JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Predictive Modelling of Personality Traits through Curriculum Vitae Analysis using Machine Learning

[1]Mr. Suraj Kumar B.P., [2]Adarsh Kumar Mall, [3]Aditya Kishore, [4]Avinash Chandra, [5]Ayush Srivastava

[1]Guide, [2]Student, [3] Student, [4] Student, [5]Student
[1]Dept. Of Computer Science and Engineering,
[1]SMVIT, Bengaluru, India

Abstract: Today, the business sector does not consider the potential employee's skills, but rather their characteristics. People are what shape professional, but also personal lives. So the recruiter needs to refine the employee's attitude. With the number of job seekers on the rise but the number of jobs continuing to decrease, it is very difficult to hand-list the best candidate for a suitable job by looking at their CV. This article attempts to explore different machine learning approaches to accurately predict human behavior through reanalysis, and concepts such as natural language processing (NLP). The Output shows that the Random Forest algorithm have better precision compared to other algorithms such as kNN, Logistic Regression, SVM and Naive Bayes. The template searches for the characters and other details of the candidates, such as skills, experience, etc. Using this system, organizations can filter the applicant pool and reduce the workload of a company's recruitment department.

I. INTRODUCTION

One of the most essential factors in determining an employee's long-term involvement in an organization is their personality. We understand the strengths of the candidate and analyze if he can influence and communicate well with others, which is beneficial to the development of an organization. Whenever there is a need for human resources in the company, they receive thousands of applications for job postings, because it is difficult for the recruiters to go to many recruitments and find the candidate is eligible for the interest using traditional methods which includes technical tests, interviews and group discussions. Then, from the first round, they analyze the candidates based on different factors, such as their suitability for the job, ability, CV that is not relevant and skills. Therefore, in order to minimize the complexity of the hiring process, we propose a new method, which will make the selection and pre-selection of candidates easier. This is using human prediction. For human prediction, we use a ML algorithm called logistic regression. Personality is one of the important factors that determine how well a person can perform in a given job. Analysis and understanding of the health of an organization is therefore essential. Our intention with this project is to make the machine more human and analyze the candidate like a human reviewer. This article attempts to explore and implement different machine learning algorithms and analyze which one is the most appropriate and the range of data provided. We also try to visualize the data and connect the various features

1.1 BIG FIVE TEST

The Big Five personality test, known as the OCEAN model, analyzes various personality traits based on five dimensions: openness (O), conscientiousness (C), extraversion (E), agreeableness (A), natural reaction (N). Each character represents a different type of person.

- Openness: This quality has acceptance, imagination, and curiosity.
- Conscientiousness: It talks about a high amount of thoughtfulness, a goal-oriented attitude and good decision-making characteristics.
- Extraversion: This also means extroversion, which is identified by excitement, talkativeness and assertiveness.
- Agreeableness: This term refers to features such as trust, affection and social behavior of an individual.
- Natural Reaction: This term includes attributes like sadness, moodiness and sudden burst of emotions.

Based on the result of each domain, we will determine the personality of a person, i.e., serious, extraverted, lively, dependable, responsible.

1.2 Generating Resume Score on the basis of Technical Skills

Technical skills scoring is a way to assess a candidate's skills and suitability for a job. This process analyzes the CV to identify the important technical skills mentioned by the applicant. These skills have technical tools skills, programming language skills, databases skills etc.

When we figure out the skills needed for a job, we often give each skill a weight based on how important it is. For example, some skills might be very important, while others are not so important. By having these skills different weights, we make sure that the really important ones count for more in the final score. Next, each skill is counted based on various factors such as depth of knowledge, years of experience, and relevant certifications or programs mentioned in the statement. This assessment allows the applicant's level of expertise in each skill to be calculated.

After assessing the individual skills, the scores are aggregated to produce a total resume score, which reflects the depth of the candidate's technical skills. This score is a valuable tool for recruiters and hiring managers, who can quickly identify the best candidates with the technical skills needed for the position. In addition, CV benchmarks can be refined by introducing machine learning algorithms that analyze patterns between CVs and jobs to continuously improve the benchmarking process and improve the rights of candidates.

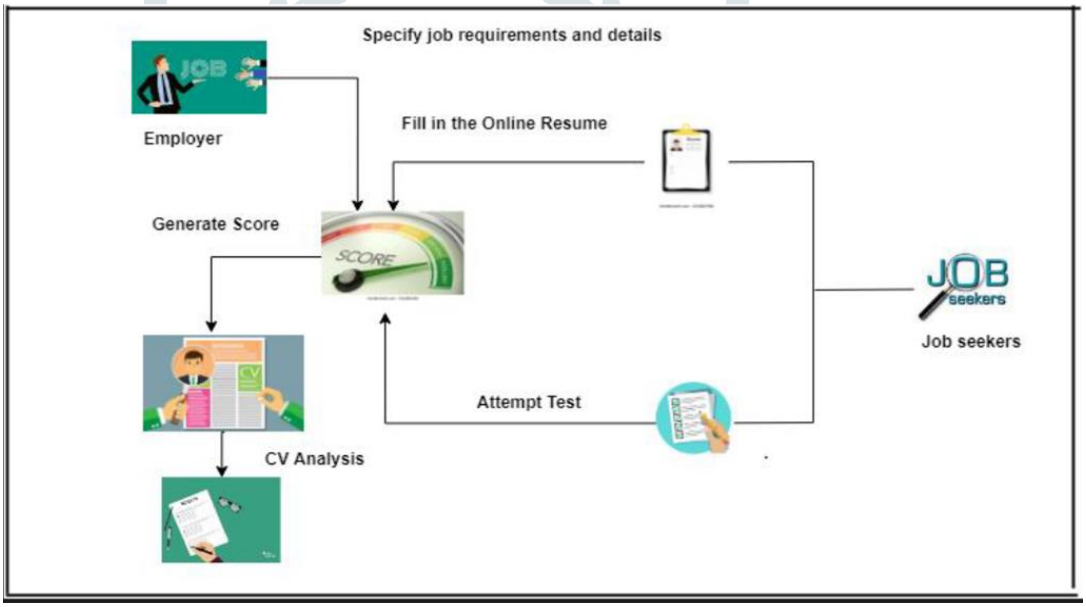


Fig: Process Architecture

2. RELATED WORK

Kalghatgi et al. [2] They presented a neural network method based on the Big Five test to predict people's personality based on their tweets posted on Twitter and extract meta-features from tweets. It is used to analyze human behavior. The authors followed four steps: collecting data from tweets, processing, organizing, and sorting. Neural networks are used to uncover patterns, this technique has limitations such as non-fake data, automatic analysis of tweets, and reliance on Twitter alone to predict not only human behavior but also user behavior.

Allan Robey et al [3] introduce a method to reduce the burden on companies' human resources departments with two parties: the organization and the candidates. The author says that the proposed system will be more efficient in selecting resumes from a huge collection for a more suitable and valid list. The main disparity between the current system and the proposed system is that the authors propose to conduct a qualification test with
personality tests for character prediction instead of just scanning the CV.

Juneja Afzal Ayub Zubeda and colleagues [4] developed a CV classification project using natural language processing and machine learning. The system organizes your resume just the way the company likes it. The authors suggest taking a look at your GitHub and

LinkedIn profiles too. This helps companies get a good feel for your skills, what you're capable of, and, most importantly, who you are as a person. It makes it easier for them to find someone who's the perfect fit.

Md Tanzim Reza and Md Sakib Zaman analyzed the CVs using Natural Language Processing and Machine Learning, creating the graduation and graduation stage by first converting the CVs to HTML and then redesigning them into the HTML code below. The model takes the data from CV and reduces it by values. They classified resumes using multiple regression models. However, the size of the dataset was too small. [5]

---

3. PROPOSED SYSTEM

3.1 Dataset

Because manual data collection is time-consuming, we collected candidate resumes from multiple websites and through personal interactions with recruiters; the total number was 708 resumes in PDF and DOCX formats.

3.2 Methodology

The purpose of our page is to determine a person's personality based on their openness, agreeableness, neuroticism, and conscientiousness scores. To achieve this, we need a way to calculate the scores on each CV. Our approach was to examine the entire profile and search for keywords related to the 'Big Five Test', as shown in Fig 1.

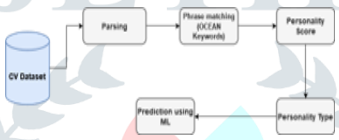


Fig: Working

When it comes to analyzing CVs, we used Pyresparser, a simple mining tool used to extract important elements from the CV such as name, email address, description and skills. Pyresparser can help in both PDF and DOCX files. The analyzed data is then saved in a CSV file.

Table 1: OCEAN Keywords

Openness	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Imaginative	Thoughtful	Cheerful	Trustworthy	Calm
Insightful	Goal-oriented	Sociable	Altruism	Strong hearted
Curious	Ambitious	Talkative	Kind	Collected
Creative	Organized	Assertive	Affectionate	Balanced
Outspoken	Mindful	Outgoing	Cooperative	Peaceful
Straightforward	Vigilant	Energetic	Empathetic	Tranquil
Direct	Control	Extroverted	Modest	Strong-willed
Receptive	Disciplined	Friendly	Sympathetic	Emotionally Stable
Open-minded	Reliable	Enthusiastic	Compliant	Serene
Adventurous	Responsible	Outspoken	Tender-mindedness	Resilient

Table 1 is given above for OCEAN keywords. Each element has a combination of 10 keywords associated with it. There are several natural language processing (NLP) libraries that can help us keep records, such as Natural Language Toolkit (NLTK), TextBlob, and SpaCY. We used SpaCY, an open source software library for natural language processing and processing large amounts of text data.

The PhraseMatcher class in spaCY is effective at matching large sets of symbols in text [7].

The keywords in Table 1 are related to the mentioned category. Our algorithm searches for keywords using the PhraseMatcher class and assigns a score from 0-10 based on the presence of OCEAN keywords in a person's resume. After assigning a score as shown in the table below, the algorithm labels each article as credible, resilient, interesting, responsible, or strong. Thus, we output a CSV file with 'Big Five' qualifiers in all columns. Each of the data points identified was labeled as reliable, overestimated, active, responsible, or critical, as displayed in Table 2 below.

Table 2: OCEAN Score and Personality Type

Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism	Personality
6	4	7	5	4	Extraverted
4	6	4	4	7	Serious
5	6	4	7	4	Lively
7	4	5	4	5	Dependable
5	7	6	6	3	Responsible

4. MODEL TRAINING AND TESTING

Before training our model, we coded the columnar structure of our data. Using the Sklearn library, we used 70% of our data for training and 30% for testing results. We are using various machine learning algorithms such as Logistic Regression, Random Forest, Support Vector Machine (SVM), Naïve Bayes and KNN to find the nature of the candidate.

- Logistic Regression

Logistic regression is a statistical method used for modeling the relationship between a categorical dependent variable and one or more independent variables. It's commonly used for binary classification problems, where the dependent variable has only two possible outcomes, such as "yes" or "no", "spam" or "not spam", etc.
The sigmoid function, also known as the logistic function, is used in logistic regression. Its formula is Y = 1 / (1 + e^(-m)), where m is the input value. The sigmoid function maps any input to a value between 0 and 1, helps in binary classification problems.

- Naive Bayes

In probability, Baye's theorem is used to calculate probabilities. The theorem forms the basis of the Naïve Bayes classifier, a classification algorithm that assumes strong independence between features. According to the algorithm, each feature of the problem contributes equally and independently to the solution.

- kNN

kNN stands for k-nearest neighbors, a machine learning algorithm that can solve regression problems. We can get an idea by comparing the proverb 'Birds with the same wings fly together' with kNN. The algorithm assumes that similar points are often located nearby.

- SVM

Support Vector machine is a supervised machine learning algorithm used for data processing and background analysis. The goal of SVM is to find a hyperplane that can easily classify data points in an N-level space (N number of features).

- Random Forest

Random forest is another method used for classification and regression. Multiple trees are used to produce the product. Clustering or bootstrap clustering is used to train the model.

After training our model on all algorithms, we discovered that our predictions were wrong. Even our best models may only be 30% accurate.

Another issue is that our training and testing statistics have different distributions. But if we put ourselves in an employer's shoes, we know that he or she would prefer to hire someone who is "responsible" and "healthy" above all else. So, our problem now becomes dual sorting problems (1- responsible or alive 0-other) [8]

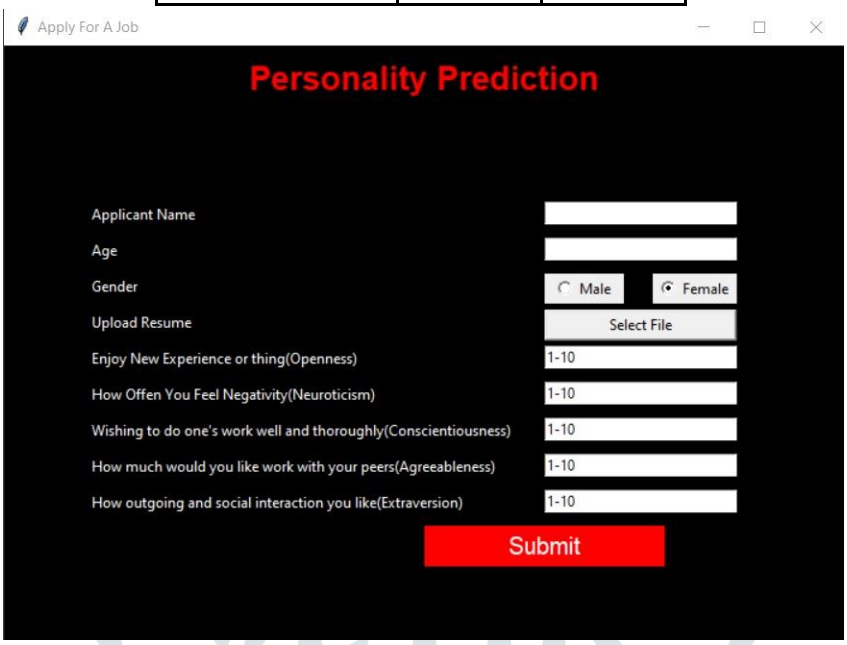
5. EXPERIMENTAL RESULTS

Now, after putting the data into our model we were able to increase the accuracy to 0.71. The normal forest algorithm gives the best accuracy, followed by Bayes, kNN, SVM, and logistic regression, as mentioned in Table 3. The forest also has the least mean square error, which measures the mean square of the difference between actual and predicted values.

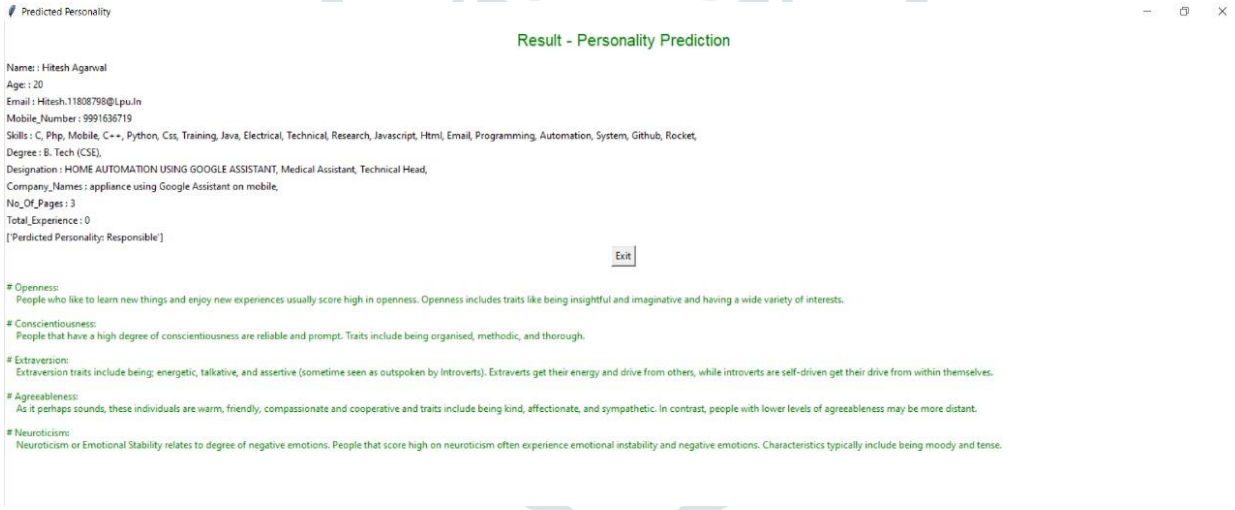
Model	Accuracy	MSE
Logistic Regression	0.62	0.37
Naive Bayes	0.65	0.37

Table 3: Accuracy and MSE Values

kNN	0.64	0.35
SVM	0.63	0.36
Random Forest	0.71	0.29



6. **OUTPUT**



Candidates should apply if they want to apply for any position. A pop up will appear and ask the candidate to fill in the details and submit their resume or CV. The candidate fills out an application, uploads a resume or curriculum vitae, and answers some background questions. The reference or CV is evaluated by the model and candidate features are predicted based on our algorithm which is displayed as output. The results obtained are as follows:

Fig: Interface Template

Fig: Result

---

7. CONCLUSION AND FUTURE SCOPE

In this paper, we have used multiple machine learning techniques like Logistic Regression, Random Forest, SVM, Naive Bayes, and KNN for human prediction using CV analysis. Thanks to Pyresparser, SpaCy and PhraseMatcher, we were able to predict the nature of different candidates.

Using Pyreparser we can easily identify the skills, designation and experience of the candidate The results shows that Random Forest has the highest accuracy of 0.71, but due to the lack of available data, it is less accurate than expected. Companies can use the proposed system to simplify the recruitment process by considering the characteristics of the possible candidates. Progress can also be made to enhance the efficiency and performance of the proposed system for more accurate prediction of individuals through CV analysis.

REFERENCES

[1] Mathuriya N, Bansal D, “An overview of the OCEAN Model”

[2] Anjana, R.: “Role of artificial intelligence in recruitment”

[3] A..Robey, K. Shukla, K. Agarwal, K. Joshi, Professor S. Joshi “Personality prediction system through CV Analysis, in IRJET vol 6, issue 02, February 2019.

[4] J. Zubeda, M. Shaheen, G. Narsayya Godavari, and S. Naseem “Resume Ranking using NLP and Machine Learning”

[5] Md Tanzim Reza, and Md. Sakib Zaman, “Analyzing CV/Resume using natural language processing and machine learning”

[6] https://towardsdatascience.com/

[7] https://spacy.io/

[8] https://www.kaggle.com/