BY USING MACHINE LEARNING AND DEEP LEARNING METHOD, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) ICCIDT – 2021 (VOLUME 09 – ISSUE 07),
[14] HTTPS://NEVONPROJECTS.COM/PERSONALITY-PREDICTION-SYSTEM-THROUGH-CV-ANALYSIS/
[15] HTTP://WWW.CS.ALBANY.EDU/~PATREY/ICSI660445/PROJECT/SURVEY_SAMPLE_REPORT.PDF
[16] ZULKIFLEY, NOR & RAHMAN, SHUZLINA & NOR HASBIAH, UBAIDULLAH & IBRAHIM, ISMAIL. (2020). HOUSE PRICE PREDICTION USING A MACHINE LEARNING MODEL: A SURVEY OF LITERATURE. INTERNATIONAL JOURNAL OF MODERN EDUCATION AND COMPUTER SCIENCE. 12. 46-54. 10.5815/IJMECS.2020.06.04.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY