2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. June 4-5, 2020

Personality Classification System using Data Mining

Sandhya Katiyar
Associate Professor
Galgotia's College of Engineering & Technology
Uttar Pradesh, India
drkatiyarsandhya@gmail.com

Himdweep Walia
Research Scholar
AIIT, Amity University
Uttar Pradesh, India
himdweep@yahoo.com

Sanjay Kumar
Professor
Galgotias University
Uttar Pradesh, India
katiyarsanjay@yahoo.com

Abstract – **Personality is one feature that determines how people interact with the outside world. Personality can be defined as a necessary element of a person's behavior. The way people interact with other people determines their personality. This paper covers the topic of Automated Personality Classification – a system that analyses the personality of a user based on certain features using Data Mining Algorithms. In this paper, a system is proposed which analyses the personality of an applicant. This system will be helpful for organizations as well as other agencies who would be recruiting applicants based on their personality rather than their technical knowledge. The personality prediction results are based on Big Five Personality traits and the classification is done using Naïve Bayes Algorithm and Support Vector Machine.**

Keywords – Naïve Bayes Algorithm, Support Vector Machine, Automated Personality Classification, Data Mining.

I. INTRODUCTION

Personality classification has been one of the most researched topics in the recent past. Personality is a combination of an individual's behavior and characteristics features that determines how he/she reacts under different circumstances. Individual Behavior can influence by all choices like a person observes regarding various things like books, clothes, music and films [1].

Personality can also affect his/her interaction with the outside world and his/her environment. Personality can also be used as an additional feature during recruitment process, career counselling, health counselling, etc. Predicting personality by analyzing the behavior of the person is an old technique. This manual method of personality prediction required a lot of time and resources. Analyzing personality based on one's nature was a tedious task and a lot of human effort would be required to do such analysis. This traditional method of predicting personality would require a lot of time and was very limited in scale. Also, this manual analysis did not give accurate results while analyzing the personality of a user from their nature and behavior. Since analysis was done manually, it affects the accuracy of the results as humans prone to be prejudice and generally see the things accordingly.

Data mining techniques are therefore used to study and analyses data and then identify any hidden patterns or information from a large data set. These techniques are used to mine user characteristics and then train the model accordingly to predict the personality of other users in the future. Using these techniques, the personality of an applicant is analyzed who is applying for a job in an organization which gives priority to one's behavior and personality rather than technical knowledge. Also, the applicant gets to know what all personality traits are in him/her and what all traits are missing. Thus, he/she can then be guided to develop those traits or to strengthen the other traits accordingly. The major purpose of this paper is to overview the data mining algorithms which are used to predict the personality of the user. In this paper, focus is on an online test which would be given by the applicant and then his/her personality would be predicted accordingly based on the Big Five Personality traits. In this way, candidates can be filtered out who are applying for a specific position in the organization. Thus, it would save the resources of the organization and they would then interview only those candidates which would be most suitable for the job.

II. LITERATURE SURVEY

Aleksandar Kartelj et. al. [2] said that reliable approaches can be used to classify the personality in various new researches by applying the concept of Automated Personality Classification. Firstly, we examined all the possible solutions and what all improvements can be made to the existing problems of Automated Personality Classification. Then we considered the extension of the Automated Personality Classification [APC] problem such as the Dynamic APC and how to remove inconsistency in textual data. This entire research was carried out in the context of social networks and related data mining mechanisms.

R. Wald et. al.[3] have used social media like twitter contents to identify human psychology. They said Twitter, a micro blogging site, is used by a number of users to share their experiences and thoughts about their day-to-day life. Although researchers have often discarded the method of predicting personality by analyzing the tweets because they are of the view that it contains very little content to predict significant information, but these tweets can be combined to make a larger picture of the user who is posting them. Select RUSBoost, a new form of ensemble learning has been used to predict psychopathy using Twitter, which uses four classification learners and four feature selection techniques.

Fazel Keshtkar et. al. [4] said that aims of developing methods for modelling student behavior based on data such as online conversations, discussions in class, etc. However methods like Intelligent Tutoring System (ITS) and

978-1-7281-7016-9/20/$31.00 ©2020 IEEE

Educational Data Mining (EDM) used individual's behavior and personality for analysis purpose. Thus, a system is developed which can be adjusted by the user and analyze student's behavior during their interaction as well.

Yago Saez et. al. [5] developed a system for analyzing the behavioral traits and cheerfulness of a user. Eysenck's theory defines too the human personality, the authors built a system wherein they would collect the text messages from different media resources such as twitter and face book and then classify them into various personality types. Although a clear link between behavioral traits and cheerfulness cannot be established, however some correlations do exist between them which could be found out in the coming future.

J. Golbeck et. al. [6] said that Social network is a platform where the users tend to reveal themselves to the outside world, sharing their behavioral information and giving imminent to other people into their lives. Personality plays an important part in many types of interactions among people; it can be used to predict the job satisfaction, professional as well as romantic relationship success. Until now, in order to accurately predict user's personality, they surveyed among different individuals with help of a survey test. However, this was highly unfeasible while collecting data from social media platforms and hence correct analysis of personality was a problem.

Nurbiha A Shukora et. al. [7] have given the concept of Online learning which became highly popular because of technological advancement that made it possible to have discussions even from a distance. Most studies that have been conducted report how effective online learning has helped students to improve their learning power while assessing the learning process simultaneously. This kind of discussion can be possible only by applying data mining technique wherein we can assess the different experiences of students which they filled online on the basis of their log files. However, it is suggested by the results that students should put more hard work to became an excellent online learner

III. PROPOSED METHODOLOGY

To counter the problems of the existing system, an Automatic Personality Classification system is proposed which uses some techniques of Data Mining to classify the human behavior of the individuals. The system uses three different algorithms to predict user behaviors and form a base from received pattern. These three algorithms are the first is Big Five Model along with advanced data mining, second is Support Vector Machine and third is Naïve Bayes theorem. By giving proposed model user personality can now be identified based on past user history and obtained traits pattern.

This system analyses rarely used user attributes and qualities and observes users past behavioral patterns and then forms its own patterns which help to predict the users personality and classify in different category among all users[10]. The proposed work is beneficial for predicting personality of applicants applying for various roles in an organization.

An Automatic Personality Classification system is designed in which every applicant is given a separate user name and password. Each applicant logins the test using his/her user name and password and then takes the survey. The survey consists of 30 questions, where each question determines one of the Big Five Personality traits.

Once the applicant takes the survey of 30 questions, he/she can see the result of his/her personality. The survey analyses the personality of the applicant based on the Big Five Personality traits.

A graph has also been displayed based on various qualifications specified by the applicants in the registration page. The graph then predicts which all candidates are suitable for a particular job opening in an organization.

The algorithms like Naïve Bayes and Support Vector Machine groups the qualities of the applicants of different streams and a graph is shown which depicts which set of candidates would be most suitable for a particular post in the organization.

This type of graph will be helpful in screening the candidates and would save the resources of the organization. The organization, on seeing the graph, can call only those applicants belonging to a particular stream whose scores are greater than the score of applicants of other streams. Thus, this type of system would be cost effective for an organization as well as provide accurate results for choosing applicants for a particular job.

A. Big Five Personality Trait

The Big Five Personality traits are the five dimensions or the domains of personality that can be used to analyse or predict the personality of a user [9]. The Big Five Factors are:

1. Openness to Experience or Imagination Capability
2. Agreeableness
3. Extraversion
4. Neuroticism or Emotional Stability
5. Conscientiousness

The Big Five Personality Model is the most widely accepted and researched model for predicting the personality of a user. The Big Five Personality traits are found in a variety of people of different ages, locations and cultures. The Big Five Personality results are very accurate and predict the true personality of a user to a large extent.

B. Support Vector Machine

This is a type of machine which is basically used for analysis the data which receive from supervised learning and identify the patterns for classification [8]. Training data set is taken and checked that whether the test data belongs to existing class or not for personality classification and classification. Data is represented by Support Vector Machine model in the form of a point commonly in space which further classified in a line or in a hyper plane. The main idea behind the support vector machine algorithm is that if a classifier performs well at the most challenging comparisons, then it

will definitely perform even better at the most easy comparisons. Support Vector Machine which is a nonlinear classifier often produces superior classification results than other classifier methods. Support Vector Machine is based on non-linearly mapping the input data to some high dimensional space where the data is separated linearly, thus giving accurate classification results.

The steps involved in Support Vector Machine are:

1. Create vectors for given question answers.
2. Then calculate the weights of the vectors.
3. Get the vectors with highest value and find value of personality
4. Finally predict personality type.

C. Naïve Bayes Algorithm:

Naïve Bayes Algorithm, which is a type of an inductive learning algorithm, is considered to be one of the most efficient and effective algorithms that is widely used in data mining. The performance of Naïve Bayes Algorithm in classifying data is quite accurate because the conditional independence assumption on which the entire algorithm is set up is rarely true for the real world applications. The application of Bayes theorem forms the basis of Naïve Bayes Algorithm [8]. A variation of Naïve Bayes Algorithm is Multinomial Naïve Bayes Algorithm which is also designed for classification purposes. The Multinomial Naïve Bayes Algorithm uses multinomial distribution in which it considers either how many times a particular word occurs or calculated weight of that particular word for classification. The Naïve Bayes algorithm combines both efficiency i.e. optimal time performances with reasonable accuracy.

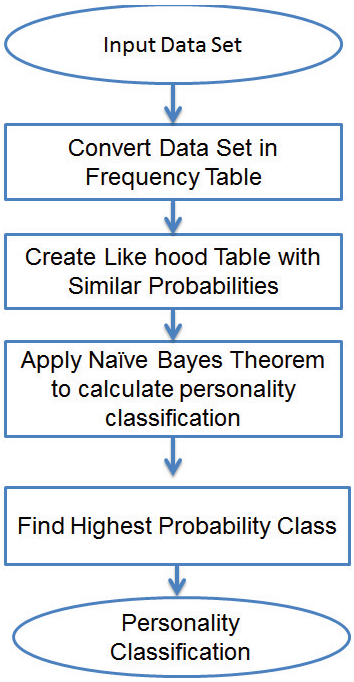

Fig. 1. Flowchart describing the proposed system

A limitation of Naïve Bayes method is it supposes that all the linguistic features are conditionally independent.

However, even if the Naïve Bayes Algorithm produces an oversimplified model, the classification results which it produces are surprisingly accurate.

In our system datasets, the results are either stored in the form of accurate or inaccurate. Hence the entire working of the system depends on the corresponding probabilities of either accurate or inaccurate. The steps involved in the Naïve Bayes Algorithm are:

Step 1: Convert dataset into a frequency table.

Step 2: Create likelihood table based on probabilities of similar types of personality traits.

Step 3: Use Naïve Bayesian equations to calculate the posterior probability of each class.

Step 4: Highest probability class is the outcome of the prediction.

IV. RESULT ANALYSIS

Based on the survey taken by the applicants, a graph is plotted for the applicants having different qualifications. The graph helps the organization to filter out applicants for a particular job opening in the organization. This helps the organization to screen the applicants based on their personality and a lot of resources are saved because the applicants which are not suitable for the job are filtered out in the first phase only. This helps the organization to conduct interviews for only those applicants which have passed the personality test. The organization need not interview the rest of the applicants which will save their time and money.

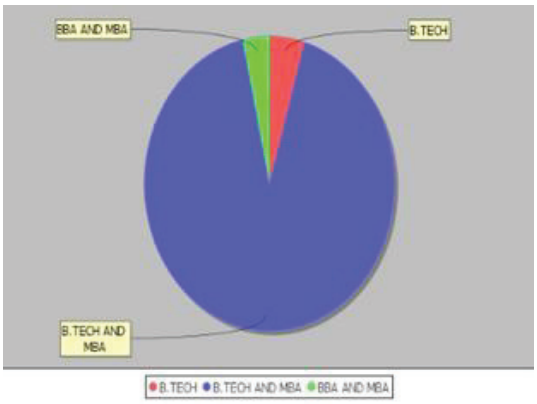

Fig. 2. Graph - 1

The graph gives us an insight as to what all applicants will be suitable for a particular job opening. Each applicant will give the personality test and based on their score applicants having similar qualifications will be grouped in one class and the graph will be plotted separately for each and every class of qualifications of the applicants. This graph will then be helpful for the organization to segregate class of applicants having highest scores and will then be interviewed for the job.

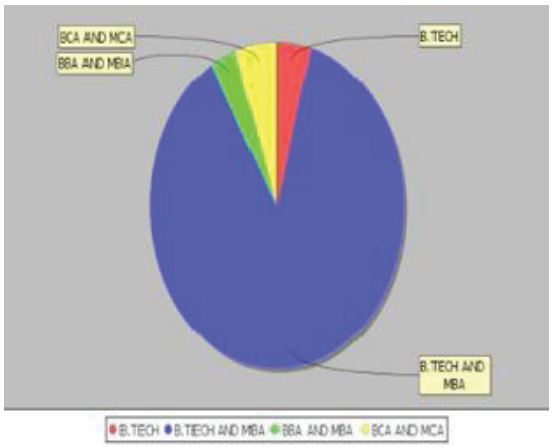

Fig. 3. Graph - 2

This Personality Prediction System will evaluate the overall personality of the applicants and will be useful for organizations which are recruiting applicants based on their personality rather than their technical knowledge. The system will help all the organizations recruiting for management and sales post and will give a boost in their revenue system because the resources for recruitment process will be saved to a large extent.

Also, the applicants would themselves get to know their own personality in a separate section provided in the system. They will get to know where they lack and what all personality traits they have to develop. Also, it will give them an insight into where they are strong and where they have to perform more.

Thus, the Personality Prediction System will help the organization by saving their resources and also the applicants by telling them their strong and weak areas.

V. CONCLUSION & FUTURE WORK

In this section, a discussion is done on how the classification algorithms performed for predicting the personality of the user. Naïve Bayes Algorithm has the best accuracy in two methods tested with an average accuracy of around 60%. Support Vector Machine method performance was a little worse than Naïve Bayes due to the difficulties of separating a class of a word as dataset was not quite accurate.

Personality analysis and prediction has increased very much in the recent times. Extracting the personality of the user using the current system is very much helpful in various fields, for instance, recruitment process, medical counseling, and likewise. Personality detection from survey means to extract the behavior characteristics of the users taking the survey. This paper focuses on providing a state-of-art review of an emerging filed i.e. personality detection from survey. This paper also discusses the state-of-art methods for personality detection and prediction.

Apart from the work done towards this system, future work mainly comprises of the following objectives:

For future work, more personality traits should be included so that a more detailed personality analysis can be shown to the user as well as to predict personality using textual data and sentiment analysis.

There can be module where user will be provided with career guidance and counseling sessions which match his personality.

REFERENCES

[1] I. Cantandir, I. Fernandez-Tobiaz, A. Belllogin, "Relating personality types with user preferences in multiple entertainment domains," EMPIRE 1st Workshop on Emotions and Personality in Personalized Services, 2013.

[2] Aleksandar Kartelj, Vladimir Filipović, Veljko Milutinović, Novel approaches to automated personality classification: Ideas and their potentials.

[3] R. Wald, T. M. Khoshgoftaar,A. Napolitano Using Twitter Content to Predict Psychopathy.

[4] Fazel Keshtkar, Candice Burkett, Haiying Li and Arthur C. Graesser,Using Data Mining Techniques to Detect the Personality of Players in an Educational Game.

[5] Yago Saez , Carlos Navarro , Asuncion Mochon and Pedro Isasi, A system for personality and happiness detection.

[6] J. Golbeck, C. Robles, K.Turner, "Predicting personality with social media," In CHI'11 Extended Abstracts on Human Factors in Computing Systems, pp. 253-262, 2011

[7] Nurbiha A. Shukora, Zaidatun Tasira, Henny Vander Meijden, "An Examination of Online Learning Effectiveness using data Mining", Science Direct – Procedia – Social and Behavioural Sciences 172 (2015) 555 – 562.

[8] C.D. Manning, P. Raghavan, H. Schutze. Introduction to Information Retrieval. Cambridge UP, 2008.

[9] P.T. Costa, R.R. McCrae, "Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI)," Psychological Assessment Resources, 1992.

[10] Crist´obal Romero, Member, IEEE, and Sebasti´an Ventura, Senior Member, IEEE, "Educational Data Mining: A Review of the State of the Art" VOL. 40, NO. 6, NOVEMBER 2010.

[11] D. Jurafsky, J. H. Martin, "Naive Bayes Classifier Approach to Word Sense Disambiguation", Chapter 20 Computational Lexical Semantics, available at (http://www.let.rug.nl/nerbonne/teach/rema-stats-meth-seminar/presentations/Olango-Naive-Bayes-2009.pdf), last seen 2015.

[12] H. Walia, A. Rana, and V. Kansal, "A Naïve Bayes Approach for working on Gurmukhi Word Sense Disambiguation", 6th international Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2017, IEEE Explorer.