A Study on Personality Prediction

Priya Yadav
AIT-CSE,Apex,
Chandigarh University
Gharuan,Punjab,India
20BCS6126@cuchd.in

Hitesh Kumar
AIT-CSE,Apex
Chandigarh University
Gharuan,Punjab,India
20BCS6157@cuchd.in

***Abstract*— Personality is an important aspect of one's perspective towards their life. It majorly impacts the decision-making and approach to solving the problem. Information about users and what they expressed through status updates are such important assets for research in the field of behavioral learning and human personality. Personality prediction to better accuracy could be very useful for society. There are many papers and research conducted on the usefulness of the data for various purposes like marketing, dating suggestions, organization development, personalized recommendations, and health care to name a few. Similar researches have been conducted in this field and it grows continually till now. Particular approaches differ concerning different machine learning algorithms, data sources, and feature sets. The goal of this project is to investigate the predictability of the personality traits of users based on different features and measures of the Big 5 model. This study attempts to build a system that can predict a person's personality based on the dataset collected through internet. The dataset contains user's information that helps model to learn and give accurate personality based information.**

Keywords—personality prediction, mining, Big 5model, behavioural learning

I. INTRODUCTION

Personality is a way a person responds to a particular situation. It is a combination of characteristics that make an individual unique. Assessment of personality over the past two decades in various researches has revealed that personality can be defined by five dimensions known as Big Five personality traits. In general, the study of personality is considered as a psychology research based on the survey or questionnaire. But this limits the research data to a smaller number of persons. Hence there is a need for something through which we can increase the number of people involved in surveys and to make the process automated.

Personality identification of a human being by their nature is an old technique. Earlier these were done manually by spending a lot of time to predict the nature of the person. Data mining is primarily used today by companies with a strong consumer focus- retail, financial, communication, and marketing organizations. A dataset is scrapped from the internet and fed to the model, using machine learning techniques to understand the personality of the users to give accurate predictions.

But these traditional methods are time consuming and very limited in scale. Our Proposed system will provide information about the personality of the user. Based on the personality traits provided by the user, System will match the personality traits with the data stored in the database. System will automatically classify the user's personality and will match the pattern with the stored data. System will examine the data stored in the database and will match the personality traits of the user with the data in the database. Then the system will detect the personality of the user. Based on the personality traits of the user, the system will provide other features that are relevant to the user's personality. 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 the recruitment process, career counseling, health counseling, 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. 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

The following paper is categorized in sections; Section II illustrates a brief of a few significant study already published study concerning this analysis. Section III gives an overview of the methodology, feature extraction, and specification of machine learning algorithms used in the study.

Section IV lays out the result of the analysis using a machine learning model. The conclusion is derived in Section V. In the end, Future work is described in Section VI followed by acknowledgment and references.

II. LITERATURE SURVEY

This section provides an outline of many researchers' past work on the sentimental analysis of social media users. Sentiment analysis deals with distinctive and categorizing views or sentiments that are present in the extracted text. Social media is catalyzing an enormous quantity of sentiment-rich information such as tweets, reviews, weblog posts, etc. Extracting emotions became meaningful using the user generated information for the analysis.

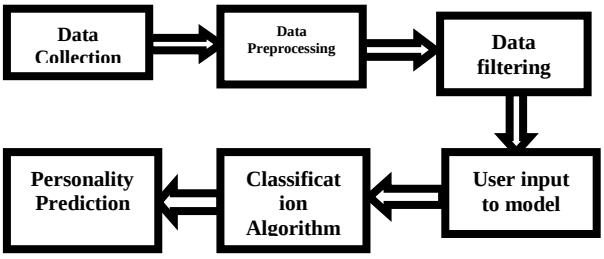
Arsa and Shubhangi, (2015) [1], aimed for developing handwritten- based personality and behavior identification systems. Supervised techniques Support vector machine and Artificial Neural network were employed into designing this model. They achieved their aim of securing the accuracy of 98.5%.In future; analysis will be performed for multiple lines.

Kedar et al.(2015) proposed [2] ,By using the supervised techniques ANN and Zernlike and Pseudo-Zernlike, Author builds a personality identification through handwriting analysis and graphology study, which is capable of determining personality of an individual by achieving the

accuracy 90%. This system has its potential in personal recruitment in marketing, medicine, and counseling etc.
Ilmini and Fernado (2016)[3], build models to identify personality traits from face image, identification of criminal behavior in criminology etc. Implementation of Supervised technique ANN and SVM into the model achieved the accuracy of 98.5%.Large dataset can improve the accuracy of the classification. More study on psychology and improved feature extraction phase may improve final results.
Sagadevan et al. (2015) [5], recognized the personality of facebook users from messages based on three Factor Personality (PEN) models. Techniques like supervised, stemming and part of speech tagging (POST) were employed into this system that managed to achieve the accuracy of 95%.In future, use of the higher negative words as cues to detect the psychoticism trait among facebook users will be implemented.

III. METHODOLOGY

Personality analysis of human behavior based on data reviews is a new area that needs more attention. The collected user's information has been developed from scratch to organize for cleanliness. Model is trained on classification algorithm mentioned below under classification and techniques. Preparation sets marked with individual highlights are provided as an aid to the classification to add additional information such as test sets. Finally, prototype is build to give user an interface to predict their personality.



FIGI Methodology Flowchart

A. Data Source

Determining the supply of data could be a vital task to advance the ultimate investigation. Online media levels are broadly divided into three general classifications as sources of information; User's personal information being vital is challenging to collect. Data collection is a not only a challenging phase but a important step too for prediction.

1. Big 5 models

The message posted on Twitter is called a tweet, limited to 140 characters. Tweets typically include the following: text, contacts, emoticons, and images. The six-second video was appended in 2012 as a tweet section. By closely logging these factors, extraction is applied to text, pictures, emoticons or emojis, and recordings. The tweets contain three records, including the hashtag (#), retweet (RT), and record ID (@).

B. Data Pre-Processing

Data preprocessing is a decisive tool for data mining algorithms. Twitter consists unstructured data set. it is a platform consisting of a pile of feelings, opinions, attitudes, emotions, etc expressed by the users over time. The projected work is to research the sentiments expressed on Twitter with the help of collected information from the users. Good quality data will lead to boot and improve the higher performance of the information preprocessing tool.
Extracting Information from Twitter is challenging. To build up approximate information, you must prepare or swap up raw data to execute a classification. Completed works include uniform packaging, hashtags, and other Twitter documents (@, RT), emoticons, URLs, stopwords, slag word decompression, and extended word stress.

C. Feature extraction

Feature Extraction is a technique to originate new features from prevailing ones. The new extracted features preserve the original information of existing features. . This approach encapsulates the original features by generating new reduced features containing the same information.
Prefabricated datasets have different individual properties. In the component extraction technique, we separate the different angles, descriptive words, action words, and objects and then identify these contexts as default or negative to indicate the intensity of the whole sentence.

A feature techniques applied to train the machine learning model is bag of words.. 2500 features were procured by applying this technique to the models' training.

1. Bag of words:

The often used feature extraction technique in NLP is Bag of words. For the training of the set recurrence of the words is given as a function. The BoW approach is implemented in this study by utilizing the Count Vectorizer from the Scikit-learn library of Python [10]. Transforming textual data to acquire numerical vectors is coined vectorization. The recurrence of words is enumerated signifying that tokens have been added to the making of token vectors. A value is assigned by the Bag of feature technique to every attribute based on the recurrence of those features.

D. CLASSIFIER AND TECHNIQUES

This section provides a brief description of machine learning models for the sentimental analysis of tweets. These machine learning models have been implemented because of their exceptional performance over traditional models for analysis. Some of the famous models like Naïve Bayes, Random Forest, Logistic Regression, Decision Tree, Support Vector Machine, and XGBoost are mentioned in this paper for accomplishment.

1. Naïve Bayes

The rule is known after the renowned statistician Thomas Bayes, who proposed the Bayesian theorem. This theorem assumes that each one of the attributes square measures does not freelance to every alternative. During this rule, the chance for every attribute concerning, bound category level

is calculated. The new document {categorified classed} victimization possibilities for each class. The classification framework is instanced as, we consider D set of tuples and each tuple consists of attribute vector X(x1, x2, x3, xn) of n dimensions. Let their unit of measurement k form of classes C1, C2, C3... Ck. The classifier predicts X belongs to C i if

$$P(C_i|X) = P(Ci|X) \text{ for } 1<=j<=k, j<>I \quad (1)$$

We can calculate Posterior probability as:

$$P(C_i|X) = \frac{P(X|C_i)\,P(c_i)}{P(X)} \quad (2)$$

2. DECISION TREE

A decision tree is a classification model based on supervised learning in which internal nodes represent the features of the data set, and each leaf represents the outcomes [12]. The leaf nodes represent the ultimate categories of the information points. This model uses data with defined labels to create the decision tree which is then applied to the test data. For each node within the tree, the most effective check condition or

$$GINI(t) = 1-\sum_j [p(j|t)]^2 \quad (3)$$

call has P to be taken. We tend to use the GINI issue to determine the best split. For a given node t, wherever p (j|t) is that the relative frequency of sophistication j at node t.

3.RANDOM FOREST

Random forest is based on ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and improves the performance of the model as it reduces the over fitting by averaging the result. It is a classifier that contains a aggregate of decision trees on various subsets of the given dataset. A greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting. Random forests are defined by the equation as follows:

$$p = \text{model}\{T_1(y), T_2(y). \ . \ . \ .T_m(y)\} \quad (4)$$

$$p = \text{model}\ \{^m\sum_{m=1} T_m(y)\} \quad (5)$$

4. Support Vector Machine

Support vector machines, based on a supervised learning algorithm is a non-probabilistic binary linear classifier. It creates the best line or decision boundary that can segregate n-dimensional space into classes [13]. For a training set of points (x i, y i) wherever x is the feature vector and y is the category, we would like to search out the maximum-margin hyper plane that divides the points with y i = one and y i = −1. The equation of the hyper plane is as follows

$$w.\ x - b = 0$$

To maximize the margin, denoted by γ is

$$\max_{w,\gamma} \gamma, s.t. \forall i, \ \gamma \le y_i (w.x_i+b) \quad (6)$$

5. XGBOOST

Xgboost could be a type of gradient boosting algorithm that produces a prediction model that's an ensemble of weak predictions calls trees. We use the association of K models by taking summation of their output in the following manner

$$\hat{y}_i = {}^k\sum_{k=1} f_k(x_i),\ f_k \in F \quad (7)$$

where F is the space of trees, x i is input and y ˆ i is final output. We have a tendency to decide to minimize the given below loss operation

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum \Omega(f_k) \quad (8)$$

$$\text{Where } \Omega(f) = \gamma T + 1/2\lambda||w||^2 \quad (9)$$

6. LOGISTIC REGRESSION

Logistic Regression is a supervised learning algorithm used for the analysis of binary information within which one or more variables are used to determine results [9]. This machine learning model is applied to predict a categorical dependant variable. To understand the correlation between the dependant variable and one or more independent variables by estimating chances by employing a logistical equation. A logistical performance or logistical curve is a sigmoid curve defined as

$$f(x) = L\,|1+e^{\wedge}(-m(v-v_O)) \quad (10)$$

IV. RESULTS

Machine learning models are trained on Google collab in python language using Sci-kit learn frameworks. The metrics executed for the evaluation of the models implemented are training accuracy, validation accuracy, and F1 score. The Results presented by the machine learning model using BoW features denotes that Support Vector Machine attains approximate accuracy of 0.97 and Logistic Regression attains a 0.96 accuracy score. These models area unit best performers once the feature set is giant as during this study. These are often applicable conditions for each SVM and LR models. RF is additionally smart in terms of accuracy with a 0.97 accuracy score. The study leverages XGBoost for racism/sexism detection and obtains an accuracy of 0.94 and F1 scores of 0.35. Results recommend that SVM and LR show higher performance once used with BoW features. Each SVM and LR model acquires an accuracy score that is comparatively higher than all other models. LR and RF each improve the accuracy with BoW features.

Model	Accuracy		F1-score
	Validation Accuracy	Training Accuracy	
Random forest	0.951	0.999	0.603
Logistic Regression	0.941	0.985	0.593
Decision Tree Classifier	0.933	0.999	0.540
Support Vector Machine	0.952	0.978	0.498
XGBoost	0.944	0.943	0.353

Table 1. Results using machine learning model with BoW features

The above-given table I displays the accuracy and f1-score of the applied machine learning techniques to the sentimental

analysis model. Here, accuracy is scaled as training and validation accuracy, where training accuracy describes how the model will be classifying two images throughout training on the training dataset and validation accuracy signifies how images with the validation dataset will be classified by the model. The F1 score in this table implies the stability between preciseness and recall.

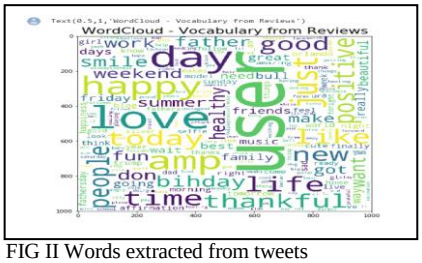


FIG II Words extracted from tweets

The development within the performance is because of easy BoW options that aid in higher coaching of machine learning models. BoW provides an easy set that will be a lot of applicable for coaching machine learning models. On average, the performance of the machine learning models are healthier with using the BoW feature.

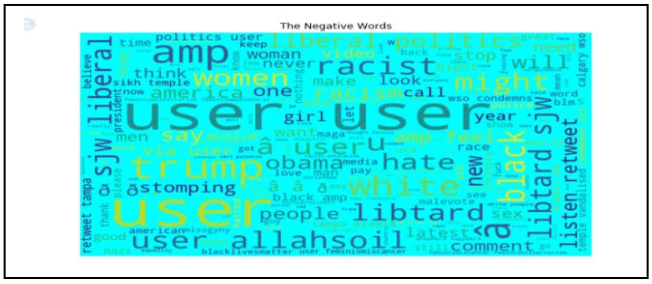


FIG III Negative words extracted from tweets

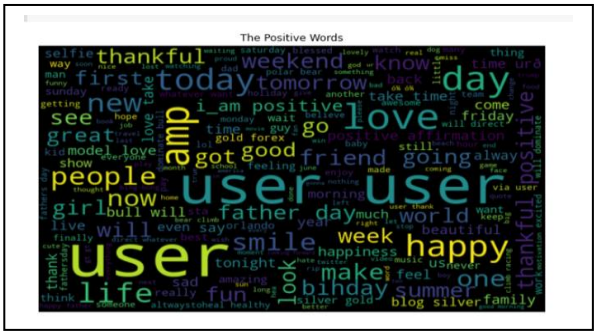


FIG III Positive words extracted from tweets

This study aims at distinguishing racist content announced within the tweets by acting sentiment analysis. For this purpose, the dataset is annotated into positive and negative categories. Positive categories indicate that the content provided in tweets isn't racist/sexist whereas the negative category portrays that these tweets are racist/sexist as they contain negative views which amplify racism. Therefore the distribution of correct and wrong predictions and accuracy provided here is concerning the negative category. The collected dataset contains a complete of 31,962 train tweets and 17,197 test tweets including 29720 and 2242 tweets for positive, and negative tweets, respectively. Training Tweets containing negative sentiments build 7.01% of the entire tweets .Results indicate that SVM shows the potential for detective work in negative tweets with the very best accuracy of 0.97, for BoW features which suggests that 3% of racist tweets are unit by SVM. Similarly, LR properly 96% of the racist techniques however attributes 4% of the racist tweets to non-racist tweets.

V. CONCLUSION

This section briefly concludes the research on personality prediction. This paper discusses diverse machine learning algorithms to approach this problem. The methodology describes data source, data preparation, feature extraction, and classifier techniques hierarchically. Extraction and preparation of tweets play a significant role in this analysis. The accuracy of each method gives the efficiency of the applied technique in this analysis.

Racist comments have become very frequent on social media platforms like Twitter and may be mechanically detected and stopped to avoid any unfolding. This study considers racism detection from a sentiment analysis perspective and marks racism amplifying tweets by distinguishing negative sentiments.

Training and testing dataset collected from GitHub and annotated by utilizing the Text Blob is employed for experiments with many machine learning. Overall, 7.01% of the collected 31,962 train tweets and 17,197 check tweets contain racist comments. Since the negative category is vital to notice racism, a separate analysis indicates that SVM and LR respectively observed 97% and 96% of racist tweets accurately whereas 3% and 4% of the racist tweets are mislabeled, respectively. Results show that machine learning models SVM and LR show higher performance getting averaged 0.97 and 0.96 accuracy scores respectively, concerning the sentiment analysis for positive and negative categories. The accuracy/ results of every methodology permit us to imagine the potency of the applied technique in several circumstances.

VI. FUTURE WORK

We believe our results would build a brand new reference for future comparative studies. We are going to explore future work diversely, like alternative network structures to extract different options. We expect to refine our work by adding new features like not only letting people know their weak points but also giving them an idea through productive video references on how to improve themselves to excel in their respective field.

Acknowledgment

This project wouldn't be feasible without the support of many people. I am immensely obliged to my research supervisor, Mrs. Lata Gupta, for her epitome guidance, and ardent encouragement. I would also like to express our gratitude for her advice and contribution to the project and preparation of this report. I was able to procure invaluable insights during this product development and to explore and learn about many other machine learning fields. I am eternally grateful to my parents and friends for their support

and much-needed encouragement throughout this research study.

REFERENCES

[1] Deepa A, Dr. Chandramouli H, V. Chandrasekhar, Abhiman J R" Social Media Sentimental Analysis" Vol11,Issue2,FEB/2020 ISSN NO:0377-9254

[2] Ernesto Lee 1, Furqan Rustam 2, Patrick Bernard Washignton, Fatima El Barakaz, Wajdi Aljedaani, And Imran ASshraf "Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweet" VOLUME 10, 2022

[3] Sahar A. El_Rahman, " Sentiment Analysis of Twitter Data", Computer and Information sciences College Princess Nourah Bint Abdulrahman University,

[4] I. Aljarah, M. Habib, N. Hijazi, H. Faris, R. Qaddoura, B. Hammo,M. Abushariah, and M. Alfawareh, ''Intelligent detection of hate speech in arabic social network: A machine learning approach,'' J. Inf. Sci., vol. 47,no. 3, May 2020, Art. no. 0165551520917651

[5] Rasika Wagh &Payal Punde ,"Survey on Sentiment Analysis using Twitter Dataset".

[6] S. Goswami, M. Hudnurkar, and S. Ambekar, ''Fake news and hatespeech detection with machine learning and NLP,' PalArch's J. Archaeol.Egypt/Egyptol., vol. 17, no. 6, pp . 4309–4322, 2020.

[7] B. Vidgen and T. Yasseri, ''Detecting weak and strong Islamophobic hate speech on social media,'' J. Inf. Technol. Politics, vol. 17, no. 1, pp. 66–78, Jan. 2020

[8] (Mtech): Department of Computer Science and IT, Dr. BAMU Aurangabad, India, ICECA 2018.

[9] Adyan Marendra Ramadhani & Hong Soon Goo, "Twitter SentimentAnalysis using Deep Learning Methods", Department of Management Information Systems Dong-A University Busan South Korea, 2017.

[10] Bing Liu, Sentiment Analysis and Opinion Mining Morgan and Claypool Publishers, May 2012.

[11] V. Kharde and S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques",International Journal of Computer Applications, vol.139,p.112016.

[12] Huma Parveen & Prof. Shikha Pandey "Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm", Dept. of Computer Science and Engineering Rungta College of Engineering and Technology Bhilai, India , 2016.

[13] Rincy Jose & Varghese S Chooralil , "Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Classifier Ensemble Approach" ,Department of Computer Science and Engineering Rajagiri School of Engineering and technology Ernakulam, India , 2016.

[14] Volume five, Issue a pair of | ISSN: 2321-9939 IJEDR1702032 International Journal of Engineering Development and analysis (www.ijedr.org) 198.