Big Five Personality Prediction from Social Media Data using Machine Learning Techniques

Suman Maloji, Kasiprasad Mannepalli, Navya Sravani. J, K. Bhavya Sri, C. Sasidhar

Abstract: : Personality has been important for a number of types of cooperation; it has useful in predicting job achievement, expert and emotional relationship achievement, and even tendency towards a variety of interfaces. To accurately examine the characters of users, a personality test must be carried out. In numerous areas of online life it is usually impractical to use character research. . We used SVM classification, Random Forest algorithm, Naïve Bayes Algorithm and Logistic regression to comparatively predict the user's personality accurately. The main goal of the paper is to evaluate the machine learning models using the four parameters- accuracy, precision, recall, f1 score and basing upon these parameters the best machine learning model will be used to classify the big five personality traits of the twitter users.

Keywords :. Social Media, Twitter, Personality, Feature extraction.

I. INTRODUCTION

In the last decade, social networking has increased drastically. A survey of social networking sites in January 2005 estimated that approximately 115 million members were present on different social networking sites[1]. Over the last five years, only 200 million of Tweets have been reached. In which, they posted their activities, opinions, interests on social media profiles. Much of the personality of a user stems from his profile through bio, status updates, profile pictures. For decades, conduct scholars have attempted to systematically clarify personality. Studies have shown associations among general characteristics and various forms of behavior following intensive work on developing and validating a generally accepted model of personality. There has been numerous relationships between personality and psychological disorders[2],job performance[3] , happiness, and even romantic success[4]. This paper uses the information that people disclose in their social media platforms to sort the distance between social media and personality characteristics. Our main problem is to address whether social media profiles can predict personality traits of a person. If so, various discoveries on the effects of personality and behavior variables can be incorporated into online user interactions as well as the use of social media profiles to help people understand each other better. We start with the Big Five Personality File Foundation, associated characteristics and web-based life work. At this time, we present our testing arrangements and strategies for Twitter profile data dissection and measurement.

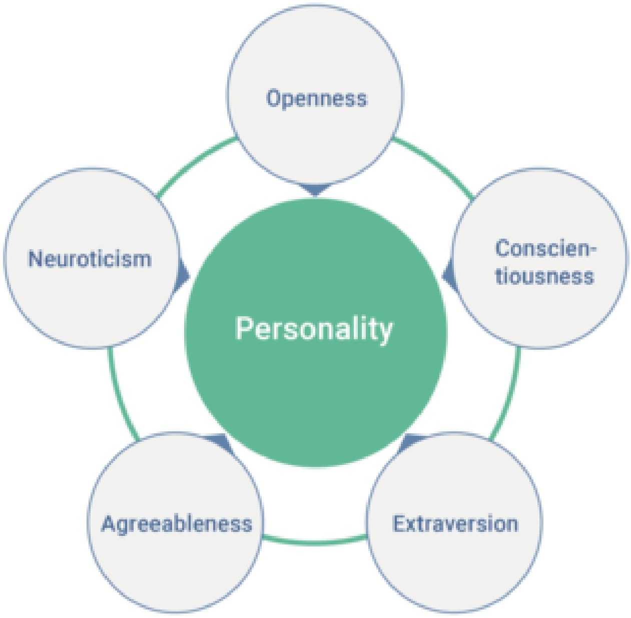


Fig.1. Big Five Personality Model

Openness means intelligent people who express their view in bold or open manner. This user expression can be identified by analysing his twitter profile and twitter messages, a person are intelligent if he uses open words or bold words in his tweets. By looking for such words we can categorize this person under Openness personality trait. LIWC dictionary contains all open or swear words by applying this dictionary on tweets messages we can predict Openness personality score. If predicted score > 0.1 then this person will put under this category. Agreeable means peoples who use words such as 'am, will have and these words can also be referred as ARTICLES or AUXILIARY VERBS' etc, will come in this category.

Revised Manuscript Received on April 27, 2020.

* Correspondence Author
Dr. Suman Maloji*, Professor,Department of Electronics and Communication Engineering.
Dr. Kasiprasad Mannepalli, Associate Professor, Electronics and Communication Engineering Department.
Navya Sravani. J, Student, Electronics and Communication Engineering Department, Koneru Lakshmaiah Education Foundation
K. Bhavya Sri, Student, Electronics and Communication Engineering Department, Koneru Lakshmaiah Education Foundation
C.Sasidhar, Student, Electronics and Communication Engineering Department, Koneru Lakshmaiah Education Foundation

MRC dictionary contains all words of this category and by applying this dictionary on user's tweets we can categorize the user's personality trait as agreeable. In the category of Neuroticism, people are considered as emotional, people who use words such as 'ugly, nasty, sad' etc., will come under this category. By looking for such words in tweets we can predict score of this category. Extroverts are usually friendly and people who have much number of friends or followers or following in twitter profile come under this category. Conscientious people are generally more determined, hardworking, and present organized ideas in their respective work fields.

II. RELATED WORK

A. Online Social Networking and Marketing Communication Insights

Despite the fact that online informal community administrations have gotten tremendously well known among overall population, there is a laxity of experimental examinations on the person's level in this area. This paper inspects the effect of character factors, for example, extraversion, confidence, feeling chasing and conclusion administration on brand correspondence and online social practices [5]. Our outcomes show that sex and extroversion anticipate online interpersonal organization size and time spent on the web; that feeling searchers invest more energy on the web and have bigger systems comparative with conclusion pioneers; and that sentiment heads are bound to convey their image utilize on the web. We likewise discover the intervening job of conclusion authority and supposition looking for in clarifying the effect of general character attributes on online brand correspondence and interpersonal interaction [6].

B. Relationship between the dimension of the big five personalities and employment

In this examination, the research has been done on directing job of self-sufficiency on the connections between the Big Five character measurements and boss appraisals of occupation execution. Drawing on data from 146 directories, results show that the main characteristics identified with the execution of the project are two character measures: Conscientiousness (r= 25) and Extraversion (r= 14). In keeping with our wishes, for the directors in high self-regulation and in low self-regulation employments, the legitimacy of conscience and extraversion was more prominent.[7] For highly independent and low self-regulated professions, the validity of Agreeableness was also higher, but the relation was adverse. These discoveries propose that level of self-governance in the activity directs the legitimacy of probably some character indicators. Suggestions for future explorations are noted.

C. The right relationship is all: the link between preferences for personality and management compartments

Singular contrasts and character variables have reappeared as a portion of the more significant exploration subjects in the applied hierarchical sciences. With the expanding predominance of official instructing and the utilization of character evaluations, more research should be done on the effect of character factors on administrative practices in the working environment. The accompanying investigation gives an applied examination of character inclinations and social appraisals gathered for a formative multilateral criticism intercession dependent on 343 ranking directors and others in an exploration driven worldwide wellbeing administrations association. Results uncovered unobtrusive character conduct connections, huge numbers of which were predictable with Myers-Briggs Type Indicator hypothesis and discoveries; contrasts by onlooker viewpoint were additionally apparent. Suggestions for HRD practice are talked about [8].

D. Improving reliability of feedback by trust-building social networks

Social trust connections between clients in interpersonal organizations address the comparability in sentiments between the clients, both by and large and in significant nuanced ways. They have been utilized in the past to make proposals on the web. New trust measurements enable us to effectively bunch clients dependent on trust. In this paper, we explore the utilization of trust bunches as another method for improving proposals. Past take a shot at the utilization of groups has demonstrated the system to be generally ineffective, however those bunches depended on similitude instead of trust. Our outcomes show that when trust groups are coordinated into memory-based communitarian separating calculations, they lead to factually huge upgrades in precision. In this paper we talk about our strategies, examinations, results, and potential future uses of the strategy[9].

E. Predicting social media connection power

The social media treats everyone the same: a trustworthy friend or an alien, with little or nothing. In fact, relationships fall across this range, a topic of social science has been researching the theme of link strength for decades. This gap between practice and theory is linked to our work. In this article, we present a predictive model to link social media data[10].

The model draws on a data set of over 10,000 social media links and is well tested, which differentiates strong and weak relations with more than 85% accuracy. Interviews complementing these quantitative findings show the relationships that we were unable to predict.

III. METHODOLOGY

A presentation is made basing on strategy for precise anticipation of the character of a client through the openly accessible data on their Twitter profiles. We will now exhibit the AI procedures that allow us to determine the personalities effectively. We start by showing the foundation on the file of the Big Five Personality and related work on life based character and on the internet. We also present results inorder to understand the link between the original character and online networking profiles. Having this in mind, we portray the AI procedures used for grouping and show how we can achieve enormous and critical upgrades on each character factor over pattern characterization.

Retrieval Number: D7946049420/2020©BEIESP
DOI: 10.35940/ijeat.D7946.049420
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication
© Copyright: All rights reserved.

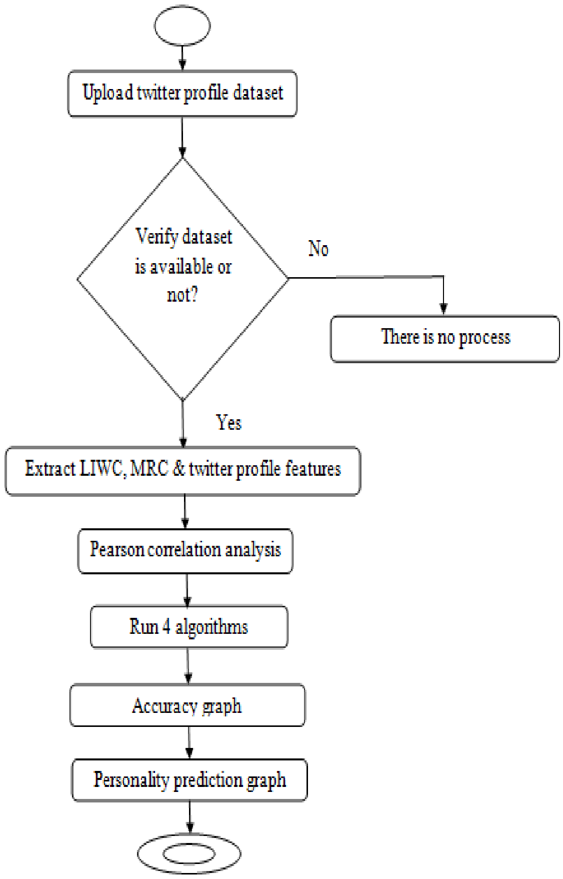
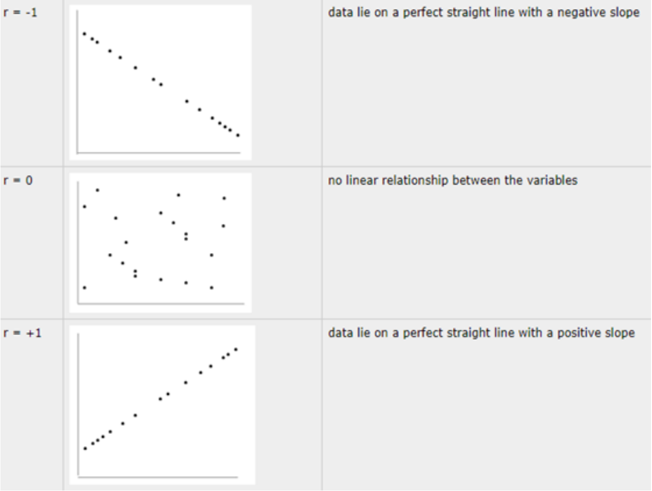

Upload twitter profile dataset
Verify dataset is available or not?
No
There is no process
Yes
Extract LIWC, MRC & twitter profile features
Pearson correlation analysis
Run 4 algorithms
Accuracy graph
Personality prediction graph

Fig.2.Flow of the project

Step 1: Loading twitter dataset and extracting LIWC and MRC features.

Step 2: Obtaining the Pearson correlation analysis

Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables[11].



r = -1		data lie on a perfect straight line with a negative slope
r = 0		no linear relationship between the variables
r = +1		data lie on a perfect straight line with a positive slope

Fig.3. Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1

Step 3: Run the four machine learning algorithms and evaluate the model's accuracy, precision, recall and f1 score. The formulas for calculating the performance metrics are

T1= True Positive

F1= False Negative

T2= True Negative

F2= False Positive

Accuracy is the ability to determine the correctness or closeness of personality categorization. The formula for accuracy can be given by

Accuracy = (T2+ T1)/ (T2+ T1+ F1+ F2)

Precision refers to the closeness/ correctness of two or more than two values. The formula for precision can be given by

Precision = T1/ (T1 + F2)

Recall is the fraction of the total amount of relevant instances that were actually retrieved. The formula for recall can be given by

Recall=T1/ (T1+F1)

F1 score is harmonic mean of Precision and Recall. The formula f1 score can be given by

F1 score= (2*T1)/ ((2*T1) +F2+F1)

In this paper we performed comparative analysis of machine learning models for Big five personality prediction. We are considering four machine learning algorithms namely Support Vector Machines (SVM), Naïve Bayes, Random Forest, and Logistic Regression.

A. Support Vector Machine

A Support Vector Machine (SVM) is a supervised machine learning classifier that is formally defined by an isolating hyperplane which is used for two-group classification [12]. Because of marked preparation information (administered learning) at the end of the day, the calculation yields an ideal hyperplane that commands new models. This hyperplane is a line isolating a plane in two sections in two dimensional spaces where it lay on either side in each class[13]. Our data has two features: red and blue. We want a classifier that, given a pair of (red, blue) coordinates, outputs if it's either red or blue. We plot our already labelled training data on a plane:

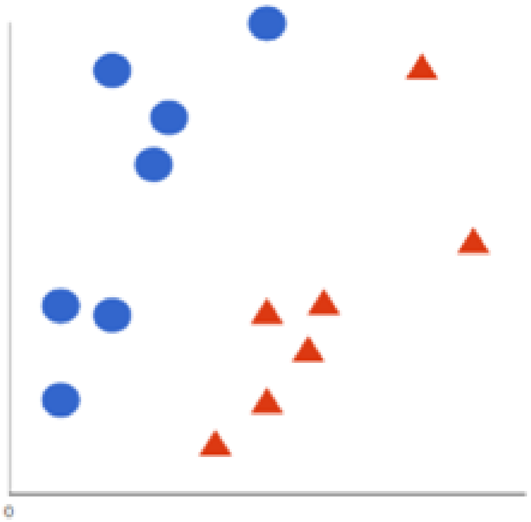


Fig. 4. Labelled data

A support vector machine takes these data points and outputs the hyperplane (which in two dimensions it's simply a line) that best separates the tags. This line is the decision boundary: anything that falls to one side of it we will classify as blue, and anything that falls to the other as red[14].

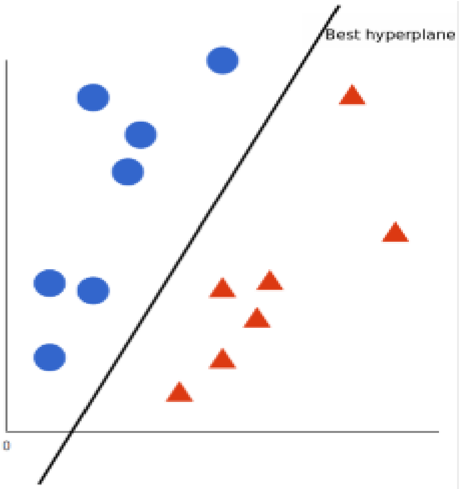


Fig. 5. In 2D, the best hyperplane is simply a line

B. Random Forest Algorithm

Random forest algorithm is a calculation of the administration of the arrangement. Random forest has found its wide spread use in various applications. The acceptability of random forest can be primarily attributed to its capability of efficiently handling non-linear classification task. Random forest is well-known for taking care of data imbalances in different classes especially for large datasets [15]. The ability to precisely classify observations is extremely valuable for various business applications like predicting whether a particular user will buy a product or forecasting whether a given loan will default or not. Decision trees as they are the building blocks of the random forest model.

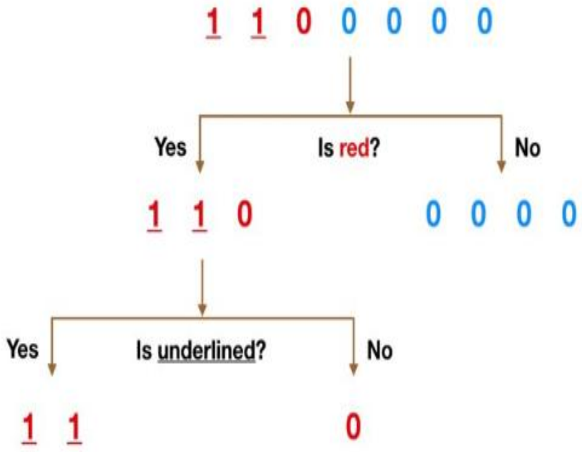


Fig. 6. Simple Decision Tree example

Random forest, as its name implies, is made up of a large number of individual decision trees which act as an ensemble. Every single tree in the random forest spits out a class prediction and the class with the most votes is the prediction of our model. The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is: A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models [16].

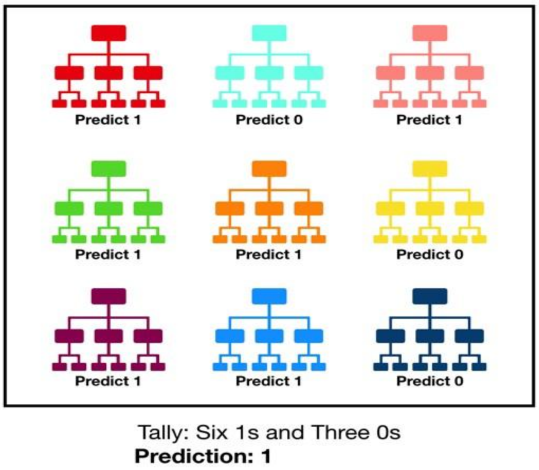
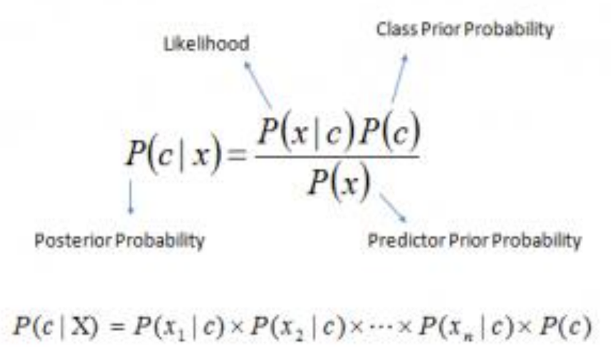


Fig.7. Visualization of a Random Forest Model Making a Prediction

C. Naive Bayes Algorithm

A Naive Bayes Algorithm is used for calculation using the Bayes hypothesis to arrange objects. It has an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c).



$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Fig 8: Naive Bayesian equation

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Steps involved in Naïve Bayes Algorithm:

Step 1: Obtain the prior probability P(c) for the given class labels.

Step 2: Calculate Likelihood probability P(x|c) with each attribute for each class.

Step 3: Substitute these values in Naïve Bayesian equation and calculate the posterior probability P(c|x).

Compared to other algorithms, Naive Bayes classifiers often used in text classification (due to better results in multi-class problems and independence rule) have higher success rate. As a consequence, spam filtering (identifying spam e-mail) and sentiment analysis (identifying positive and negative consumer sentiments in social media data) are commonly used [17].

D. Logistic Regression

Logistic regression is a measurable procedure used to predict the likelihood of a coupled reaction that depends on at least one free factor. It implies that strategic relapse, given a specific component, is used to predict a result that has two qualities, e.g. 0 or 1, pass or come up short, yes or no, and so on.

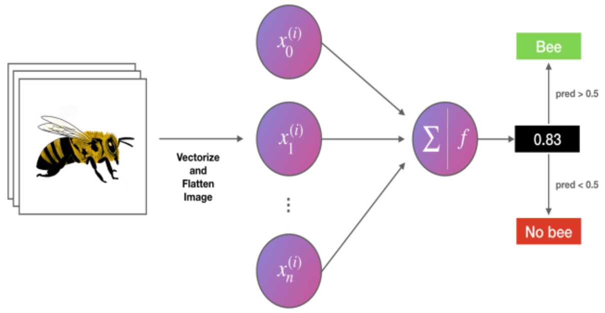


Fig.9. Simple Logistic Regression Model

There are three types of Logistic Regression techniques. They include:

- Binary Logistic Regression: The categorical response has only two 2 possible outcomes.

Example: Spam or Not

- Multinomial Logistic Regression: Three or more categories without ordering.

Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)

- Ordinal Logistic Regression: Three or more categories with ordering.

Example: Movie rating from 1 to 5

To predict which class a data belongs, a threshold can be set. Based upon this threshold, the obtained estimated probability is classified into classes. Decision boundary can be linear or non-linear. Polynomial order can be increased to get complex decision boundary [18].

$$\mathrm{Cost}(h_\theta(x), Y(\mathrm{actual})) = -\log(h_\theta(x)) \text{ if } y=1$$
$$-\log(1-h_\theta(x)) \text{ if } y=0$$

Fig.10. Cost Function of Logistic Regression

E. Dataset Description

The twitter dataset is openly available social media dataset called 'Twitter Profile' which of 10,000 tweets that includes columns like: Username, Tweet text, Followers, Following, Density, Hashtag, Tweet word length. To run this project we used twitter dataset which contains tweets and user details in JSON format. The dataset is available inside 'tweets' folder and each file contain user details and tweeted data of 1 user.

IV. RESULT AND DISCUSSION

This paper aims to predict human personality by considering five features such as Openness, Agreeableness, Neuroticism, Extroversion, and Conscientious.

Table. I. Precision, Recall, F1 score, Accuracy values

Algorithm	Precision	Recall	F1 score	Accuracy
SVM algorithm	0.78	0.88	0.82	0.88
Naïve Bayes algorithm	0.74	0.88	0.80	0.875
Random forest algorithm	0.14	0.38	0.20	0.375
Logistic Regression algorithm	0.78	0.62	0.65	0.625

From the results, it is a clear evident that among the machine learning algorithms used- SVM algorithm tops the accuracy report when compared with the other algorithms, producing an accuracy of 88% while Naïve Bayes algorithm has produced an accuracy of 87.5%, Logistic Regression algorithm has produced an accuracy of 62.5% and Random Forest algorithm has produced least accuracy of 37.5%. When considering three parameters recall, precision and f1 score, the SVM algorithm gives best results among the all four machine learning algorithms.

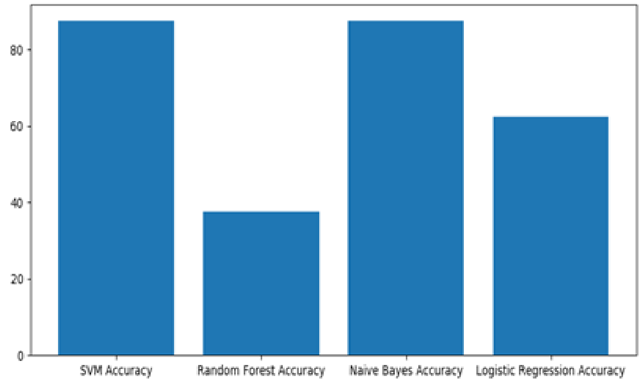


Fig.11.Accuracy Graph

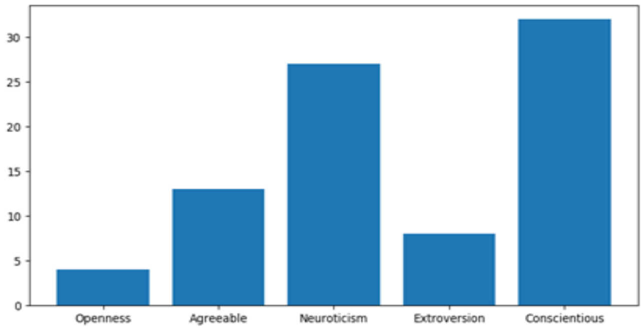


Fig.12. Big Five Personality Traits Classification Graph

DOI: 10.35940/ijeat.D7946.049420
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. CONCLUSION

The work has clearly explained the classification of users personalities basing upon of Big Five Personality traits from the open data that they share on Twitter.

The results have clearly explained the comparisons between the four machine learning models and concluded that, among SVM, Random forest algorithm, Naïve Bayes algorithm and logistic regression- SVM algorithm has topped the table by producing an accuracy of 88% with its classification mechanism.

Also, among the different tweets accessed from the database, users seem to present more conscientiousness with their routine lives when compared with the other four personality traits among the Big Five Personality traits prediction. In future, more work could be extended by including optimization, developing a web application or a mobile application to analyze personalities of different people from the text they provide depending on their moods which helps the model to give more accurate results with less computational time.

ACKNOWLEDGEMENT

We thankfully acknowledge management of KL deemed to be University for providing their extendible support and required resources for successful completion of this work.

REFERENCES

1. J. Golbeck. Computing and Applying Trust in Web-based Social Networks. PhD thesis, University of Maryland, College Park, MD, USA, 2005.
2. L. Saulsman and A. Page. The five-factor model and personality disorder empirical literature: A meta-analytic review* 1. Clinical Psychology Review, 2004.
3. M. Barrick and M. Mount. The Big Five personality dimensions and job performance: A meta-analysis. Personnel psychology, 1991.
4. P. Shaver and K. Brennan. Attachment styles and the “Big Five” personality traits: Their connections with each other and with romantic relationship outcomes. Personality and Social Psychology Bulletin,11992.
5. Acar and M. Polonsky. Online Social Networks and Insights into Marketing Communications. Journal of Internet Commerce, 2008.
6. M. Back, J. Stopfer, S. Vazire, S. Gaddis, S. Schmukle, B. Egloff, and S. Gosling. Facebook Profiles Reflect Actual Personality, Not SelfIdealization. Psychological Science, 2010.
7. M. Barrick and M. Mount. Autonomy as a moderator of the relationships between the Big Five personality dimensions and job performance. Journal of Applied Psychology, 1993.
8. S. Berr, A. Church, and J. Waclawski. The right relationship is everything: Linking personality preferences to managerial behaviours. Human Resource Development Quarterly, 2000.
9. Zhou, Ming & Dresner, Martin & Windle, Robert. Revisiting feedback systems: Trust building in digital markets. Information & Management, 2009.
10. Schoen, Harald & Gayo-Avello, Daniel & Metaxas, Panagiotis & Mustafaraj, Eni & Strohmaier, Markus & Gloor, Peter. The power of prediction with social media. Internet Research: Electronic Networking Applications and Policy, 2013.
11. University of the West of England, Bristol (2020). Data Analysis. Retrieved from http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442
12. M. E. Mavroforakis and S. Theodoridis, "Support Vector Machine (SVM) classification through geometry," 2005 13th European Signal Processing Conference, Antalya, 2005.
13. K. Mannepalli, P. Sastry, M. Suman, ―Emotion recognition in speech signals using optimization based multi-SVNN classifier‖, J. King Saud Univ. – Computer Inform ci, 2018
14. Bruno Stecanella (2017, June 22). An introduction to Support Vector Machines (SVM). Retrieved from https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/
15. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha and S. Kundu, "Improved Random Forest for Classification," in IEEE Transactions on Image Processing, 2018.
16. Tony Yiu (2019, June 12). Understanding Random Forest, How the Algorithm Works. Retrieved from https://towardsdatascience.com/understanding-random-forest-58381e0602d2
17. Sunil Ray (2017, September 11). 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R. Retrieved from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
18. Saishruthi Swaminathan (2018, March 15). Logistic Regression — Detailed Overview. Retrieved from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
19. De Raad. The Big Five personality factors: The psycholexical approach to personality. Hogrefe & Huber G ”ottingen, 2000.

AUTHORS PROFILE

Dr. Suman Maloji is a professor in the Department of Electronics and Communication Engineering. He has published many number of articles in various international and national journals. His research interests are Speech coding, Speech compression, speech, and speaker recognition.

Dr. Kasiprasad Mannepalli, Associate Professor of Electronics and Communication Engineering department. He received his Doctoral degree in the field of speech signal processing from Koneru Lakshmaiah Education Foundation. Currently, he is guiding four members for their Doctoral degree. His research interests are Emotional speech recognition, Accent recognition, and pathological speech processing. His research interests are Signal, Image and Speech processing.

Navya Sravani. J, is a student of Electronics and Communication Engineering department. She is currently pursuing final year in Bachelor of Technology from Koneru Lakshmaiah Education Foundation. Her research interests include Signal, Image and Speech processing; Data Science and Machine Learning.

K. Bhavya Sri , is a student of Electronics and Communication Engineering department. She is currently pursuing final year in Bachelor of Technology from Koneru Lakshmaiah Education Foundation. Her research interests include Signal processing and Machine Learning.

C.Sasidhar , is a student of Electronics and Communication Engineering department. He is currently pursuing final year in Bachelor of Technology from Koneru Lakshmaiah Education Foundation. His research interests include Signal , Image processing and Machine Learning.