International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

Personality Prediction from Stack Overflow by using Naïve Bayes Theorem in Data Mining

Sandhya Katiyar, Sanjay Kumar, Himdweep Walia

Abstract: ***This paper presents a system that provides an automatic process which defines how to evaluate designation and predict personality of individuals as wel for any jobl. This system is being set up via a web application, where the system manipulates data from social media sites like Twitter and Blog-Forums i.e. "Stack Overflow" of individuals. Myers-Briggs personality type is used for predicting the personality behavior of individuals. The technical skills in the form of raw data have been classified by using Gaussian Naive Bayes classification and for personal behavior; data mining & machine learning are used. This system is used to analyze the personality of an individual and hence using this information to build a team for a job.***

Keywords: Data Mining, Gaussian Naive Bayes Classifier, MBTI, Myers-Briggs Personality Type. , N-Gram Algorithm, Stack overflow, Twitter, TF-IDF Algorithm, Word Vector Algorithm

I. INTRODUCTION

Human behaviors or Personality can be measured by individual characteristic. Personality is a psychological construct and can be identified by common pattern of behavior [1]. Common patterns of behaviors like attitude, continuous happiness, sadness, work priorities, strength, weaknesses, encouragements & motivations and health issues also affect the personality analysis. Based on analysis of individual's characteristics, various applications have been developed such as behavior analysis system[2], customer satisfaction analysis, opinion mining, predictive analysis[3], customer interaction analysis[4], predictive system [5]and so on [6].

It is observed that personality analysis by traditional method like written test, interviews and survey are time consuming, costly and lead to inefficient results. As people use social media very frequently and share their personal and behavioral nature and The large amount of personal information include thoughts, feelings, aspirations, goals, failures, successes, fears, and dreams, as well as one's likes, dislikes, and favorites on social networking platforms make it more easy to analyze the individual's behaviors and conclude their personality character through the social sites[7], which have focused much awareness from different analysis[8], [9], [10].

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Sandhya Katiyar*, Department of Information Technology, Galgotias College of Engineering &Technology, Greater Noida, India, E-mail:drkatiyarsandhya@gmail.com
Sanjay Kumar, School of Computing Science & Engineering, Galgotias University, Greater Noida, India, E-mail:drkatiyarsanjay@gmail.com
Himdweep Walia, Amity Institute of Information Technology, Amity University, Noida, India, E-mail: himdweep@gmail.com

Thus it leads to the development of automated system of personality classification that Analysis of a user's personality attributes from widely available information on social media.

The personality prediction concept has always been a "red hot" topic since the advent of the intellectual class which has been thinking on how to save "man's hours" as much as possible. One of the most important contributors to this concept is Brig Myers. She was very impressed with the concepts of Carl Jung and just referenced the ideas of Carl Jung and thus implemented Carl's ideas along with her own insights. She created a test which came into existence after 20 years of rigorous hard work and research with and mother and many other companions. In an era this topic (personality prediction) did not saw a considerable acceleration that was required but it kept on having a "tortoise walk". The research is still being in process in the 21st century and thousands of articles have been written on it. The personality type of an individual is analyzed based on a questionnaire that is framed considering the various personality traits exhibited by an individual. With the help of this questionnaire the people can find their "best match" and thus can realize the best personality type that will help her/him get success with at-most precision. As said, Briggs Myers was inspired by the ideas of Carl; the three original pairs of references in his topology are Sensing and Intuition, Introversion and Extroversion, Thinking and Feeling. After studying all the three pairs Briggs Myers added a fourth pair. Extraversion or Introversion is defined how an individual apply his energy on people or things and where these directions can be further used in the outer world or alone in the inner world [11]. Sensing or Intuition implies sensing learn to the individual that how to deal with information in different ways like user can use that information directly or by narrating in some other way or adding any new meaning to make it more interactive.[11]. Thinking or Feeling is the way by which one can take decision refers to decision making. It may be any logic, any particular approach or supervised or unsupervised or considering other people and special circumstances [11]. Judging or Perceiving defines the quality that how an individual see the things from his own point of view and how he can relate that thing from outer world- directly or indirectly and basis of his perception how he judges the things[11]. The information provided by individual is suitable for any profession, is verified based on the entries made by the users on "Stack Overflow". The goal of this system to reduce the time spent in personality classification of Individual traits. The system is not to guide the recruiters for making decisions on their capabilities but only aid them in tedious process of recruitment. The remainder of this paper will be organized as follows: Section II will present literature survey.

Section III will discuss the architecture of system employed in this paper. Section IV presents the details of the Myers-Briggs personality type. Section V presents our methodologies and algorithms to implement the system. Finally, section VI provide conclusion.

Web application is used for design and prototype implementation of the proposed system.

II. LITERATURE REVIEW

2.1 Using social media, evaluating emotional intelligence

The ranking of candidates can be done easily by analyzing their tweets for emotional aptitude and for professional eligibility; analysis can be done on resumes provided. Creating a set of tweets addressed as cluster, on these cluster meta-attributes extraction and multi label classification is done. The features extracted by the meta-attribute extraction are classified into linguistic and profile-centric. Then these Meta attributes are mapped to a subset of the big five personality traits. Some learning to rank algorithms like Support vector and linear regression are employed to work on the training set on the previous selection of candidates [12].

2.2 Data mining techniques to predict personality in an educational system

Personality behavior of student is considered in computer based learning environment. The personality can be predicted based on the chat interaction log in an educational game. A supervised system using data mining, machine learning algorithms and natural language processing [18] can be used to predict the personality of students. Using this system prediction can be made on accuracy, precession, recall and F-measure, where machine learning algorithms like Naive Bayes and SVM are used for classification [13].

2.3 Automated personality prediction

An automated personality prediction can be set up via textual content a person wrote and meta-information about the person through social means. LIWC and MRC are needed to be found out which correlates with personality characteristics. LIWC (Linguistic Inquiry and Word Count) holds a database of many words and these words relate to the personality characteristics. MRC is psycho-linguistic database which include of words classified by various measures as, imagery, concreteness, frequency of usage, etc.

Comments from the Internet site youtube.com could be used in a predictive way. The analysis of public data flows from social networks is yet another resource that is currently not used for solving the APC problem. To analyze behavioral impact like, shyness, judgmental, and frankness to experience, comments and news portal can be used. Similarly youtube.com comments can also be used, use of the analysis of public data flows from social network [14].

1. MYERS - BRIGGS TYPE INDICATOR (MBTI)

MBTI (The Myers–Briggs Type Indicator) is a survey form result of which gives the introspection of an individual and indicates different psychological preferences of people observe the things as per their perception around them and analyze to judge accordingly.

It was observed by taking all statistics that it was made for normal people which tells the difference of natural occurrence [11]. "The suppositionon which MBTI based is that we all have some common categories in our behavior which make our personality and some mechanism to predict our traits too, and these common categories may be our way of giving personal information, strengths, weakness, happiness, and dedication.

For any particular instance commonly 4 types of behaviors are described by four letters: These are

ESTJ: Extraversion (E), Sensing (S), Thinking (T), Judgment (J), INFP: Introversion (I), Intuition (N), Feeling (F), Perception (P)

Figure 1: 16 Personality types [15]

Different behavioral parameters used earlier by MBTI have few correlation matrixes with above different personality behaviors, these personality behaviors and characteristics are generally uploaded by the people in the social

media. McCrae ve his personality prediction theory based on Costa five theory models [11]. McCrae defines correlation matrix and scales these personality traits as well to find out the suitability of an individual for particular designation. With help of correlation matrix one can predict the ability and experience of an individual l[16]. In more recent years, it is observed that there is extremely use of mobile and people share their information in, social media therefore research on Personality Prediction has been done in two directions: Personality Recognition via smart devices and Personality Recognition on web [17].

Table1. Correlation between MBTI and Big-Five [11]

	Extraversion	Openness	Agreeableness	Conscientiousness	Neuroticism
E – I	**-0.89**	0.02	-0.02	0.05	0.10
S – N	0.05	**0.85**	0.02	-0.05	-0.03
T – F	0.13	0.01	**0.55**	-0.12	0.04
J – P	0.12	0.20	-0.05	**-0.69**	0.10
If behavior are closer to1.0 or -1.0, the higher the prediction probability.					

III. GENERAL ARCHITECTURE

The most recognized approach for solving the Automatic Personality Classification problem is described in the following steps [12]:
A. Gathering the corpus data,
B. Determination of the personality characteristics of the participants, and
C. Building the model.

We are working on a system which has its principle laid on collecting data from social media platform like twitter and a question and answer site i.e. Stack Overflow, here we can find huge amount of behavioral data exhibited by a person which can actually predict about the traits of a person's behavior. By using the following data we classify individuals using personality classification. Our project focuses on using learning algorithms and data mining to extract the characteristics exhibited by the user and learning from the various patterns occurred. Now as a result of the learning, the system can predict the user personality based on the past experiences.

Basically, the system analyses the huge amount of behavioral characteristics and from the patterns observed, it stores its very own characteristic patterns that has already observed in a database. The above proposed system predicts the personality of any new user based on the behavioral data (personality) stored by the classification of previous user data. The above system proposed is very useful and important for various social networks and various online ad selling companies who uses the user's personality and proposes ads according to the findings. This system is also very useful for various government investigating agencies who checks and predicts the personality of a particular suspected person according to the results shown.

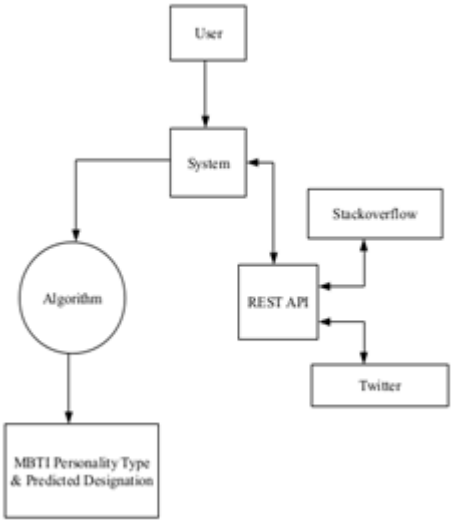


Figure 2: System Architecture

IV. SYSTEM WORKING

General working of the system is as followed:-

a. First of all the user will input their basic information along with Twitter User-Handle and User-ID (Numeric) of Stackoverflow.com to the system.
b. Now, The system will fetch their data from Twitter Database and Stack overflow Database using REST API.
c. At last, in the system the data (fetched from databases) is transferred to algorithm to calculate the MBTI Personality type and Predicted Designation

V. METHODOLOGY & ALGORITHM

This paper describes the process of identification the personality of the people based on common attributes uploaded by them in twitter. We will be using only information contained within the tweets for our objective so user will not feel that his privacy is being compromised. Personality prediction can be found by multi label classification if we consider different available combinations of personality attributes among the many described in the MBTI. We have used Naive Bayes Classifier in our system which will emphasize our solution approach.

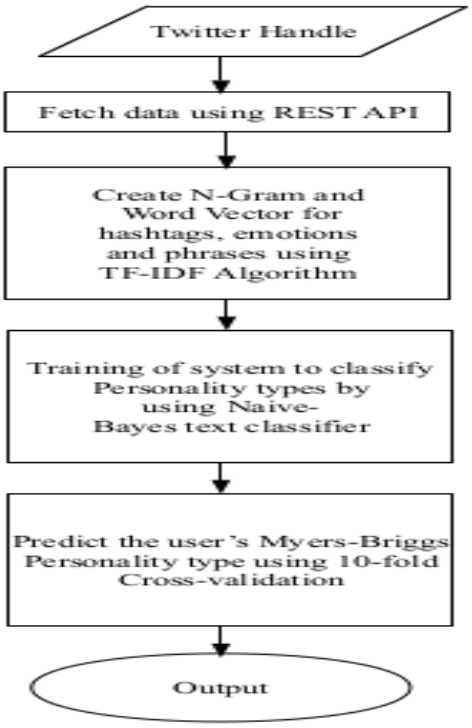


Figure 3: Flow Chart of Induced Steps

6.1 Fetch data using REST API

General data from the user will be collected including their Stack overflow User-ID and User-Handle of Twitter. Now data is extracted from Twitter using its REST API call providing the User-Handle of Twitter. Similarly, it is done with Stack overflow. The data returned is in the raw form is collected.

6.2 Creation of n-gram and word vectors using TF-IDF algorithm

Now, the data fetched from the user account is saved into "user.sav" file on which we apply tf-idf algorithm to determine the frequency of the words that help us to generate the data, appropriate for Naive Bayes Classification where, TF-IDF stands for term frequency - inverse document frequency.

The tf - idf weight is a weight often used in information retrieval and data mining.

It is an analytical calculation that is proposed to reflect how important a word is to a document in a collection or corpus.

6.3 Classification of data using Gaussian Naive Bayes classifier

Since, the input variables are having real-time values; therefore we choose to apply the Gaussian Naive Bayes to train the algorithm to fetch output on its own when a certain sample input variables are provided.

Where, Gaussian Naive Bayes for a input variable x and data is segmented into different classes, now the probability distribution for any observed value v (to be predicted) will belong to a class C_K is :-

$$p(x = v/C_K) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

1.4 PERSONALITY PREDICTION AND CROSS VALIDATION

Previously, the personality types being used are expected to be 5 for Big5 or MARS model. But in our research, we've re-structured the personality types to 16 in total by implementing the basic four personality traits to each one of Myers Briggs personality types. As of choosing between introvert and extrovert; sensitive and intuition; thinking and feeling, judging and perceiving. The output of the classifier is then been mapped to one of these 16 personality types to seek the personal characteristics of an individual, which is already been stated by Myers-Briggs and we're also implementing the mapping of David Keirsley for the designation for each one of the 16 personality types.

Where, 10-fold cross-validation is a prediction methodology toconstruct the predictive models by categorizing the original dataset into a sampling set to preparethe model, and a sample set to evaluate it.

VI. RESULT ANALYSIS

The output vector,

{[(0.0, 409), [(0.0, 441)], [(1.0, 449)], [(1.0, 466)]}

is generated by Gaussian Naive Bayes Classifier and also mapped in one of the Myers-Briggs 16 Personality types i.e., ENTP which defines the characteristics of ENTP like "Frank, decisive, assumes leadership quality. This paper reduces illogical, inefficient procedures and policies of personality prediction by developing and implementing comprehensive systems to solve organizational problems for recruiting their staff and providing well informed, well explained knowledge and passing it on to others.

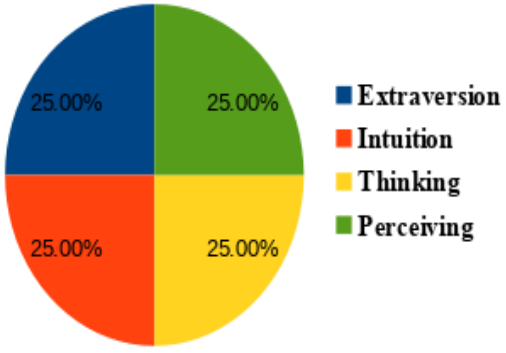
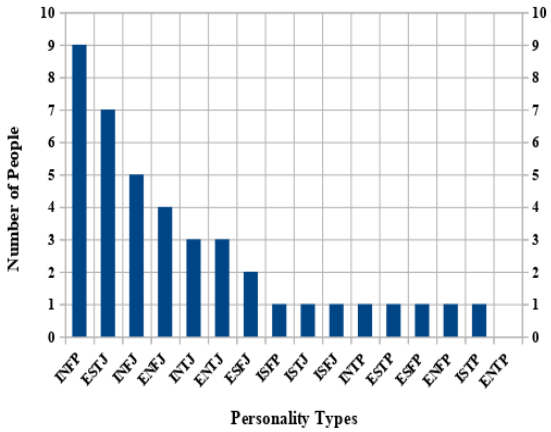


Figure 5: Pie Chart for Personality of individuals

Here in this graph 1, we examined personality traits on 40 Individuals. The output we got was very promising, consistent and true to great measure. The output is been plotted in the graph followed:

Personality Type→	Extraversion	Intuition	Thinking	Perceiving
weight age %	25	25	25	25



Graph 1: Personality Prediction as an end result

This result shows personality prediction of 40 individuals; earlier published research has done the same for lesser personality types. So, this research provides a better overview for personality prediction with maximum personal trait combinations. And this research is also capable to provide a technical trait which was not given in earlier researches. A prototype is developed for this research and the developed code is consistent, robust, reliable, usable, maintainable, readily portable to different systems, scalable and not vulnerable for individual's data (Twitter/Stack overflow).

VII. CONCLUSION & FUTURE SCOPE

The purpose of this paper is to serve to provide a personality prediction and an autonomous process of designation evaluation. The research we undergone for this paper, add on for data mining and machine learning as a result in the form of raw data classified using Gaussian Naive Bayes classification. The researched prototype we implemented enables us to predict an individual's personality from Myers-Briggs 16 Personality types, by analyzing his/her twitter handles and same for stack overflow for technical traits.

REFERENCES

1. Vinciarelli A and Mohammadi G2014 A survey of personality computing *IEEE Transactions on Affective Computing***5**273-291
2. Roshchina A,Cardiff J and Rosso P 2011 A comparative evaluation of personality estimation algorithms for the twin recommender system *3rd International Workshop on Search and Mining User-Generated Contents* Glasgow ISBN: 978-1-4503-0949-3
3. Enos F, Benus S, Cautin R L, Graciarena M, Hirschberg J, and Shriberg E 2006 Personality factors in human deception detection: comparing human to machine performance *INTERSPEECH 2006 – ICSLP* 813-816.
4. Luyckx K and Daelemans W Personae: 2008 A corpus for author and personality prediction from text*6th International Conference on Language Resources and Evaluation* 2981-2987.

5. Golbeck J and Hansen D2011Computing political preference among twitter followers *SIGCHI Conference on Human factors in Computing Systems***36**1105-1108
6. Celli F, Pianesi F, Stillwell D, and Kosinski M 2013 Workshop on computational personality recognition: shared task*7th International AAAI Conference on Weblogs and Social Media*
7. RainieLand Wellman B 2012 Networked: the new social operating system*MIT Press Cambridge S* **JSTOR**
8. Gosling S D, AugustineAA, Vazire S, Holtzman N, and GaddisS, 2011 Manifestations of personality in online social networks: self-reported facebook related behaviors and observable profile information*Cyber psychology Behavior and Social Networking***14**483-488.
9. Buchanan T and. Smith J L 1999Using the internet for psychological research: personality testing on the world wide web*British Journal of Psychology***90**125144.
10. Sumner C, Byers Aand Shearing M2011Determining personality traits & privacy concerns from Facebook activity*Black Hat Briefings***11**197-221.
11. Varvel T, Adams S G and Pridie S J 2003A study of the effect of the myers-briggstype indicator on team effectiveness*American Society for Engineering Education Annual Conference & Exposition* .124.1-8.124.9.
12. Kartelj A, Filipović V and Milutinović V2012 Novel approaches to automated personality classification: ideas and their potentials*IEEE 35th International ConventionMIPRO 2012*ISBN 978-1-4673-2577-6
13. Keshtkar F, Burkett C, Li H and Graesser A C2014 Usingdata mining techniques to detect the personality of players in an educational game*Studies in Computational Intelligence*125-150.
14. Ombhase M, Gogate P, Patil T, Nair K and Hegde G 2017Automated personality classification using data mining techniques*Researchgate*10.13140/RG.2.2.35949.59363.
15. https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/
16. Menon V M and Rahulnath H A2016A novel approach to evaluate and rank candidates in a recruitment process by estimating emotional intelligence through social media data*International Conference on Next Generation Intelligent Systems*ISBN: 978-1-5090-0870-4
17. Xue D, Hong Z, Guo S, Gao L, Wu L, Zheng J and Zhao N2017 Personality recognition on social media with label distribution learning*IEEE ACCESS*(99):1-1.
18. Pratama B Y and Sarno R2015Personality classification based on twitter text using naive bayes, knn and svm*ICoDSE 2015*ISBN: 978-1-4673-8430-8

AUTHORS PROFILE

Dr. Sandhya Katiyar is working under the department of Information Technology in Galgotias College of Engineering and Technology Greater Noida (U.P.) since July 2007. She obtained her Ph.D (Computer Science and Engineering) in 2018 under faculty of engineering, M.M University, Mullana, Ambala, M.Tech. (Computer Science and Engineering) From Kurukshestra University Kurukshetra (Haryana) in 2007. She supervised 4 M. Tech students and many BTech students. Her research area includes Wireless Sensor Networks, Reliability Theory, Artificial Intelligent and Cryptography.

Dr Sanjay Kumar is working as professor in school of Computing Science and Engineering Galgotias University Greater Noida India. He obtained his Ph.D. (Computer Science and Engineering) in 2015 from M.M University, Mullana, Ambala, MTECH (CSE) in 2005 from University School of Information Technology, GGSIP University Delhi and MIS (Computer Science and Engineering) in 2002 from Dr.B.R.Ambedkar University Agra.

His research area includes Biometric Security, Big Data, Data Analytics, Software Engineering, Wireless Communications, Mobile Ad hoc & Sensor based Networks and Network Security.