Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

Personality Prediction Via CV Analysis using Machine Learning

Atharva Kulkarni[1], Tanuj Shankarwar[1], Siddharth Thorat[1]
[1]Student, Dept. of Computer Engineering,
Thakur College of Engineering and Technology,
Maharashtra, India

***Abstract* - The corporate world today does not focus just on the skills a potential employee possesses but also their personality. Personality is what helps one be successful in professional as well as personal life. Hence, the recruiter must be aware of the personality traits a person has. With an exponential increase in job seekers but a decrease in the number of jobs, it is difficult to manually shortlist the best fit candidate for a suitable job by looking at the CV. This paper attempts to examine different machine learning approaches forefficiently predicting personality through CV analysis using Natural Language Processing (NLP) techniques as well. Results show that the Random Forest algorithm achieved better accuracy when compared to other algorithms such as kNN, Logistic Regression, SVM and Naive Bayes.**

Keywords- Personality Prediction, CV, MachineLearning

1. INTRODUCTION

The word ‘Personality’ derives from the Latin word persona which refers to a mask worn by actors to act. However, Personality is much more than a mask now, it could possibly determine whether a person is suitable for a particular job profile. It tells us if a human is capable enough to lead, influence and communicate effectively with others. The first step of recruitment is the job application which consists of personal details, experience, and most importantly CV. Companies typically receive thousands of applications per job opening and have a dedicated team ofscreeners to select qualified candidates. It is very difficult for human beings to manually go through the CV of all applicants. Many candidates get filtered out in the first round itself on the basis of suitability, improper CV, not being skilled enough. Hiring the right candidate is a very difficult task as no candidate is perfect, some might not be skilled enough or some might not have the right personality. Hence, we propose a way in which the process of shortlisting gets streamlined and faster by personality prediction.

CV’s can reflect upon the professional qualifications of a person but do not reflect upon the personality of a person. Personality is one of the vital factors which suggests how a person would be able to work in a designated role, hence personality analysis and understanding is key. Our objective doing this project is to make the machine more human, and analyse the candidate in such a way that an actual human reviewer would.

This paper tries to explore and implement various machine learning algorithms and analyse which one among them provides the best accuracy with a wide array of data provided. We also attempt to visualise the data and form a connection between various factors.

2. BACKGROUND

There are various tests that help to determine personality types such as the Big Five, Rorschach test, and MBTI test. In this paper, prediction of personality is done by considering the Big Five Test as it is said to be widely used, reliable and also it has direct relation between performance mainly focusing on five cognitive domains [1].

2.1 Big Five Test

Big Five Personality test also known as OCEAN Model analyses personality types of individuals based on five dimensions - Openness(O), Conscientiousness (C), Extraversion (E), Agreeableness (A), Neuroticism(N). With each of the dimensions signifying a different personality type. It uses keywords to identify traits and analyze in which personality a person fit.

- Openness: As the word suggests, This quality features characteristics such as openness and imagination and curiosity.
- Conscientiousness: Conscientiousness talks about a high amount of thoughtfulness, a goal- oriented attitude and good decision-makers.
- Extraversion: Extraversion also means extroversion is identified by excitement, talkativeness and assertiveness.
- Agreeableness: Agreeableness refers to features such as trust, affection and social behaviour of an individual.
- Neuroticism: Neuroticism includes attributes like sadness, moodiness and sudden burst of emotions.

As these five dimensions cover almost all avenuesneeded to know someone it is the right method that forms the basis of a person’s overall personality.

3. RELATED WORK

Kalghatgi et al. [2] presented a Neural Network Approach based on the Big Five Test to predict the personality of individuals depending on tweets published on Twitter by extracting meta-attributes from tweets. Which are used to analyze one’s social behaviour. The authors followed a four-step process which is Data Collection from tweets, Preprocessing, Transformation and Classification. Although neural networks are used to predict personality there are limitations such as countering fake information, automatic analysis of tweets and relying on just Twitter is not enough to predict someone’s personality but only user behaviour and trends.

Allan Robey et al [3] proposed a system to reduce the load on the Human Resource department of companies by having two sides: organization and candidate oriented. The authors claim that the proposed system will be more effective to shortlist CVs from a large pool making sure that the ranking is fair and legal. The main difference between the existing system and the proposed system is that instead of just scanning the CVs, the authors propose to conduct an aptitude test and a personality test for personality prediction.

Juneja Afzal Ayub Zubeda et al [4] worked on a project to rank CVs using Natural Language Processing and Machine Learning. The system ranks CVs in any format according to the company’s criteria. The authors propose to consider candidate’s Github and LinkenIn profile as well to get a better understanding making it easier for the company to find a suitable match based on skillsets, ability and most importantly, personality.

Md. Tanzim Reza and Md. Sakib Zaman analyzed CV of individuals using Natural Language Processing and Machine Learning by first converting CVs to HTML and then reverse engineering to HTML code following which, segment finalization and qualification feature extraction has been done. The model extracts data from a CV and segments them based on the values. They have classified the CVs using multivariate logistic regression. However, the size of the dataset was very less.[5]

4. PROPOSED SYTEM

4.1 Dataset

As manual data collection is time-consuming, we collected candidate resumes through a lot of websites and personal interaction with potential job seekers taking the total count to 708 CVs. The collected CVs were in PDF and DOCx format.

4.2 Methodology

The objective of our paper is to predict the personality of a person based on their score of openness, extraversion agreeableness, neuroticism and conscientiousness. For achieving this, we needed a way to calculate the scores directly from every CV. Our approach as shown in *Fig.1* was to parse the entire resume and search for keywords relating to the ‘Big Five Test’.

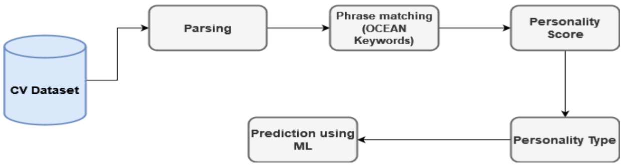

Fig-1: - Workflow of Proposed System

For parsing CVs, we have used pyresparser- a simple resume parser used for extracting important features such as name, email id, description, skills from CVs. Pyresparser supports PDF and DOCx files. The parsed data is then stored 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	Organised	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 given above houses various keywords of OCEAN. Each trait is associated with a set of 10 keywords that relate to it. There are many Natural Language Processing (NLP) libraries like Natural Language Toolkit (NLTK), TextBlob, SpaCY which could help us in parsing the resume data. We have used SpaCY- an open-source software library for advanced natural language processing and is helpful to handle large amounts of text data.

The PhraseMatcher class in spaCY is highly efficient in matching large sequences of tokens in documents [7]

The keywords in Table 1 will be matched by the mentioned class. Using PhraseMatcher class, our algorithm searches for the keywords and gives a score from a range of 0-10 according to the occurrence of OCEAN keywords in one’s CV. After assigning scores as shown in the table below, the algorithm labels each data point as dependable, extraverted, lively, responsible, or serious. Thus, we output a CSV file with degrees of the ‘Big five’ traits as all the columns. Each datapoint has been labelled as either dependable, extraverted, lively, responsible, or serious as given 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

3. MODEL TRAINING AND TESTING

Before training our model, we label encoded the Personality column of our dataset. Our final dataset had 708 rows and 6 rows. Using the sklearn library, we have used 70% of our data for training purposes and 30% for testing the results. For predicting the personality of a prospective candidate, we have used various machine learning algorithms like Logistic Regression, Naive Bayes, Random Forest, Support Vector Machine (SVM) and KNN.

- Logistic Regression

It is an algorithm analogous to Linear Regression, except it predicts whether something is True or False. It is a popular algorithm for solving classification problems like Binary Classification (Pass/Fail, Rain/No Rain).

- Naive Bayes

In probability, Baye’s theorem is used to compute the conditional probability. The theorem forms the basis of the Naive Bayes classifier, a classification algorithm that assumes strong independence assumptions between the features. According to the algorithm, each feature in the problem makes an equal and independent contribution to the outcome.

- kNN

kNN stands for k-nearest neighbours, a supervised machine learning algorithm capable of solving both regression and classification problems. Intuitively we can think of the proverb ‘Birds of the same feather flock together as similar to kNN. The algorithm assumes that similar data points usually occur in closeproximity.

- SVM

Support vector machine is a supervised machine learning algorithm used to deal with data for classification and regression analysis. The goal of SVM is to find a hyperplane in N-dimensional space (N- number of features) that can easily classify the data points.

- Random Forest

Random forest is another ensemble technique used for classification and regression tasks. It uses multiple decision trees to produce the output. Bagging or bootstrap aggregation are used to train the random forest algorithm's "forest.

After training our model on all of the algorithms, we realised that our predictions turned out to be rather poor. Even our best models could only find an accuracy of about 30 percent.

Another factor was that our training and testing datasets had very different distributions. While our training data was a little imbalanced, the testing data was even more imbalanced. But when we put ourselves in the shoes of an employer, we realise that he would want to hire someone who is ‘responsible’ and ‘lively’ more than anything else. Thus our problem now turns into a binary classification problem (1- responsible or lively 0-others) [8]

5. EXPERIMENTAL RESULTS

Now after feeding the data to the models, we managed to spike the accuracy to about 0.71. Random Forest algorithm provides the best accuracy followed by the likes of Bayes, kNN, SVM and Logistic Regression as seen in Table 3. As expected, Random Forest also has the least mean squared error, which measures the average of the square of the difference between actual and estimated values.

Table 3:- Accuracy And MSE Values

Model	Accuracy	MSE
Logistic Regression	0.62	0.37
Naive Bayes	0.65	0.37
kNN	0.64	0.35
SVM	0.63	0.36
Random Forest	0.71	0.29

6. CONCLUSION AND FUTURE SCOPE

In this paper, we have used various Machine Learning Algorithms such as Logistic Regression, Naive Bayes, Random Forest, SVM and KNN for Personality prediction using CV Analysis. Using pyresparser, spaCy and PhraseMatcher we were able to predict the personalities of various candidates. The results indicate Random Forest has the maximum accuracy of 0.71 however due to lack of available data the accuracy is much lesser than it was anticipated. The proposed system can be used by various companies in order to streamline the recruitment process by considering the personality of potential candidates. Future work can also be done to increase the efficiency and performance of the proposed system in order to predict personality using CV analysis more accurately.

7. REFERENCES

[1] A Demetriou, L. Kyriakides, and C. Avraamidou, “The missing link in the relations between intelligenceand personality” in Journal of Research in Personality,vol. 37 issue 6, December 2003, pp 547-581.

[2] M. Kalghatgi, M Ramannavar, and Dr. N. S. Sidnal, “Neural Network approach to personality prediction based on the Big-Five Model” in IJIRAE, vol2 issue 8, August 2015, pp 56-63.

[3] A..Robey, K. Shukla, K. Agarwal, K. Joshi, Professor S. Joshi “Personality prediction system through CV Analysis, in IRJET vol 6, issue 02, February2019.

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

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

[6] https://towardsdatascience.com/

[7] https://spacy.io/

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